Evaluating racial clusters as a unit of aggregation for neighborhood effects.

Extended Abstract

Draft. Do not cite.

Jonathan Tannen
Ph.D. Candidate, Princeton University

March 24, 2014

Introduction

The study of neighborhoods has recently focused on identifying the 'correct' level of aggregation (Clapp and Wang, 2006; Hipp, 2007; Flowerdew et al., 2008; Nau, 2013). While our theory of neighborhoods is complex and nuanced, our quantitative study of them has often been limited to aggregating observations to pre-selected units, such as the Census tract. We certainly don't believe these units to represent the true, on-the-ground neighborhood divisions, but we assume they are correlated with the spatial effects of neighborhoods and thus provide relatively-explanatory, extremely-convenient proxies. Neighborhood theories, meanwhile, have emphasized the social structure of neighborhoods, often divided by race and class boundaries. These perspectives imply an emergent form of neighborhood boundary rather than a fixed one, defined only by humans' use of the space and perhaps able to move over time. This paper proposes a method for using data-identified clusters to identify such emergent neighborhood boundaries, and evaluates the performance of these clusters–versus census geographies–as a hierarchical level in predicting crime rates in Philadelphia in 2008-2012.

Neighborhoods as Demographic Clusters

Neighborhoods as areas of demographic segregation forms a strong thread through the history of neighborhoods literature. The "natural areas" of Park and Burgess (1925) consisted of residents with similar social characteristics, segregated by similar responses to market forces and by the experienced ecology. Suttles (1972) showed that even within small units of space, ethnicity divided the residents' conception and use of a neighborhood. The recent renaissance of neighborhood literature has focused intently on the effect of segregation and hyper-concentration of poverty within neighborhoods, building on the theories of Wilson (1987) and Massey and Denton (1993).

Neighborhoods have many theoretical definitions; they are, at a single time, collections of similar people, spatially bounded social processes, and areas of correlated outcomes. A challenge in neighborhood literature is merging all of these intuitive definitions. The hidden logic in much of the early research is that these areas will often be the same: areas of similar demographics will serve as the divided area within which social processes operate, and will be the same spaces in which we observe similar outcomes. Here, I construct cluster definitions based on similar household race and ethnicity–the first intuitive concept–and measure the correlation in crime outcomes–the third concept–without directly measuring social processes.

Block-level racial demographics of American cities exhibit very strong racial clustering. Figure 1 maps household race and ethnicity for a region in South Philadelphia in the 2010 Census, illustrating sharp spatial transitions between racially extremelydifferent regions. These regions do not map directly onto Census tract boundaries; they often span multiple tracts or have boundaries that split a tract.

[Figure 1 about here.]

Consider a resident living in, e.g., an all-White block close to one of the demographic boundaries clearly visible on the map. The question I ask here is 'What areal unit best captures the spatial effects this resident experiences?' Is their experience best proxied by the traits of the entire tract? By their exact block? Or by the homogeneous region in which their block resides, delineated by the sharp boundary clearly visible to the eye? My hypothesis is that the boundaries of these clusters are consistent with boundaries in social and activity spaces of residents' experiences, and thus will capture an important fraction of the spatial correlation of outcomes.

The comparative strength of these boundaries has important implications for the theoretical understanding of neighborhood mechanisms, but also methodological implications for neighborhood effects research. The problem of mis-specified spatial aggregation units is termed the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). The MAUP states that using arbitrary or misaligned spatial units can lead to severe bias or highly varying results; this lack of robustness to an arbitrary decision is problematic. As proposed in the very first paper on the MAUP (ibid.), a solution is to look for the units that best capture spatial correlations in outcomes, and provide the best fit in a multi-level analysis. We want to identify boundaries that are both real on the ground and provide robust, explanatory units. These racial boundaries are justified by theories of race and neighborhoods; should they also capture spatial correlations in neighborhood outcomes, they would be very appealing neighborhood-level units for aggregation.

Method

The process of aggregating polygons into clusters hasa strong tradition termed 'Regionalization' in the Geography literature. The method I use here, however, comes from the problem of Image Segmentation in Machine Learning: the distance dependent Chinese restaurant process (ddCRP) (Blei and Frazier, 2011; Ghosh et al., 2011). This method is very similar in idea to the graph-based spanning tree algorithms of Regionalization, first proposed by Maravalle and Simeone (1995), but has the benefit of a relatively simple generative model in a Bayesian framework, allowing for straightforward extensions and for incorporating the full set of tools developed for Bayesian techniques.

ToThe ddCRP is a bayesian non-parametric probability distribution over partitions. Each block i is given an assignment c_i, the index of a neighboring block or itself.

The assignments form a directed graph; disconnected clusters in the graph define the cluster assignment z_i. Figure 2 shows a sample ddCRP cluster realization.

[Figure 2 about here.]

Let G be the neighbor graph of the blocks. The probability of block i connecting to block j, $i \neq j$, is a function of the distance between them. Here, I use a neighbor distance function with window 1, meaning that blocks can only connect to their immediate neighbors or themselves. I have slightly modified the original ddCRP so that all directed loops confer a probability of α (rather than only nodes pointing to themselves), so that the probability of a single assignment is given by

$$p(c_i = j | c_{-i}, G, \alpha) \propto \begin{cases} \alpha, & i = j \\ 1, & i \sim j, j! \rightarrow i \\ \alpha, & i \sim j, j \rightarrow i \\ 0, & otherwise, \end{cases}$$

where $i \sim j$ symbolizes that blocks i and j are neighbors, and $j \rightarrow i$ symbolizes that c_{-i} has a path from j to i, and thus connecting $c_i = j$ would form a loop. The number of clusters is equivalent to the number of directed loops in the full graph (it's mechanically impossible to have a non-directed loop), including blocks pointing to themselves.

Conditional on these cluster assignments, each block draws the number of households of each race and ethnicity, X_i, from a multinomial distribution with cluster-level parameter, p_z. The p_z have a Dirichlet prior parametrized by p_0. The generative model is

- $c_{1:N} \sim ddCRP(G, \alpha)$. This defines cluster assignments $z_{1:N}$.
- $p_z \sim Dirichlet(p_0)$.
- $X_i \sim Multinomial(p_{z_i})$,

in which I use the vague prior $p_0 = \vec{\mathbf{1}}_{N_{race}}$. I sample from the posterior of this model conditional on observed X_i using Gibbs sampling.

Clustering Results

The ddCRP model outlined above was fit on 2010 Census block-level household race and ethnicity data for Philadelphia. Race and ethnicity combinations waere divided into 8 groups: Non-hispanic (NH) White, NH Black, NH Asian, NH Native

American/American Indian, NH Hawaiin/Pacific Islander, NH Other, NH Two or more, and Hispanic. While the ddCRP is a non-parametric model–meaning that the number of clusters is not fixed ahead of time–the parameter α adjusts the number of clusters by penalizing new loops, and thus new clusters. I used Gibbs sampling to sample four chains for values of α of $1, 10^{-10}, 10^{-20}, 10^{-100}, 10^{-200}$, chosen to give an array of numbers of clusters.

Figure 3 presents a map of cluster assignments drawn for each value of α. The number of clusters identified ranges from a mean of 31 for $\alpha = 10^{-200}$ to a mean of 1186 for $\alpha = 1$; for comparison, Philadelphia has 384 tracts, 1,336 block groups, and 18,872 blocks.

[Figure 3 about here.]

These clusters suggest a city that is highly segregated; significantly more segregated than tract or block-group indices represent. Figure 4 presents Theil's H index of segregation calculated for each level of clustering. The index is based on information theory entropy, and ranges from 0 (each group perfectly matches the proportions of the entire city, and is thus entirely non-segregated) to 1 (the units are perfectly segregated). The indices for the race-based clusters are much higher than those calculated for the census geographies. This is intuitive; by creating groups of blocks with similar racial proportions, we will obviously measure more racial segregation[1]. What is surprising, though, is how few clusters are needed to reach such high segregation indices: for example, we can achieve the level of segregation of 1,336 block groups ($H = 0.396$) with fewer than 100 clusters. It is well established that at smaller scales, segregation indices increase dramatically (Wong, 1997), yet here we can very effectively segregate the city with very few divisions.

[Figure 4 about here.]

Crime Analysis

The fact that a differently drawn set of lines on a map achieves higher segregation would not be important unless these lines in some way represent an on-the-ground truth better than another. For evidence of that social importance, I turn to analyzing crime data.

Crime is a well-studied neighborhood-level phenomenon, experiencing strong neighborhood correlations (e.g. Leventhal and Brooks-Gunn, 2000; Sampson et al., 2002).

[1] More subtly, clusters are not necessarily of a single race, as they may be defined as groups of blocks of similarly mixed populations. But broadly, the intution that clusters will yield higher segregation indices is true.

To assess the explanatory power of cluster boundaries versus census tracts, I measure the correlation among crime rates within each of these aggregating units; if the cluster boundaries better capture the correlations in crime rates, then whatever neighborhood mechanism is causing that correlation is should be bound within those clusters as well.

Crime data comes from the Philadelphia Police Department Crime Incidents dataset via OpenDataPhilly (Philadelphia Public Interest Information Network, 2013). I separately aggregated Homicides and Aggravated Assaults in the years 2008-2012 to the census blocks. Figure 5 presents the crime counts for blocks as a function of block proportion non-Hispanic white.

[Figure 5 about here.]

I model crime counts as a multilevel poisson with log link function, with both cluster and tract crossed random effects. The count Y_i of a given category of crime for block i, with cluster membership z_i and tract membership t_i is first naively modeled as

- $Y_i \sim Poisson(\mu_i)$,
- $\mu_i = exp\{\alpha + \gamma_{z_i} + \lambda_{t_i}\}$,
- $\gamma_z \sim Normal(0, \sigma_z^2)$,
- $\lambda_t \sim Normal(0, \sigma_t^2)$,
- $\sigma_{z/t}^2 \sim Inv.Gamma(1, 1)$.

where γ_z and λ_t are the hierarchical effects of cluster z and tract t, respectively, and capture the intra-unit correlation among blocks. The variance of the aggregation-unit's effects–σ_z^2 or σ_t^2–is a measure of correlation among blocks in the same group; aggregation units with a higher-variance effect have lower between-group correlations, and are dividing the outcomes more effectively. I fit the model by selecting a single clustering assignment from each cluster chain above (four for each α), and fitting the crime model for each with a single chain in the rstan package for R (Stan Development Team, 2014), yielding four estimates of σ_z^2, σ_t^2 for each α. Figure 6 presents the posterior distribution of the variance for each α and each crime. As a sample calculation, for $\alpha = 10^{-10}$, which yields on average 142 clusters, the mean of the four chains' cluster- and tract-effect variances for aggravated assault are 0.49 and 0.35, respectively. This means that the standard deviation in the effects are 0.70 and 0.59. Given the log-link function in the model, a block in a cluster one standard deviation above the mean has a predicted aggravated assault rate $e^{0.70} = 2.0$ times higher, and a block in a tract one standard deviation above the mean has a predicted rate $e^{0.59} = 1.8$ times higher. This difference is significant at the 99% level.

With the broadest clustering ($\alpha = 10^{-200}$, $N(clust) \approx 31$), the tract effects have a higher variance than the cluster effects. By the next scale of clustering ($\alpha = 10^{-100}, N(clust) \approx 45$), the variance in cluster-effects surpasses that of tract-effects, and gets subsequently much larger. At the level of a similar number of clusters and tracts ($N(tract) = 384$), the cluster variances are much higher, and thus the clusters are better capturing spatial correlations. This is entirely unsurprising: we have not controlled for block-level race, by which the clusters were defined. Any effect of block-level race would be captured in this cluster-level correlation.

[Figure 6 about here.]

To account for this, I extend the model to control for block-level race and ethnicity, as well as total population, population density, and area zoned for commercial and residential use. Block-level race and ethnicity was divided into dummy variables for bins of 0-5

- $\mu_i = exp\{\alpha + \beta X_i + \gamma_{z_i} + \lambda_{t_i}\}$,

where X_i is the vector of block-level covariates, and with the rest of the model the same as above.

[Figure 7 about here.]

The controls had opposite effects for aggravated assault and homicide. The tracts are now generally out-performing the clusters for large scales of aggregation, though they are roughly on par when the aggregation scale yields a similar number of clusters and tracts (n(tracts) = 384)[2]. For homicide, however, the benefits of clusters are striking: the clusters better capture spatial correlations in homicide then tracts at all scales. For $\alpha = 10^{-10}$, a mean of 142 clusters is produced; the cluster traits have an estimated variance of 0.21, versus 0.15 for tracts. This corresponds to a cluster with an effect one standard deviation above the mean expecting a 58% higher homicide rate, compared to a 47% higher homicide rate in a tract one standard deviation above the mean tract.

Conclusion

This project recognizes the importance of fine-scale racial boundaries in cities, and measures their explanatory power in spatial correlations in crime. The ddCRP

[2] Ongoing analyses include selecting an α that yields a similar number of cluster and tracts (n(tract) = 384), and comparing block groups as an alternate to tracts.

model provides a formal, Bayesian method to identify these racial clusters. The clusters strongly outperform tracts in capturing spatial correlations in homicides, providing evidence that the social processes which yield these spatial correlations are bounded by these same racial boundaries. The evidence for aggravated assault is less conclusive: tracts outperform the clusters at large-scale clustering, but the two perform similarly at similar aggregating scales. I am currently exploring the features of aggravated assault and homicide that yield such different results; the evidence suggests different spatial correlating mechanisms, bound differently by racial boundaries.

References

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

Clapp, J. M. and Wang, Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics*, 59(2):259–284.

Flowerdew, R., Manley, D. J., and Sabel, C. E. (2008). Neighbourhood effects on health: does it matter where you draw the boundaries? *Social Science & Medicine*, 66(6):1241–1255.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B., and Blei, D. M. (2011). Spatial distance dependent chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems*, pages 1476–1484.

Hipp, J. R. (2007). Block, tract, and levels of aggregation: Neighborhood structure and crime and disorder as a case in point. *American Sociological Review*, 72(5):659–680.

Leventhal, T. and Brooks-Gunn, J. (2000). The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychological bulletin*, 126(2):309.

Maravalle, M. and Simeone, B. (1995). A spanning tree heuristic for regional clustering. *Communications in statistics-theory and methods*, 24(3):625–639.

Massey, D. S. and Denton, N. A. (1993). *American Apartheid: Segregation and the making of the underclass*. Harvard University Press, Cambridge, MA.

Nau, C. (2013). Monte carlo simulation-based recommendations for reducing the risk of bias in multilevel models introduced by mis-measuring the neighborhood. Presented at the conference for the Population Association of America, May 2013.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Concepts and techniques in modern geography, no. 38. Elsevier Science Geo Abstracts.

Park, R. E. and Burgess, E. W. (1925). *The city*. University of Chicago Press, Chicago, IL.

Philadelphia Public Interest Information Network (2013). OpenDataPhilly. http://www.opendataphilly.org/. Last accessed on Nov 1, 2013.

Sampson, R. J., Morenoff, J. D., and Gannon-Rowley, T. (2002). Assessing “neighborhood effects”: Social processes and new directions in research. *Annual review of sociology*, pages 443–478.

Stan Development Team (2014). Stan: A c++ library for probability and sampling, version 2.2. http://mc-stan.org/.

Suttles, G. D. (1972). *The social construction of communities*, volume 111. University of Chicago Press Chicago.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press, Chicago, IL.

Wong, D. W. (1997). Spatial dependency of segregation indices. *The Canadian Geographer/Le Géographe canadien*, 41(2):128–136.

List of Figures

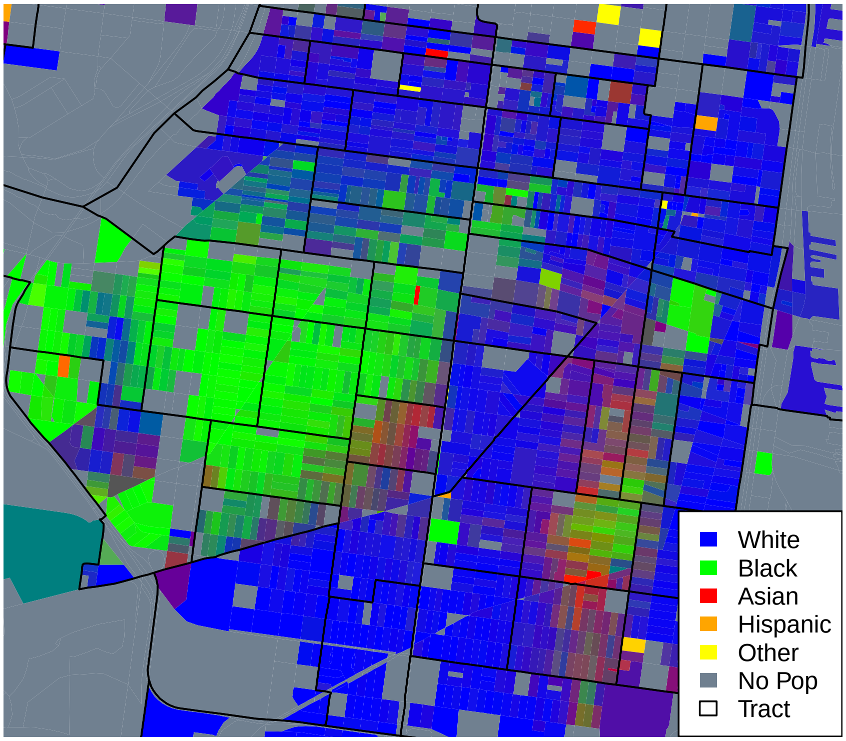

Figure 1: Map of household race and ethnicity in South Philadelphia, 2010 U.S. Census. Colors are weighted averages of household race and ethnicity on the RGB scale. White, Black, Asian, and Other are only non-Hispanic households.

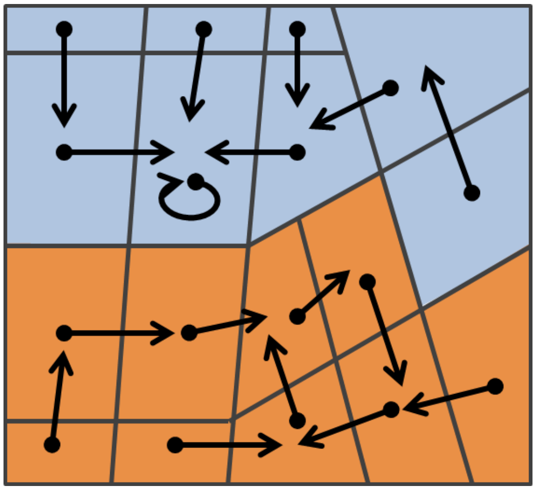

Figure 2: A sample cluster assignment of the ddCRP.

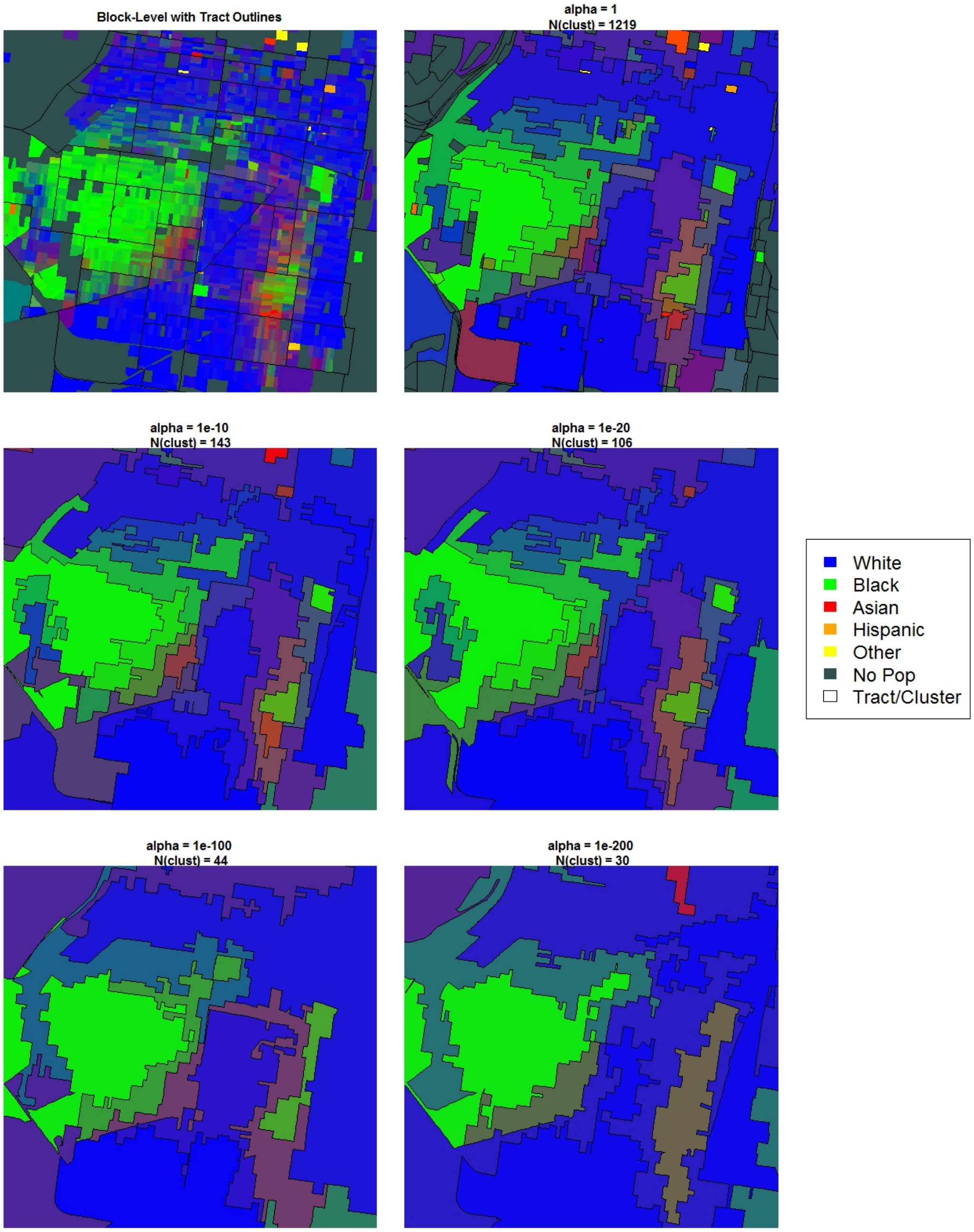

Figure 3: A sample from one chain of cluster results for each α.

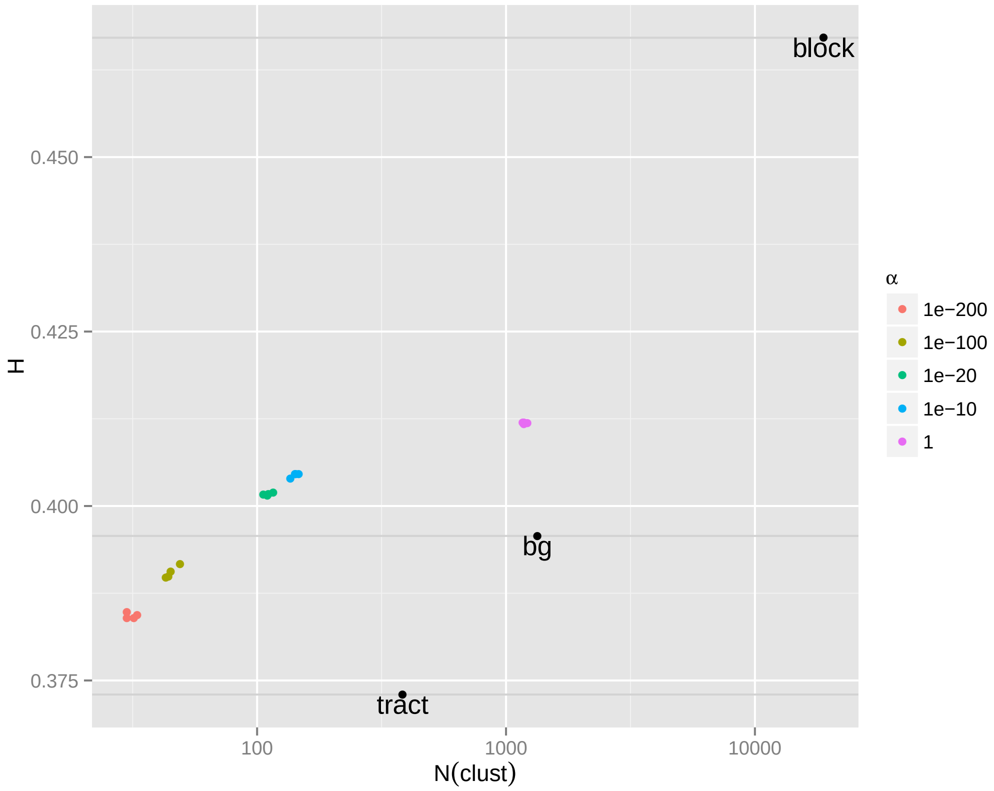

Figure 4: Theil's H index calculated for sample cluster assignments from four chains of the ddCRP model, at various α.

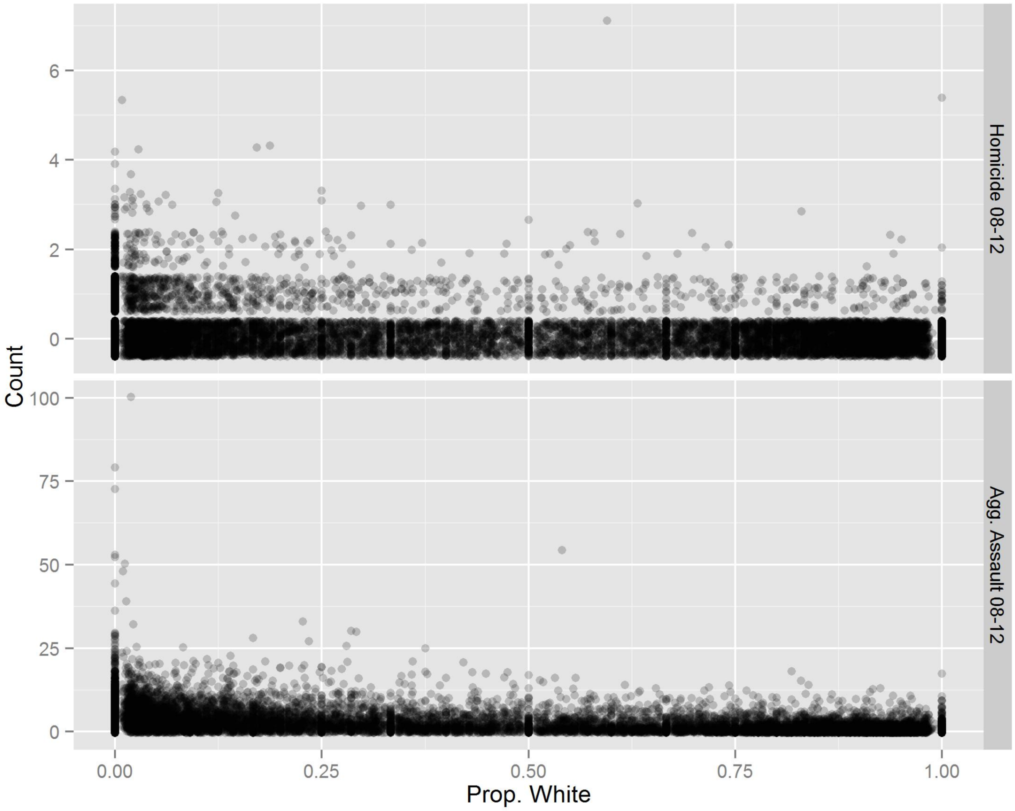

Figure 5: Crime counts from Philadelphia Police Dept. data for 2008-2012, aggregated to block-level and jittered.

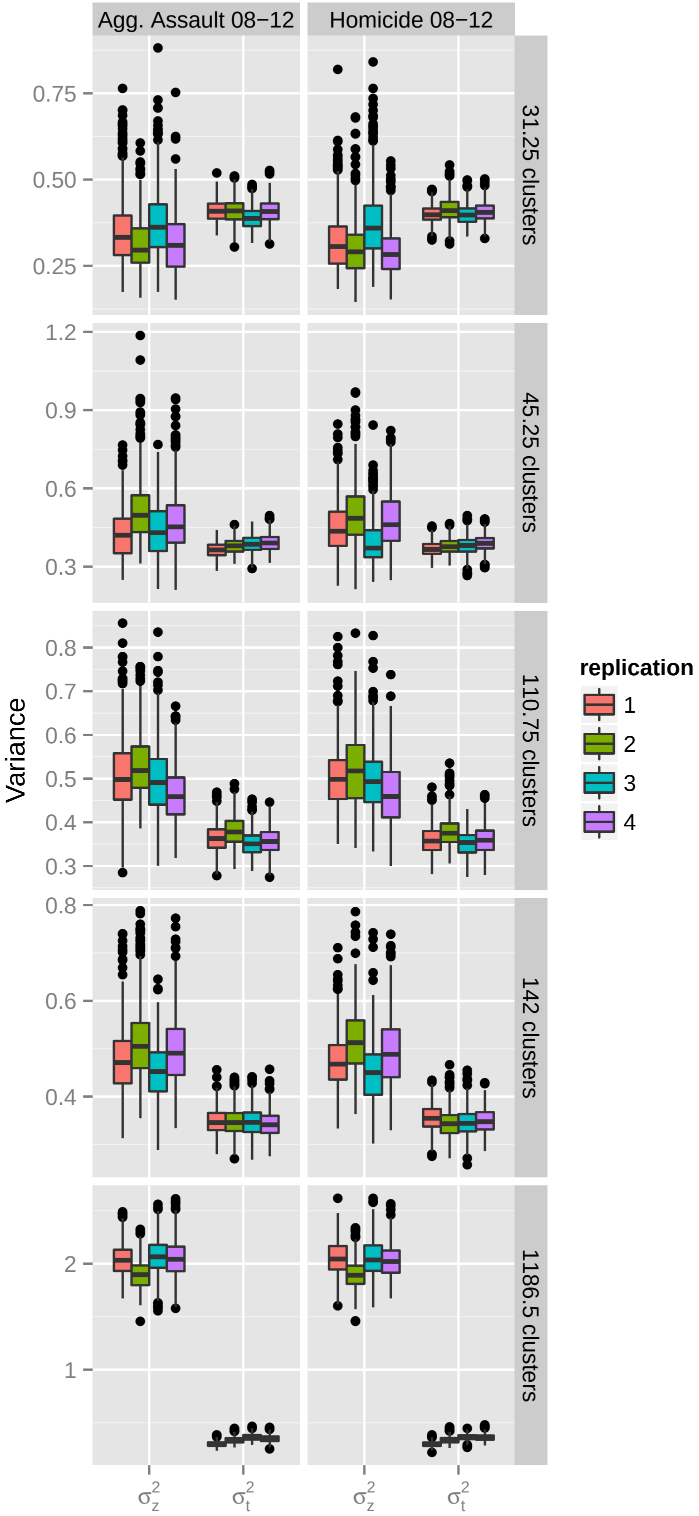

Figure 6: Variance of cluster- and tract-effects for naive crossed-effects models. Rows are labelled by the mean number of clusters among the four ddCRP chains.

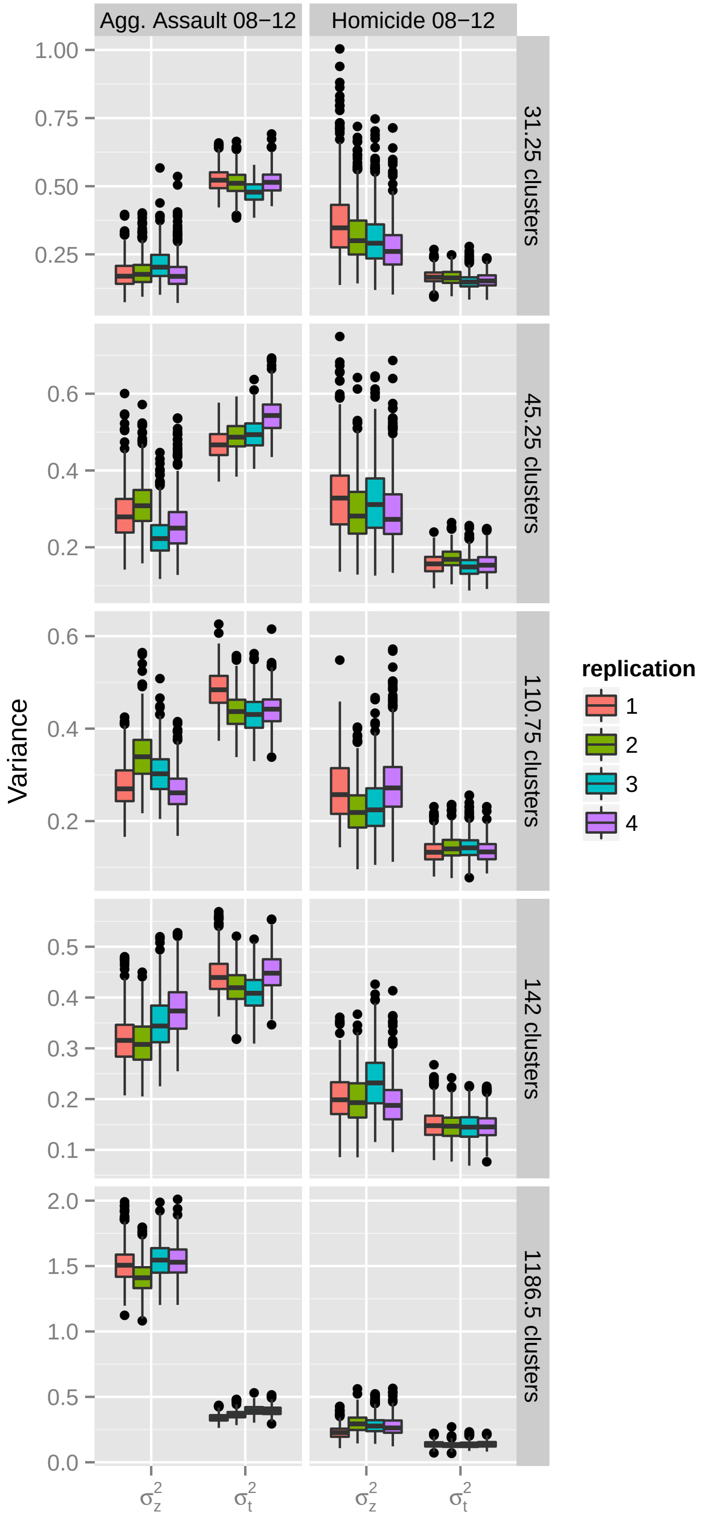

Figure 7: Variance of cluster- and tract-effects for crossed-effects models controlling for block-level race/ethnicity, population, population density, and zoning areas. Rows are labelled by the mean number of clusters among the four ddCRP chains.