Crime Rates Won’t Work: Analyzing Crime for Small Areas Taking into Account More Than Population

John R. Ottensmann
Indiana University-Purdue University Indianapolis
john.ottensmann@gmail.com
urbanpatternsblog.wordpress.com
August 2019

Abstract

Crime rates—numbers of crimes divided by the population living in an area—have problems when used for small areas. Some small areas include substantial nonresidential areas that contribute to the risk of crime, can be the location of crimes, but that have no populations. Negative binomial models to predict counts of the numbers of crimes in small areas are used to incorporate multiple measures of the risk or exposure to crime that cannot be accomplished using crime rates. Population, several measures of employment, and numbers of students in small areas from a transportation planning dataset all contribute to exposure and the prediction of crime in Indianapolis. Because these data are specific to Indianapolis, models using generally available data from the Census Transportation Planning Products and only data from the census of population are evaluated as alternatives. As the initial exposure data were available for the entire metropolitan area, alternative crime rates using these data are estimated and compared with the traditional population-based crime rates for 14 municipalities in the metropolitan area.

Introduction

Levels of crime in areas have traditionally been calculated and presented as crime rates, the numbers of crimes divided by the populations of those areas. This certainly seems warranted and reasonable. After all, it would hardly be appropriate to compare the number of homicides in Wyoming with the number in California given the great differences in their populations. And crime rates do make sense for comparing crime over time for the nation and for comparing crime among states and metropolitan areas.

However, as one looks at smaller and smaller areas, the use of the traditional crime rates becomes more and more problematic. Consider crime rates for blocks within

1 I wish to acknowledge the contributions made in discussions about this research by Tom Stucky and Seth Payton, colleagues at Indiana University-Purdue University Indianapolis.
a city. Numerous blocks in commercial and industrial areas will have no persons residing there—zero populations. Crimes can, however, still occur in those areas. Crime rates cannot be calculated for such blocks—division by 0 is impossible. Some studies have addressed the problem by simply dropping areas with zero populations. But this means excluding some crime from the analysis, crime in areas very different from those that are included.

At least when using very small areas that can have zero populations one comes face to face with the problem since crime rates cannot be calculated. The problem becomes more insidious when using areas that all have some resident population. The crime rates can be calculated for each area, but the validity of those crimes rates can be highly questionable. This is the definitely the case for census tracts, the small areas very frequently used for the analysis of patterns of crime within cities. Census tracts are defined to have an average of around 4,000 people living in them. Their populations can vary, but nearly all census tracts will have significant resident populations. But census tracts are also delineated to include the entire area of a city (in fact, of the entire nation). This means that some tracts will necessarily include territory that is nonresidential—commercial, industrial, parkland, agricultural land, undeveloped vacant land. Crimes can be committed in those areas as well. And those crimes will be added to the crimes committed in the residential portions of the tracts.

A simple hypothetical example will serve to illustrate the problem. Consider two residential areas located next to one another that have the same land area, population, and number of crimes. Looking just at those areas, they would have identical crime rates. One of these residential areas is designated as a census tract. The other area is adjacent to a commercial area having the same land area. This area has significant levels of employment and other visitors. Because that area must be included within some census tract, it has been included as part of a tract that includes the second of the two residential areas. With the activity in the nonresidential area, crimes are likely to be committed there as well. For this example, let us assume that the number of crimes in that nonresidential area is the same as the numbers in each of the two original residential areas. The second census tract would then have twice the number of crimes as the first. Since it has the same population, the crime rate in that tract would be twice the crime rate in the first. That difference results not from any differences in the incidence of crime in the residential areas but rather from the presence of the nonresidential area and the crime committed there.

More crimes occurred in the tract including the nonresidential areas because there were more people in that tract, at least at some times of the day, who would be potential victims of violent crime. And that tract also included more property that could be targets of property crime. What is more relevant for understanding levels of crime is the extent of such risks for crime—the potential victims or targets. In this paper, the
amount of such risk is referred to as the exposure to crime, much as accidents and numbers having a disease are related to the exposure to the risk factor(s).

The role of risk or exposure in relation to crime explains how the (resident) population of an area can be an inadequate measure of that exposure. For a nonresidential area, risk is not related to residential population. Population is zero for a completely nonresidential area and is an underestimate of the risk for areas that are partially residential and partially nonresidential. This also explains why traditional crime rates calculated using residential population are reasonable for very large areas such as states and metropolitan areas but begin to fail as areas become smaller. The larger areas will also have nonresidential areas that contribute to the risk of crime, to exposure. But the ratios of activity in such areas contributing to exposure to the resident populations will be reasonably similar, making population a reasonable measure of relative exposure and making crime rates useful. But as areas become smaller, the ratios of nonresidential areas and activity with their risk of crime to the populations will vary, ultimately greatly. This produces variation in traditional crime rates that render them inappropriate.

This problem with crime rates has been raised for a long time, with examples from the literature discussed in the following section. Addressing the problem, however, has been another matter. Doing so, of course, requires data relating to the factors associated with the risks of crime. But it also requires a methodological approach that can employ multiple measures of risk or exposure, and crime rates are not the answer. It is certainly possible to calculate alternative crime rates using measures of exposure other than resident population, as numbers of studies have done. But a crime rate requires dividing the number of crimes by a single number for the denominator, a single measure of exposure. It is much more likely that multiple measures will be needed to estimate exposure across areas that are both residential and nonresidential. The section following discusses the use of models for the prediction of counts of crime rates that allow the use of multiple measures of exposure. The paper examines the relationship of different measures of exposure to crime, but the different models use some common data, described next. The first sets of models use data on population, employment (of various types) and students, developed for transportation planning, for the prediction of crime in the City of Indianapolis in different sets of small areas. As these data were specific to the Indianapolis area, the next section of the paper presents models using alternative measures of exposure using data that is generally available for areas across the United States. Finally, the exposure data used in the first section was available for the wider metropolitan area. It is used to estimate alternative exposure-based crime rates for 14 municipalities in the area that have reported crime data, comparing these to the traditional population-based crime rates. Finally a series of papers used a dataset with estimates of what were described as 24-hour average
populations or ambient populations for small areas. An appendix discusses issues with that data and the papers that used it.

**Concerns About Crime Rates**

The possibility that residential population may not be the most appropriate basis for calculating crime rates has been considered sporadically for at least half a century. The initial approach has been to identify alternative measures of opportunities for crime or the risk for crime, calculate crime rates using those measures, and compare those with the traditional population-based crime rates.

Boggs (1965) suggested that rates of occurrence for crime in areas within cities should be based upon the environmental opportunities for specific types of crime within those areas. She observed that high crime rates reported for central business districts reflect the small numbers of residents in those areas. Crime occurrence rates were calculated using a variety of opportunity measures ranging from the business-residential land use ratio to miles of streets (for vehicle thefts), with high levels of variation between those rates and rates calculated using population.

Phillips (1973) argued that employment reflects the intensity of activity better than land use, which was used by Boggs. Employment-based crime rates were compared to traditional rates across areas within a city. Harries (1981) defined crime rates as crimes divided by “some measure of risk or opportunity.” He calculated numerous crime rates using a wide variety of measures associated with the likelihood of crime, though it is questionable as to whether all of these can be considered to be measures of risk.

An alternative approach to the question of the effect of using population for crime rates was first offered by Gibbs and Erickson (1976). They argued that traditional crime rates in central cities were in part higher because their populations are underestimates of the number of persons spending time in those areas and thus of the potential number of offenders and victims. They illustrated the effect by showing that traditional central city crime rates were related to the ratio of central city to metropolitan area population. Stafford and Gibbs (1980) extended the analysis to more cities and added the proportion of retail sales in the cities a measure of city dominance within the larger area.

More recently, Stults and Hasbrouck (2015) did the most comprehensive analysis of this effect on city crime rates, using census data on inbound and outbound commuting for all cities with populations of 100,000 or more. They first used populations adjusted by this commuting to compute alternative crime rates. These are compared with the standard rates using resident population and significant changes in the rankings are seen for some cities using the adjusted populations. They also used population change due to the in- and outbound commuting as a predictor in
multivariate models predicting the traditional crime rates. The found not only was this measure significant in the models but its inclusion produced changes in the coefficients of other predictors of crime in the models.

The effect of using population for computing crime rates for cities was noted in a recent report on improving crime statistics produced by the National Academies (Lauritsen and Cork 1916):

Crime rankings simply list cities in order of their FBI reported crime rates, typically by using an index of crimes or index of violent crimes. Cities may score near the top of such rankings because they actually have higher crime rates, or because they are more likely to record their crimes, have higher crime-reporting rates by victims, or have cities that are relatively small in proportion to their surrounding metropolitan areas, thus capturing more incidents in the numerator of their crime rates without additional population in the denominator [emphasis added].

The routine activities theory of crime proposed by Cohen and Felson (1979) provides a basis for considering these effects. They focus on “suitable targets” along with offenders and guardianship as the basis for criminal activity. The question raised by the critiques of traditional crime rates is the extent to which population is an inadequate measure of those targets. Cohen and Felson’s primary focus was on crime trends at the national level and changes in routine activity patterns, including time spent outside the home. They did not give equivalent attention to variations in targets across areas and the effects this could have on population-based crime rates for those areas.

Interest in the problems with traditional crime rates for cities and smaller areas has increased with the publication of a series of papers by Andresen (Andresen 2006, 2010, 2011; Andresen and Jenion 2010). He calculated crime rates for small areas within Vancouver, British Columbia and for municipalities within the metropolitan area using an alternative measure termed “ambient population.” Ambient population is a term used for the 24-hour average populations estimated for small areas for the entire world by the LandScan project at Oak Ridge National Laboratories (2019). The appendix to this paper discusses the LandScan data, questioning its suitability for such application, and identifies additional issues with the Andresen papers.

Andresen’s work has led to numerous papers addressing these questions. Felson and Boivin (2015) have some of the best data for examining the issues, data on population flows to census tracts for four trip purposes—work, shop, education, and recreation. These data are from a transportation planning survey for a “major city in eastern Canada” (obviously Montreal). They used basic descriptive statistics, including simple correlations, to relate these population flows to the numbers of different types of crime. They presented the results of quadratic regressions of each type of flow on
numbers of violent and property crimes. What stands out most, however, is what they
did not do. While they included some characteristics of the populations of the tracts in
the correlation matrix with crime and the population flows in their analysis, they did
not include the total tract populations. The quadratic regression models of flows on
crimes did not include the population characteristics as controls in more complete
multivariate models.

Boivin (2013) used resident population and employment in census tracts to
predict the number of crimes along with additional control variables typically used in
the prediction of crime. Confusingly, tract employment was called the ambient
population, adapting that term from Andresen but using it to mean something very
different. This may have been the first of a number of papers in the criminological
literature using ambient population in various ways with the only common meaning
being that ambient population was something other than the resident population. (The
appendix includes further discussion of this.)

Subsequent papers have made use of the newer, massive bodies of data
(sometimes referred to as “big data”) that have resulted from the use of new
technologies. Malleson and Andresen (2015) used geolocated Twitter message data as an
alternative to population for the calculation of crime rates and the identification of
crime hotspots. Malleson and Andresen (2016) extended this by using additional
measures including mobile phone counts for areas and the workday populations
provided in the 2011 United Kingdom census. These are all compared with resident
population by considering the rank-order correlations between the number of crimes
and the various measures. Hipp, et al. (2019) used Twitter data at different times of day
to predict crimes at those times for city blocks.

Count Models and Crime Exposure

The problem with using crime rates if factors in addition to population affect
levels of crime is that rates are counts of crime divided by some measure that the counts
are assumed to be proportional to. Computing a rate requires the single value for that
measure, the denominator of the crime rate. But if levels of crime are influenced by
multiple factors, how can one compute a rate? If some combination of those factors can
be assumed to be the quantity to which crime is proportional, then a rate could be
calculated. But how would one figure out the combination?

The solution to the problem comes with using models predicting the numbers of
crimes, not crime rates. These models are estimated using Poisson regression or variants
such as negative binomial regression as the distributions of counts does not meet the
requirements for ordinary least squares regression. Arguments for and examples of
using count models in the prediction of crime include Osgood (2000) and Osgood and
Chambers (2000).
To see how count models can be used to address the problem with crime rates, we begin by looking at how the models deal with the simpler situation in which the level of the counts are assumed to be affected by a single measure. This quantity to which counts are assumed to be proportional is called the exposure. The use of this term may have arisen from the use of count models in epidemiology where the incidence of a disease or of accidents might be assumed to be proportional to the length of time an individual had been exposed to the risk factor or for aggregate data, counts for groups of persons, the number of persons in each group exposed to the risk factor. Winkelmann (2008) says that the basic assumption of exposure in count models is proportionality, that the counts are proportional to the exposure.

We will consider how exposure is addressed in a Poisson model. The approach is comparable in negative binomial models, but these are more complex to address the problem of over dispersion. The Poisson model incorporating exposure can be expressed as follows:

\[ c_j = \exp (x_j \beta + O_j) \]

where \( c_j \) is the count of crimes in area \( j \), \( x_j \) is a vector of predictors of crime for area \( j \), \( \beta \) is a vector of regression coefficients for those predictors to be estimated, and \( O_j \) is the offset for area \( j \). The offset is the log of the exposure variable, \( E_j \). Substituting the log of exposure into the model and rearranging gives

\[ c_j = \exp (x_j \beta) E_j \]

showing that the count is proportional to the value of the exposure variable and that exposure is incorporated in the model by including the log of the exposure variable as the offset. Note that the log of the exposure variable, the offset, is included as just another term in the model with the exception that its coefficient is assumed to have the value of 1. Cameron and Trivedi (1998) have suggested testing whether the count can be considered to be proportional to the exposure variable by including the log of exposure as a regular independent variable, estimating its regression coefficient, and doing the hypothesis test of whether the estimated coefficient is equal to 1. This was done using population as the potential measure of exposure in each of the models presented in this paper. The null hypothesis was that the coefficient was equal to 1 and that the count was proportional to the population, This was rejected in every case, producing the conclusion that crime was not simply proportional to population.

Error in the measure of exposure in the model can result in biased estimates of the regression coefficients for other variables in the model. For error in an independent
variable in a linear model with the classical errors-in-variable assumption, the result can be inconsistent, biased estimates of the other regression coefficients (Wooldridge 2006). Cameron and Trivedi (1998) have argued that the problem of error in the exposure variable is more analogous to error in the dependent variable in a linear model. If the error is correlated with the error term, the estimates will be inconsistent and biased.

The error associated with using only population as the exposure variable can be addressed by incorporating multiple measures of exposure in the Poisson or other count model. Cameron and Trivedi (1998) have proposed that exposure can be considered to be an unobserved variable with a set of observed variables being used as proxies for exposure. Let the offset be a function of a set of proxy variables such that

\[ \hat{O}_j = z_j \gamma \]

where \( \hat{O}_j \)-hat is the predicted offset for area \( j \), \( z_j \) is a vector of proxy variables for area \( j \), and \( \gamma \) is a vector of regression coefficients to be estimated. Substituting this for the offset in the equation for the Poisson model gives

\[ c_j = \exp(x_j \beta + z_j \gamma) \]

Therefore the proxy variables for exposure can be included as normal independent variables in the model with regression parameters to be estimated, with no constraints on the estimation of those parameters. A robust example of the use of a large number of proxies for exposure has been provided by Dionne et al. (1995).

Since the proxies for exposure are any variables that can predict the offset, the functional form for the inclusion of the variables is not necessarily specified. However, if the variables to be included as proxies can be considered to be measures of portions of the exposure such that the number of crimes would be partially proportional to those variables, it would be logical to include the variables in log form in the model. This will mostly be done here.

**Data for All Analyses**

The models predicting crime to be presented here use various measures of exposure. The sources of data and the variables used for each model will be described in the sections with those models. This section describes the data that are used in all of the models, the crime data and the population data from the American Community Survey.

Given the different sources of the data, analyses use 2 different sets of areal units. Census tracts, used in the reporting of data by the Census Bureau, are delineated to
have populations that average around 4,000 persons, meaning that they have smaller areas where population densities are higher and are larger in less dense areas, typically areas farther from the center of the city. A total of 198 census tracts are used that have 90 percent or more of their areas within the area being included in the analyses—Indianapolis-Marion County with the exception of the 4 excluded municipalities of Beech Grove, Lawrence, Southport, and Speedway.

The second set of areas are the traffic analysis zones (TAZs) used for transportation planning. These areas are much smaller than census tracts. A total of 1,171 TAZs are mainly within the area being included in the main analysis. In addition to being much smaller than the census tracts, the basis for their delineation is quite different, not completely related to population. Instead, the TAZs have smaller sizes in areas with higher densities of trip origins and destination. The smallest TAZs are primarily in and near the central business district. This means that significant numbers of TAZs have zero residential population, meaning that crime rates could not even be calculated for those areas. Like census tracts, they tend to become larger as one moves out from the center as the density of trip origins and destinations declines. But smaller TAZs can be found in outlying areas with high concentrations of commercial activity.

Crime Data

The counts of the numbers of crimes have been derived from data for all of the Part I criminal offenses reported by the Indianapolis Metropolitan Police Department for the Uniform Crime Reports. The target year for the analysis is 2010. However crimes for 2009, 2010, and 2011 are included to reduce the random variation in numbers of crimes for very small areas. The data were provided as geocoded point locations of individual crimes. These data could then be aggregated to obtain counts for the small areas used in the analysis, the census tracts and the traffic analysis zones (TAZs).

The jurisdiction of the Indianapolis Metropolitan Police Department is all of the consolidated City of Indianapolis and Marion County. The county does included 4 small excluded municipalities that retained independence after the city-county consolidation. These units have their own small law enforcement agencies as well. Criminal offenses reported by the Indianapolis Metropolitan Police Department represent only a portion of all crimes in those areas. These areas have been excluded from the analysis.

An issue for most analyses of crime for small areas is the presence of spatial autocorrelation, the correlations of levels of crime in an area with those in nearby areas. This can result in the estimates of the standard errors of the regression coefficients in a model being too small. To (at least partially) correct for this, a spatial lag variable is included in each of the models. The spatial lag variable is the count of the number of crimes in the areas (census tracts or TAZs) contiguous to each area. (Queen contiguity was used, meaning areas were contiguous if they at least shared a vertex.) In the models
that were estimated, the natural log form of the spatial lag of crime generally performed better and was used.

A note on the area of Indianapolis-Marion County included in the analysis. The original City of Indianapolis prior to consolidation in 1970 included a much smaller area in the center of the county. Like most large cities in the Northeast and Midwest, the area within those city limits had largely been developed prior to World War II, with most new development since that time occurring in the suburban areas outside. With city-county consolidation, the territory of Indianapolis was expanded to include many of those newer suburban areas along with significant amounts of vacant land that would allow for continued new development in more recent decades. Thus the area included in the present analysis includes a much broader range of physical and social environments than are found in many analyses of crime in large cities.

Population Data

The population data are for census tracts and come from the American Community Survey (ACS). The ACS is conducted annually and obtains data for a sample of the population. To get a sufficient sample size for smaller areas such as census tracts, the ACS reports results for the totals for 5-year periods. With the target year for the analyses being 2010, the ACS data for the period 2008-2012 was used, the period centered on that target year. The data were downloaded from the Census website using American FactFinder (U.S. Bureau of the Census 2014a). Estimates for the TAZs were obtained by apportioning the tract counts in proportion to the area of each census tract within each TAZ.

The analyses include 3 characteristics of the population computed directly from the ACS data: the proportion of the population African-American, the proportion of the population Hispanic, and the proportion of males aged 15 to 24. The fourth measure is an index of disadvantage developed using multiple highly correlated variables. The variables included were the proportion of households with children with a female householder only, the proportion of the workforce unemployed, median household income, and the proportion of persons with incomes below the poverty level. The index was derived using principal components analysis. The correlations of the variables with the resulting index ranged from 0.82 to 0.91 for the census tracts and 0.77 to 0.87 for the TAZs. The index accounted for 77 percent of the variation in these variables for the census tracts and 73 percent for the TAZs. Other variable were considered for the models but were excluded due to problems with multicollinearity.
Predicting Crime Using Data from Transportation Planning for Exposure

Transportation planners necessarily use a wide variety of data to predict travel behavior. The data used for the measures of exposure were assembled by the Indianapolis Metropolitan Planning Organization (MPO), the body responsible for transportation planning for Indianapolis and the surrounding area. (The federal government mandates the creation of such organizations covering urban areas beyond the major cities for all Urbanized Areas, areas, which have populations over 50,000.) These data will be referred to as the MPO data.

The MPO data are for the year 2010. The data are for the traffic analysis zones (TAZs) covering Indianapolis-Marion County and 8 surrounding counties. As mentioned above, the analysis area within Indianapolis-Marion County includes 1,171 TAZs. The MPO dataset includes some data on the population residing in each TAZ (and not surprisingly, their number of automobiles). More important for the present analysis, it includes data on the numbers of persons employed in each of the TAZs and the number of students in educational institutions at all levels. The employment data are very detailed, broken down by the 21 2-digit categories in the North American Industrial Classification System, NAICS.

Four exposure variables are used in the analysis. The population residing in each TAZ (the base for traditional crime rates) is naturally the first. The second is the total employment in each area. Those employees are potential targets of crime, as are the various forms of property associated with the presence of that employment and those establishments. In addition, the presence of these activities will draw additional people into the areas, increasing the potential for criminal activity. This latter effect serves as the basis for the third exposure measure. Most retail establishments and many types of service businesses (think restaurants as an example) will draw much larger numbers of people to their places of businesses. This necessarily must be the case or they would not stay in business! The number of employees working in such establishments serves as a reasonable measure of the extent of such activity. Therefore, the total employment in the retail and service industry categories is the third exposure variable. Finally, students represent another group that could be the target of criminal activity, and their numbers are included as the final measure of exposure.

This section presents 2 sets of models using these measures of exposure in the prediction of crime. The first uses the TAZs as the units of analysis, with the exposure variables directly from the MPO data. The population characteristics are those values estimated from the tract data as explained above. Census tracts are the units for the second set of analyses. For these, the MPO data have been aggregated to the tracts. Where TAZs are split by tract boundaries, the TAZ population, employment, and students are allocated to the tracts in proportion to the area of the TAZ in each tract. Such splitting is less frequent than in many situations in which data are estimated for a
different set of areas as many of the TAZ boundaries coincide with tract boundaries. Therefore, the error introduced by this process is less than it might have been.

Analysis Using Traffic Analysis Zones

The first set of models predicting the number of crimes uses the MPO exposure data with their original areal units, the traffic analysis zones (TAZs). The counts of the Part I criminal offenses for 2009 to 2011 come from the Indianapolis Metropolitan Police Department as described above. Data for the population characteristics from the American Community Survey were obtained for census tracts and values were estimated for the TAZs.

Five models predicting the total number of crimes in the TAZs have been estimated, with the results presented in Table 1. Negative binomial regression was used for the estimation as the models using Poisson regression showed significant overdispersion. The first model estimated, the base model, includes the 4 population characteristics as predictors—proportions African American, Hispanic, and males aged 15-24, and the disadvantage index. The (natural) log of the crime lag variable for the contiguous tracts is also included to address the problem of spatial autocorrelation. The model is statistically significant as shown by the likelihood ratio test versus the null model. The only predictors statistically significant are proportion Hispanic and the spatial lag variable. This model does not include any of the measures of exposure.

The subsequent models each successively add one additional variable to account for the variation in the exposure across the tracts. Each of these variables is the log of the exposure variable. The first exposure variable to be added is population, with its log included in the model. The model is not only statistically significant as shown by the likelihood ratio test versus the null model, it represents a statistically significant improvement versus the base model, shown by that likelihood ratio test. And the log of population is statistically significant. Adding the population exposure variable produces changes to the coefficients for the population characteristics, making the proportion of males aged 15 to 24 statistically significant along with proportion Hispanic.

The following models add exposure variables beyond the traditional population measure of exposure. Next is total employment. This produces a statistically significant improvement over the model including only population as shown by that likelihood ratio test. Log of total employment is also statistically significant. Once again, the addition results in changes to the coefficients for the population characteristics and their significance. The coefficient for the proportion of males aged 15 to 24 falls and is no longer statistically significant. On the other hand, the coefficient for the disadvantage index increases more than 3 times and becomes statistically significant. The proportion African American is also statistically significant.
Table 1. Predicting Crime Using the MPO Exposure Variables for Traffic Analysis Zones (standard errors in parentheses).

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Base Model</th>
<th>Total Population Added</th>
<th>Total Employment Added</th>
<th>Retail and Service Employment Added</th>
<th>Students Added</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion African American</td>
<td>0.153</td>
<td>0.154</td>
<td>0.247 **</td>
<td>0.250 *</td>
<td>0.262 *</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(0.130)</td>
<td>(0.108)</td>
<td>(0.104)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Proportion Hispanic</td>
<td>1.239 **</td>
<td>1.417 ***</td>
<td>1.516 ***</td>
<td>1.508 ***</td>
<td>1.497 ***</td>
</tr>
<tr>
<td></td>
<td>(0.401)</td>
<td>(0.381)</td>
<td>(0.333)</td>
<td>(0.317)</td>
<td>(0.315)</td>
</tr>
<tr>
<td>Proportion Males Aged 15-24</td>
<td>-0.744</td>
<td>2.067 **</td>
<td>0.888</td>
<td>0.928</td>
<td>1.090</td>
</tr>
<tr>
<td></td>
<td>(0.712)</td>
<td>(0.659)</td>
<td>(0.581)</td>
<td>(0.559)</td>
<td>(0.559)</td>
</tr>
<tr>
<td>Disadvantage Index</td>
<td>0.035</td>
<td>0.030</td>
<td>0.097 ***</td>
<td>0.102 ***</td>
<td>0.103 ***</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.021)</td>
<td>(0.018)</td>
<td>(0.017)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Log Crime Lag Variable</td>
<td>0.814 ***</td>
<td>0.686 ***</td>
<td>0.416 ***</td>
<td>0.365 ***</td>
<td>0.362 ***</td>
</tr>
<tr>
<td></td>
<td>(0.043)</td>
<td>(0.042)</td>
<td>(0.038)</td>
<td>(0.037)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>Log Population</td>
<td>—</td>
<td>0.161 ***</td>
<td>0.227 ***</td>
<td>0.229 ***</td>
<td>0.222 ***</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>(0.010)</td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Log Total Employment</td>
<td>—</td>
<td>—</td>
<td>0.310 ***</td>
<td>0.151 ***</td>
<td>0.135 ***</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>(0.014)</td>
<td>(0.015)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Log Retail and Service</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.184 ***</td>
<td>0.192 ***</td>
</tr>
<tr>
<td>Employment</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>(0.016)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Log Students</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.032 ***</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.828 ***</td>
<td>0.326</td>
<td>-0.250</td>
<td>0.151</td>
<td>-0.206</td>
</tr>
<tr>
<td></td>
<td>(0.231)</td>
<td>(0.220)</td>
<td>(0.040)</td>
<td>(0.185)</td>
<td>(0.186)</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-6,994</td>
<td>-6,892</td>
<td>-6,688</td>
<td>-6,632</td>
<td>-6,627</td>
</tr>
<tr>
<td>Likelihood Ratio Test vs Null</td>
<td>474.0(5)</td>
<td>676.9(6)</td>
<td>1,085.7(7)</td>
<td>1,197(8)</td>
<td>1,207(9)</td>
</tr>
<tr>
<td>Model - Chi²(df)</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Likelihood Ratio Test vs Prior</td>
<td>203.9(1)</td>
<td>408.7(1)</td>
<td>111.0(1)</td>
<td>10.8(1)</td>
<td>***</td>
</tr>
<tr>
<td>Model - Chi²(df)</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

*Significant at the 0.05 level   **Significant at the 0.01 level   ***Significant at the 0.001 level
The procedure continues, adding the retail and service employment exposure variable and then the students exposure variable. Each produces a statistically significant improvement to the model and each coefficient is statistically significant. It is noteworthy, however, that the coefficients for the population characteristics are not dramatically changed by these additions, and the pattern of significance for the coefficients remains unchanged. It was the addition of the total employment variable along with the population variable as measures of exposure that made the difference with the population characteristics.

The results in Table 1 clearly show the benefit of using the multiple measures of exposure. Each additional measure of exposure produces a statistically significant improvement to the model. And having at least total employment along with population as a measure of exposure proved important for producing stable estimates of the coefficients of the population characteristics.

Further analyses were undertaken predicting the numbers of the different types of the Part I crimes. (The results are not shown here.) Only some of the exposure variables were significant for criminal homicides and rapes, likely due to the smaller numbers of these crimes.

Exposure for robberies was especially interesting. The coefficient for population was the same as for all crimes. The coefficient for total employment was very small and not significant. But the coefficient for retail and service employment was far higher than for population or for either of the employment coefficients in the model for all crimes. Criminals rob convenience stores and restaurants; they do not rob offices and factories.

Aggravated assaults showed an exposure pattern similar to all crimes with one notable exception. The coefficient for students was over twice the magnitude of that coefficient when predicting all crimes. Students are a much more important component of exposure for aggravated assaults.

For burglaries, the coefficient for population was much higher than for all crimes, presumably reflecting the importance of residential burglaries. The coefficient for total employment was very small, while retail and service employment retains a higher coefficient. Presumably the latter types of establishments are more likely targets for burglaries. Finally, the coefficient for students was extremely small, suggesting little or no exposure to burglary there.

The pattern for larcenies is more similar to the pattern for all crimes, with a somewhat lower coefficient for population and a somewhat higher coefficient for retail and service employment. This may reflect the incidence of theft in retail establishments. Vehicles thefts, on the other hand, had a slightly higher coefficient for population and slightly lower coefficients for both employment variables. The coefficient for students was extremely small.
Analysis Using Census Tracts

Research on crime typically uses areal units such as blocks, block groups, and census tracts, not traffic analysis zones. The choice of areal units will affect the results obtained from an analysis. Numerous authors have written about the modifiable areal unit problem (MAUP), that model results will vary with the delineation of the areas used for the data and analysis (see, e.g., Openshaw 1984a, 1984b; Fotheringham and Wong 1991). It is uncommon, however, for analyses using different sets of areal units to be undertaken and presented in the course of research addressing issues other than the MAUP, as will be done here.

Crime exposure presents a rather unique situation where the scale of analysis and the MAUP is especially significant and worthy of attention. For very large areas, such as states, population is a very reasonable measure to use for exposure to crime, as it is likely to be very highly correlated with other measures of exposure. For very small areas, on the other hand, population is a partial and inadequate measure, failing completely for small areas in which the population is zero and in which crime can still occur. Therefore, considering the relative performance of the measures of exposure using different units of analysis can be important.

To do a comparison of the effect of the choice of areal units on the effect of measures of exposure, models comparable to those estimated for the TAZs have been estimated using census tracts as the units of analysis. The MPO data for the exposure measures have been estimated for the census tracts. The population characteristics are as originally obtained from the American Community Survey. Table 2 presents the results for the models predicting crime for census tracts for the base model using only the population characteristics and for the full model including all 4 of the exposure variables. Results for the comparable models estimated using the TAZ data are shown for comparison. The census tracts are much larger than the TAZs. The study area encompasses 198 census tracts and 1,171 TAZs, making the census tracts over 5 times larger on average than the TAZs.

The census tract model that included all of the exposure variables was a statistically significant improvement over the base model. All 4 of the exposure variables were statistically significant. This model had substantial differences from the model estimated using the TAZ data. Considering first the exposure variables, the coefficient for the log of population was more than twice as large for the census tract model compared to the TAZ model. The other 3 exposure variables had smaller estimated coefficients for the census tract model. While all were statistically significant at the 0.05 level, the levels of significance were lower than for the TAZ model for 2 of the measures.

These differences reflect at least in part the differences in the sizes and delineations of the census tracts and the TAZs. Census tracts are used for reporting
Table 2. Predicting Crime Using the MPO Exposure Variables for Census Tracts versus Traffic Analysis Zones (standard errors in parentheses).

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Census Tracts</th>
<th>Traffic Analysis Zones (TAZs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base Model</td>
<td>All Exposure Variables</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion African American</td>
<td>-0.154</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>(0.157)</td>
<td>(0.103)</td>
</tr>
<tr>
<td>Proportion Hispanic</td>
<td>1.097 *</td>
<td>0.389</td>
</tr>
<tr>
<td></td>
<td>(0.453)</td>
<td>(0.315)</td>
</tr>
<tr>
<td>Proportion Males Aged 15-24</td>
<td>2.301 *</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>(1.012)</td>
<td>(0.708)</td>
</tr>
<tr>
<td>Disadvantage Index</td>
<td>0.087 **</td>
<td>0.213 ***</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Log Crime Lag Variable</td>
<td>0.532 ***</td>
<td>0.395 ***</td>
</tr>
<tr>
<td></td>
<td>(0.110)</td>
<td>(0.074)</td>
</tr>
<tr>
<td>Log Population</td>
<td>—</td>
<td>0.494 ***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.055)</td>
</tr>
<tr>
<td>Log Total Employment</td>
<td>—</td>
<td>0.078 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.032)</td>
</tr>
<tr>
<td>Log Retail and Service Employment</td>
<td>—</td>
<td>0.126 ***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.031)</td>
</tr>
<tr>
<td>Log Students</td>
<td>—</td>
<td>0.009 **</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.007)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.984 ***</td>
<td>-1.349 *</td>
</tr>
<tr>
<td></td>
<td>(0.767)</td>
<td>(0.673)</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1.468</td>
<td>-1.385</td>
</tr>
<tr>
<td>Likelihood Ratio Test vs Null Model - Chi²(df)</td>
<td>81.5(5) ***</td>
<td>246.9(9) ***</td>
</tr>
<tr>
<td>Likelihood Ratio Test vs Base Model - Chi²(df)</td>
<td>—</td>
<td>165.5(4) ***</td>
</tr>
</tbody>
</table>

*Significant at 0.05 level  **Significant at 0.01 level  ***Significant at 0.001 level
population statistics and are delineated to include an average population of around 4 thousand persons. The minimum tract population for this dataset was 1,125. TAZs, on the other hand, are specifically drawn to identify concentrations of trip origins and destinations which can include concentrations of employment. Of the 1,171 TAZs in the dataset, 145 had zero population. So the TAZs vary more than the census tracts in terms of employment, likely contributing to the increased importance of employment as measures of exposure for the TAZs.

Comparing the results for the population characteristics also shows the significance of the MAUP. Proportions African-American and Hispanic are statistically significant in the full TAZ model. They are no longer significant in the census tract model. This presumably reflects the greater racial and ethnic heterogeneity within tracts compared to the TAZs.

The larger coefficient for the total population exposure variable should not, however, lead to the conclusion that population is a sufficient measure of exposure for analyses conducted at the census tract level. With the TAZ analyses reported in Table 1, we saw how adding exposure variables, at least through the addition of total employment, produced changes in the regression coefficients and the significance for the population variables. The same occurred for the analyses using the tract data. In the model including population as the only exposure variable (not shown), proportion males aged 15 to 24 was statistically significant as it was in the base model. The same was true for the model with population only using the TAZ data in Table 1. However, when total employment was added as an additional exposure variable, the coefficient for proportion males aged 15 to 24 declined and the variable becomes not significant. The same thing happened in the TAZ analysis.

Incomplete specification of exposure can produce results that are different from those obtained when exposure is more completely specified in the model. Using population as the only measure of exposure appears to be an incomplete specification of exposure at both the TAZ and the census tract levels of analysis.

Predicting Crime Using Generally Available Data for Exposure

The Indianapolis MPO data proved to be very effective in accounting for exposure in predicting crime. While every Urbanized Area will be included in a metropolitan planning organization, not all of those MPOs will have assembled comparable data. Even if they have, it may not be available. Furthermore, for research on crime involving multiple urban areas, data from different MPOs, even if available, will not necessarily be comparable.

This section considers two approaches for addressing exposure in models predicting crime beyond the basic residential population, models that make use of generally available data. The first—and more data intensive—uses data on workers by
place of employment from the Census Transportation Planning Products. The second approach, which may be more accessible to researchers looking at crime, makes estimates of additional exposure to crime from the standard population data from the census and American Community Survey.

*Census Transportation Planning Products Data*

The Census Transportation Planning Products (CTPP) (American Association of State Highway and Transportation Officials 2019; U.S. Federal Highway Administration 2019) are special tabulations of American Community Survey data developed specifically for transportation planning. These include data for census tracts and/or traffic analysis zones (TAZs) for population by place of residence, for workers by place of work, and worker flows between home and work. The second dataset by place of work provides employment data suitable for extending exposure to crime.

The CTPP data used here have been downloaded from the American Association of State Highway and Transportation Officials (AASHTO) website (2014). The TAZs used for these data are aggregations of the TAZs used for the MPO data, 643 in Marion County for the CTPP data versus 1,261 for the MPO TAZs. Boundary files for the CTPP TAZs, needed to manipulate the data, were downloaded from the census website (U.S. Bureau of the Census 2014b).

The CTPP workplace data include the numbers of workers classified by industry. The industry classification is less detailed and not comparable to that used for the MPO data. To have comparability with the previous MPO analysis, only the total employment is used as a measure of exposure. It would be possible to incorporate numbers of workers in specific industries as further exposure measures, but this has not been done here.

Once again, the data are for TAZs (though different from the MPO TAZs) while the population characteristics from the American Community Survey are for census tracts. Given that most researchers examining crime are likely to be using one of the sets of small areas delineated for the census, the choice was made to do the analysis using census tracts. Thus the numbers of workers in the TAZs was used to estimate the numbers for the census tracts, apportioning the TAZ workers according to the areas of each zone falling in each census tract.

Negative binomial regression was again used to estimate models incorporating the CTPP data. As before, the models predict the total number of criminal offenses. The same population characteristics and the log of the spatial lag of crime variable are included as in the prior models. The log of the CTPP workers in the census tracts is included as the second measure of exposure in addition to population in the full model.

The question here is the extent to which including this measure of exposure improves the prediction over using only population as the exposure measure. Table 3
presents the results for the model including only population and the model also including CTPP workers. For comparison with the earlier results, the results for the model (for census tracts) using MPO total employment are also included. The model including CTPP workers performs better than the model using only population as the measure of exposure based on the likelihood ratio test. The regression coefficient for the log of CTPP workers is statistically significant.

Comparing the models using CTPP workers and MPO total employment shows that the results are very similar. The regression coefficient for the CTPP workers is 0.175 and the coefficient for MPO employment is 0.181. The coefficients for the other exposure measure in the models, the log of population, are nearly identical, 0.595 and 0.593. The regression coefficients for the other variables in the models are not substantially changed and show the same patterns of statistical significance.

Because one of the models is not nested in the other, a likelihood ratio test is not appropriate for comparing the performance of the models. As an alternative, the Akaike Information Criterion (AIC) is used for this purpose. The value of the AIC is slightly lower for the model using MPO total employment—2,798 to 2,813—indicating that model does slightly better. But the difference is small. Using CTPP workers as a second measure of exposure does nearly as well as MPO total employment, making this a viable option for research in all urban areas.

Using Only Population Data for Estimating Additional Exposure

Researchers studying crime may choose not to pursue the approach of using the CTPP data to extend exposure to crime beyond residential population. This section presents a method that uses only the traditional population data typically employed in research on crime, in this case, data for census tracts from the census and the American Community Survey.

The rationale for the approach presented here is derived from urban economics. The starting point is the traditional monocentric model which begins with the assumptions that employment is located at the center of the urban area and that workers seek to minimize the cost of transportation to work while maximizing the living space available (Muth 1969; Mills 1972). This produces a negative exponential decline of population density with distance from the center. Cities have, of course, evolved to have multiple centers of employment, and studies have shown that population densities also decline with distances from those centers (Griffith 1981; Gordon, Richardson, and Wong 1986; McMillen and MacDonald 1998). Population densities have also been shown to decline with decreasing accessibility to all employment (Song 1994; 1996).

The key here is that population densities show regular patterns with respect to employment locations and that densities in nearby areas therefore tend to be similar.
Table 3. Predicting Crime Using Census Transportation Planning Package Employment and Tract Population Data Only (standard errors in parentheses).

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Population as Only Measure of Exposure</th>
<th>CTPP Workers Included</th>
<th>Contiguous Tract Max Pop Density Included</th>
<th>Total MPO Employment Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion African American</td>
<td>-0.156 *(0.126)</td>
<td>-0.081 *(0.109)</td>
<td>-0.184 *(0.122)</td>
<td>-0.082 *(0.105)</td>
</tr>
<tr>
<td>Proportion Hispanic</td>
<td>0.113 *(0.377)</td>
<td>0.343 *(0.331)</td>
<td>0.196 *(0.366)</td>
<td>0.420 *(0.321)</td>
</tr>
<tr>
<td>Proportion Males Aged 15-24</td>
<td>1.734 ***(0.809)</td>
<td>0.124 *(0.756)</td>
<td>1.681 ***(0.797)</td>
<td>0.159 *(0.725)</td>
</tr>
<tr>
<td>Disadvantage Index</td>
<td>0.193 ***(0.025)</td>
<td>0.228 ***(0.022)</td>
<td>0.209 ***(0.025)</td>
<td>0.233 ***(0.021)</td>
</tr>
<tr>
<td>Log Crime Lag Variable</td>
<td>0.485 ***(0.088)</td>
<td>0.372 ***(0.078)</td>
<td>0.416 ***(0.087)</td>
<td>0.333 ***(0.076)</td>
</tr>
<tr>
<td>Log Census Tract Population</td>
<td>0.690 ***(0.066)</td>
<td>0.595 ***(0.058)</td>
<td>0.720 ***(0.065)</td>
<td>0.593 ***(0.056)</td>
</tr>
<tr>
<td>Log CTPP Workers</td>
<td>—</td>
<td>0.175 ***(0.022)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Maximum Density Contiguous Tracts Minus Tract Density (in thousands)</td>
<td>—</td>
<td>—</td>
<td>0.058 ***(0.016)</td>
<td>—</td>
</tr>
<tr>
<td>Log Total MPO Employment</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.181 ***(0.020)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.226 ***(0.802)</td>
<td>-1.872 ***(0.692)</td>
<td>-2.121 ***(0.773)</td>
<td>-1.644 *(0.665)</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1,424</td>
<td>-1,398</td>
<td>-1,418</td>
<td>-1,390</td>
</tr>
<tr>
<td>LR Test vs Null Model - Chi²(df)</td>
<td>168.9(6) ***(4)</td>
<td>222.8(7) ***(4)</td>
<td>1841.4(7) ***(4)</td>
<td>237.5(7) ***(4)</td>
</tr>
<tr>
<td>LR Test vs Pop Only Model - Chi²(df)</td>
<td>—</td>
<td>53.8(1) ***(4)</td>
<td>12.5(1) ***(4)</td>
<td>68.691 ***(4)</td>
</tr>
<tr>
<td>Akaike Information Criterion - AIC</td>
<td>2,865</td>
<td>2,813</td>
<td>2,854</td>
<td>2,798</td>
</tr>
</tbody>
</table>

*Significant at the 0.05 level  **Significant at the 0.01 level  ***Significant at the 0.001 level
This ideally refers to net population densities, the densities of those areas that are residential. The tests of these assumptions almost always are forced to use gross population densities, the populations of areas such as census tracts divided by their total areas, which can include substantial nonresidential areas. This becomes a source of error in the tests of the model. But this also serves as the basis for estimating exposure to crime other than from residential population.

The argument starts with the assumption that net residential densities in a limited area are relatively constant. Differences in the gross densities observed for areas such as census tracts are therefore assumed to result from the presence of nonresidential areas in those tracts. These differences can then be taken as a measure of exposure to crime due to factors other than residential population.

This is operationalized as follows: The maximum population density of the queen-contiguous tracts (sharing an edge or a vertex) to any tract is taken as an estimate of the residential population density. The assumption here is that the maximum density tract is completely residential; there will be some error here, of course. The contiguous tracts and the maximum density were determined using the publicly available GeoDa software (Anselin 2003, 2019).

The difference between the maximum density for the contiguous tracts and the density of tract under consideration is then assumed to result from the presence of nonresidential land in that tract. (If the density of the tract under consideration is higher, it is of course assumed that the tract is entirely residential and the difference is set to zero.) This difference is then taken as a measure of exposure to crime associated with the activity in the nonresidential area of the tract.

Because tracts differ in land area, it would seem logical that the difference in density should be multiplied by the area of the tract. This did not work, however. The nonresidential portion of a tract does not necessarily consist only of commercial, industrial, and similar uses presenting significant additional exposure for criminal activity. The nonresidential land could be parks or undeveloped land that contribute very little to exposure. Some of the largest tracts in terms of area include large areas of park or undeveloped land. The result was that multiplying the density difference by tract area produced a poor estimate of additional exposure to crime. The difference in the densities worked far better.

This general approach bears some similarity to the approach taken by Roman (2004), who used block populations plus the mean block population for the entire city as the exposure measure. In effect, she was using the entire city as the wider area rather than the more immediate surrounding area.

The exposure variable used here is the difference (if positive) between the maximum population density in the contiguous tracts and the population density of the tract. When including this measure in the model predicting crime, the natural form of the variable, not the log form, performed better. Table 3 also presents the results for the
model including this difference in densities in addition to population as an exposure variable.

The model using the maximum contiguous tract density difference did perform better than the model using only population as the measure of exposure, as shown by the likelihood ratio test, which was statistically significant. The regression coefficient for the density difference was likewise statistically significant. Because the units are different (density versus log of workers or employment), direct comparisons of the regression coefficient with the other models is not possible.

Despite this significant, the overall regression results were much closer to those for the population-only model than to the 2 models with the other measures of exposure. The regression coefficient for population was 0.72 for this model and 0.69 for the population-only model versus about 0.59 for the other models. Other regression coefficients were also closer to those for the population-only model. And the coefficient for the proportion of males aged 15 to 24 remained slightly statistically significant as it was in that population-only model. The AIC value likewise showed the performance of this model to be much closer to the population-only model, with a value of 2,854 versus 2,865. The models using CTPP workers and MPO employment had considerably lower (better) AIC values of 2,813 and 2,798. The conclusion must be that including the density difference as a measure of exposure produces some improvement, but it is not nearly as effective in capturing exposure to crime as employment.

The Effect of Exposure on Crime Rates for Cities and Towns

The introduction included the statement that crime rates calculated using population were reasonable for use in considering the incidence of crime for the nation, for states, and for metropolitan areas. I deliberately stopped at that point. For cities and towns within metropolitan areas, crime exposure can vary such that the comparison of population-based crime rates can be problematic. A large city at the center of a metropolitan area will have large numbers commuting to places of employment in the city and additional persons coming to commercial establishments and other attractions that tend to be concentrated in such cities. This additional activity increases crime exposure—opportunities for criminal activity—in comparison with many of the other cities and towns in the metropolitan area. The result will be more crimes and higher crime rates based on population in the large city and some other cities and towns that do not reflect the true underlying incidence of criminal activity.

Gibbs and Erickson (1976) made this argument that population used for conventional crime rates for central cities underestimates the potential number of victims or offenders. Since the central cities can have widely varying shares of the total Urbanized Area or Metropolitan Statistical Area populations, the effect of such underestimates should vary with that share, which they proceeded to test. Stafford and
Gibbs (1980) extended the analysis by adding to the city/urban area population ratio a further measure of central city dominance, the proportion of the area’s retail sales in the central city. Andresen (2010) examined the extent to which municipal crime rates were affected by the choice of residential population versus the ambient population from Oak Ridge National Laboratories as discussed above and in the appendix.

The MPO data for the traffic analysis zones (TAZs) cover a 9-county area centered on Indianapolis and Marion County. These counties include over 70 cities and towns, many of which are quite small. Of these, 16 municipalities had reported Uniform Crime Report (UCR) data for 2010. The total Part I crimes reported for 14 of these places are used in the analysis (U.S. Federal Bureau of Investigation 2014). The 2010 populations are from the census. The municipalities of Beech Grove and Speedway reported UCR data but are not being included. These are 2 of the 4 excluded municipalities within Marion County. While they have their own police forces and have reported their crime statistics, these areas are also within the jurisdiction of the Indianapolis Metropolitan Police Department (IMPD) which also reports some crimes for these areas. The small numbers involved have a negligible effect on the total for IMPD and crime in the City of Indianapolis. But the numbers of IMPD-reported crimes for Beech Grove and Speedway may result in those jurisdictions’ reports of total crime being underestimates of the crime within those areas.

The MPO counts for population, total employment, retail and service employment, and students in the TAZs—the measures used in the initial exposure analysis—were estimated for the 14 cities and towns. Counts for each TAZ were aggregated in proportion to the area of the TAZ within the each jurisdiction’s boundary.

The assumption being made is that the coefficients estimated for the full TAZ model in Table 1 can be used to estimate the total exposure for the cities and towns. This is a stretch, but the idea is to get a general idea of how differences in exposure might affect crime rates as opposed to rates based only on population. The objective is not to determine accurate exposure-based crime rates.

Then comes the question as to how one uses the MPO totals for population, employment, and students for the cities and towns and the coefficients estimated from the TAZ model to develop measures of total exposure for the cities and towns. How to do this was not obvious. A variety of approaches were tried that yielded results that did not make sense. The approach finally adopted was to simply multiply the city and town exposure totals—population, total employment, and so forth—by their respective coefficients estimated from the TAZ model and then sum the results.

Table 4 presents the estimated total exposure computed by summing the products of the estimated regression coefficients times the MPO totals for the 14 cities and towns. The 2010 populations are also shown for comparison. The values for the total exposure are much smaller than the populations, totaling less than half for the 14 municipalities. To make it easier to compare the estimated total exposure to the
Table 4. Population and Estimated Exposure for Cities and Towns in the Indianapolis Metropolitan Area.

<table>
<thead>
<tr>
<th>City or Town</th>
<th>Census Population 2010</th>
<th>Estimated Total Exposure</th>
<th>Adjusted Exposure</th>
<th>Percent Change Population to Adjusted Exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson</td>
<td>56,129</td>
<td>18,145</td>
<td>54,299</td>
<td>-3.3</td>
</tr>
<tr>
<td>Brownsburg</td>
<td>21,285</td>
<td>5,773</td>
<td>17,277</td>
<td>-18.8</td>
</tr>
<tr>
<td>Carmel</td>
<td>79,191</td>
<td>29,355</td>
<td>87,846</td>
<td>10.9</td>
</tr>
<tr>
<td>Fishers</td>
<td>76,794</td>
<td>20,997</td>
<td>62,833</td>
<td>-18.2</td>
</tr>
<tr>
<td>Franklin</td>
<td>23,712</td>
<td>5,414</td>
<td>16,202</td>
<td>-31.7</td>
</tr>
<tr>
<td>Greenfield</td>
<td>20,602</td>
<td>5,116</td>
<td>15,310</td>
<td>-25.7</td>
</tr>
<tr>
<td>Greenwood</td>
<td>49,791</td>
<td>15,539</td>
<td>46,501</td>
<td>-6.6</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>820,445</td>
<td>291,423</td>
<td>872,096</td>
<td>6.3</td>
</tr>
<tr>
<td>Martinsville</td>
<td>11,828</td>
<td>3,535</td>
<td>10,580</td>
<td>-10.6</td>
</tr>
<tr>
<td>Noblesville</td>
<td>51,969</td>
<td>13,315</td>
<td>39,847</td>
<td>-23.3</td>
</tr>
<tr>
<td>Plainfield</td>
<td>27,631</td>
<td>8,923</td>
<td>26,704</td>
<td>-3.4</td>
</tr>
<tr>
<td>Shelbyville</td>
<td>19,191</td>
<td>5,642</td>
<td>16,883</td>
<td>-12.0</td>
</tr>
<tr>
<td>Westfield</td>
<td>30,068</td>
<td>8,670</td>
<td>25,944</td>
<td>-13.7</td>
</tr>
<tr>
<td>Zionsville</td>
<td>14,160</td>
<td>3,500</td>
<td>10,473</td>
<td>-26.0</td>
</tr>
<tr>
<td>Total</td>
<td>1,302,796</td>
<td>435,348</td>
<td>1,302,796</td>
<td>—</td>
</tr>
</tbody>
</table>

population, the values are adjusted by the ratio of the 2 totals, making the total adjusted exposure equal to the total population. The percent change from the population to the adjusted exposure is shown in the final column.

Most of the cities and towns had adjusted exposures that are smaller than their populations. Only 2 of the areas showed increases, resulting from their relatively larger shares of employment and students. Indianapolis, as expected, was one of those areas. Its increase was relatively modest, about 6 percent. However this results from the fact that Indianapolis constitutes such a large share of the totals, so the adjustment of the exposure values necessarily produces only a modest difference for the city. Carmel, with an increase of 11 percent, is a suburb immediately north of Indianapolis and has by far the greatest amount amount of office employment in the Indianapolis suburbs along
with significant retail and service employment as well. Many of the declines in adjusted
exposure relative to population, on the other hand, are quite substantial. A half dozen
cities and towns have adjusted exposure values more than 15 percent below their
populations.

The primary purpose is to consider the extent to which using estimated exposure
for crime rates makes a difference compared with the standard crime rates calculated
using population. Table 5 provides the results. It presents the total numbers of Part I
UCR crimes for the cities and towns along with 2 sets of crime rates. First are the usual
population-based crime rates, number of crimes per 100,000 population. Second are the
comparable crime rates calculated using the adjusted exposure shown in the previous
table. The percentage difference produced by using exposure for the crime rates is given
in the final column.

Table 5. Crime Rates Using Population and Estimated Exposure for Cities and Towns
in the Indianapolis Metropolitan Area.

<table>
<thead>
<tr>
<th>City or Town</th>
<th>UCR Crimes</th>
<th>Population Crime Rate</th>
<th>Exposure Crime Rate</th>
<th>Percent Change Population to Exposure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson</td>
<td>3,271</td>
<td>5,828</td>
<td>6,024</td>
<td>3.4</td>
</tr>
<tr>
<td>Brownsburg</td>
<td>308</td>
<td>1,447</td>
<td>1,783</td>
<td>23.2</td>
</tr>
<tr>
<td>Carmel</td>
<td>939</td>
<td>1,186</td>
<td>1,069</td>
<td>-9.9</td>
</tr>
<tr>
<td>Fishers</td>
<td>748</td>
<td>974</td>
<td>1,190</td>
<td>22.2</td>
</tr>
<tr>
<td>Franklin</td>
<td>954</td>
<td>4,023</td>
<td>5,888</td>
<td>46.3</td>
</tr>
<tr>
<td>Greenfield</td>
<td>398</td>
<td>1,932</td>
<td>2,600</td>
<td>34.6</td>
</tr>
<tr>
<td>Greenwood</td>
<td>1,951</td>
<td>3,918</td>
<td>4,196</td>
<td>7.1</td>
</tr>
<tr>
<td>Indianapolis</td>
<td>55,590</td>
<td>6,776</td>
<td>6,374</td>
<td>-5.9</td>
</tr>
<tr>
<td>Martinsville</td>
<td>972</td>
<td>8,218</td>
<td>9,187</td>
<td>11.8</td>
</tr>
<tr>
<td>Noblesville</td>
<td>903</td>
<td>1,738</td>
<td>2,266</td>
<td>30.4</td>
</tr>
<tr>
<td>Plainfield</td>
<td>720</td>
<td>2,606</td>
<td>2,696</td>
<td>3.5</td>
</tr>
<tr>
<td>Shelbyville</td>
<td>964</td>
<td>5,023</td>
<td>5,710</td>
<td>13.7</td>
</tr>
<tr>
<td>Westfield</td>
<td>448</td>
<td>1,490</td>
<td>1,727</td>
<td>15.9</td>
</tr>
<tr>
<td>Zionsville</td>
<td>197</td>
<td>1,391</td>
<td>1,881</td>
<td>35.2</td>
</tr>
</tbody>
</table>
Of course the 2 areas in which the adjusted exposure was greater than the population had lower crime rates and vice versa. The Indianapolis crime rate drops from around 6,800 to 6,400, nearly 6 percent. Carmel saw its very low crime rate go from about 1,200 to under 1,100. Looking at changes in the other direction, the exposure crime rate of 5,900 for Franklin far exceeded its population-based crime rate of 4,000. The crime rate for Noblesville jumped from 1,700 to 2,300 when exposure was used. Nine of the 16 cities and towns saw increases in their crime rates exceeding 10 percent.

One might argue that yes, these might be seen as substantial differences in crime rates when exposure is used in place of population. But there are great differences in the crime rates among these places, and those with the highest population crime rates generally have the highest exposure crime rates, and the same holds true for the low crime areas. These are very different kinds of places with very different levels of criminal activity.

The significance of considering exposure in the calculation of crime rates becomes more clear when comparing areas that are more similar to one another. Table 6 lists the population and exposure crime rates for the 6 cities and towns having the lowest crime rates among the 14 areas—population crime rates less than 1,800 crimes per 100,000 persons. These suburban municipalities are contiguous to one another and, as a group, to the city of Indianapolis. They lie in an arc to the north and northwest of the city and are the most affluent suburban communities in the metropolitan area. Most of these are primarily residential suburbs with the exception of Carmel.

Carmel has a low population crime rate and with its high office, retail, and service employment has an even lower exposure crime rate, putting it in the position of having the lowest rate among the 6 suburbs. Fishers, on the other hand, started with the


<table>
<thead>
<tr>
<th>City or Town</th>
<th>Population Crime Rate</th>
<th>Exposure Crime Rate</th>
<th>Percent Change Population to Exposure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brownsburg</td>
<td>1,447</td>
<td>1,783</td>
<td>23.2</td>
</tr>
<tr>
<td>Carmel</td>
<td>1,186</td>
<td>1,069</td>
<td>-9.9</td>
</tr>
<tr>
<td>Fishers</td>
<td>974</td>
<td>1,190</td>
<td>22.2</td>
</tr>
<tr>
<td>Noblesville</td>
<td>1,738</td>
<td>2,266</td>
<td>30.4</td>
</tr>
<tr>
<td>Westfield</td>
<td>1,490</td>
<td>1,727</td>
<td>15.9</td>
</tr>
<tr>
<td>Zionsville</td>
<td>1,391</td>
<td>1,881</td>
<td>35.2</td>
</tr>
</tbody>
</table>
lowest population crime rate—it was the only area below 1,000—but its crime rate increases to nearly 1,200 when exposure is used, putting it above Carmel. Three of the communities have population crime rates in the 1,400 to 1,500 range which increase to 1,700 to 1,900 when exposure is used. And Noblesville, with the highest population crime rate of the 6 at 1,700 sees a jump of over 30 percent to nearly 2,300, a not insubstantial change.

Of course differences in exposure do not account for much of the variation in levels of crime and crime rates across diverse areas. However crime rates that take differences in exposure into account can be substantially different from the standard rates using population. And these differences may be quite significant when making certain types of comparisons.

Conclusions

The measures from the Indianapolis MPO transportation planning data—population, total employment, retail and service employment, and students—each play a significant role in predicting crime and contributing as measures of exposure. Including these measure of exposure also changes the coefficients and pattern of significance for the population characteristics in the model. This demonstrates the importance of correctly and fully taking exposure to crime into account in models predicting crime. Failure to do so, with the resulting error in exposure, can bias other estimates in the model and alter conclusions regarding the causes of crime.

Including total employment from the Census Transportation Planning Products (CTPP) accounted for exposure about as well as total employment with population from the MPO data. And this was the point at which the improvements to exposure produced relatively stable estimates of other coefficients in the model, which remained essentially unchanged with the addition of further exposure measure. The CTPP breakdown of employment by industry was not comparable to the MPO data and was not considered here. But it holds the potential for providing further improvements for exposure. Trying to improve the treatment of exposure beyond population solely by the use of census population data was less successful, however.

The current effort should be considered to be only a start in understanding exposure to crime. Other variables can be considered as potential proxies for exposure. Boggs’ (1965) early study of crime rates employed data on land use as measures of activity. More recent work on the relationship of land use to crime (e.g., Stucky and Ottensmann 2009) show more recent interest is such measures.

Exposure to crime and its measurement will vary, likely significantly, with the context and the type of crime. Boggs (1965) used different variables to calculate alternative crime rates for different types of crimes. Stucky and Ottensmann (2009)
showed differences for types of crime in relation to land use. Recent examples are the studies by Boivin (2013) and Felson and Boivin (2015). This study mentions variations in the estimates of exposure coefficients for different crimes, but detailed results were not presented. The performance of proxies for exposure will likewise vary with the spatial scale and the size of the areal units of analysis, as shown here in the comparisons using the traffic analysis zones (TAZs) and the census tracts. And it is worth further noting that with very large areas such as states and metropolitan areas, population alone may serve as a perfectly adequate measure of exposure.

The large quantities of data made available by the use of new technologies have the potential to advance understanding and measurement of exposure to crime. Some early studies have looked at Twitter message data and mobile phone location data as predictors of crime (Malleson and Andresen 2015, 2016; Hipp, et al. (2019); Sheard, Malleson, and Birkin 2019). The magnitude of such datasets can produce unrealistic expectations regarding their efficacy for the measurement of exposure and the prediction of crime, however. Twitter users are just a fraction of the population and the most active users accounting for a large share of all messages are a yet smaller and more unrepresentative sample. Mobile phone location data would seem to be more representative as mobile phone use becomes closer and closer to being ubiquitous. But with respect to exposure to crime, a visitor to a downtown area for a few hours for shopping or recreation is likely to have much greater exposure to at least some types of crimes per hour of time present than an office worker spending much of the day in what is likely the relatively safe, low-risk environment of an office.

Further research on exposure to crime is important for 2 reasons. First, understanding what contributes to the risk of crime and exposure is itself an interesting and valuable question to answer. But at least as important is the need to properly account for exposure in models seeking to address other questions regarding the factors associated with criminal activity. As shown in this paper, error in the way exposure is accounted for in a model can result in different, biased estimates of the coefficients of other variables of interest.

Research must not fall into the trap of attempting to find the “best” measure of exposure, the single measure that is the most effective predictor of crime, the variable that can be used as the most appropriate denominator for an alternative crime rate. If this paper has shown anything of value, it is that multiple measures serve as proxies for and contribute to exposure to crime. The methods described and used in this paper provide a framework for continuing to examine exposure.
References


Appendix
The LandScan Ambient Population Data and the Andresen Papers

This appendix addresses 2 related issues. Papers by Andresen used the LandScan ambient population data for the analysis of crime. The first section discusses problems with the LandScan data, raising questions about its suitability for this purpose. A second shorter section briefly shows that the argument made by Andresen for the superiority of the LandScan data is a logical fallacy such that the papers would not support his claim even if the LandScan data were acceptable.

Problems with the LandScan Ambient Population Data

A series of papers by Andresen makes use of the LandScan ambient population data from Oak Ridge National Laboratories as an alternative to population for calculating crime rates, arguing for its superiority (Andresen 2006, 2010, 2011; Andresen and Jenion 2010). The papers suggest that the LandScan data is a better measure of crime exposure than population. The validity of Andresen’s claim is dependent in large part on the suitability of the LandScan data as a measure of exposure to crime and, for that matter, as a measure of what it purports to measure. This section argues that the LandScan data is flawed and not appropriate for this (and other) purposes.

The LandScan Data

Oak Ridge National Laboratory has been developing and providing the LandScan datasets showing the distribution of what they call “ambient population” for small areas for the entire world (Oak Ridge National Laboratory 2019). Ambient population is sometimes described as the average 24-hour population in an area as opposed to the population resident in the area, which is the way in which censuses count population. (The issue of ambient population is discussed in more detail in the following section.)

LandScan is a global dataset providing estimates of ambient population for 30 arc-second (1/120 degree of latitude and longitude) grid cells. These are approximately 1 kilometer or 0.6 mile square at the equator, becoming narrower at higher latitudes as the lines of longitude converge. The LandScan data are freely available only to agencies of the United States government and, only very recently, to researchers at educational institutions. All other users may have to pay to use the data.

The population counts are developed by taking the census populations for (in many instances) the smallest administrative units available in each country (which may be quite large) and allocating that population to the grid cells using data such as land cover, roads, slope, urban areas, village locations, and high resolution imagery. These
data are used to develop probabilities or likelihoods of the population for each cell which are then used to do the allocation. Documentation of the process is sparse. The LandScan website includes only general descriptions of the methodology and the data employed (Oak Ridge National Laboratory 2019). The most detailed description of the LandScan methodology that I have found is an early article by Dobson, et. al. (2000). A more recent paper (Rose and Bright 2014) acknowledged that “a frequent critique is the lack of documentation of and transparency into the process by which the data is developed each year.” But that paper goes on to describe the LandScan methodology in the same general way as prior works and the website, giving no more (and perhaps even less) detail.

The Issue of Ambient Population

Ambient population is defined variously on the LandScan website as the average population over 24 hours or the average day/night population count (Oak Ridge National Laboratory 2019). These could be interpreted in different ways. This also begs the question of whether this refers to weekdays or all days of the week. This imprecision in the definition of ambient population is less critical, however, because no evidence is provided that they have any actual data on ambient population for any areas that might have been used either for the estimation or validation of their population distribution models. I know of some attempts to estimate daytime populations for small areas, but none that proceed to come up with anything comparable to their ambient population.

Furthermore, the use of the term “ambient” to describe the population they are estimating is curious. The word ambient refers to the surroundings or the background of something. Google searches show the term used for the LandScan data and by studies using or referring to the use of that data. (At one point it was used by a product called Golaem Crowd, a software tool used for the computer animation of crowds for use as backgrounds for movies, commercials, and computer games. In those instances, the crowd is truly the ambient background for the primary action, human or animated.)

A clue to the origin of the use of “ambient population” comes from the statement on the LandScan website acknowledging their funding, which comes from the Department of Defense. The LandScan data was likely developed for the United States military for use in the estimation of collateral civilian casualties as a result of military operations. For such purposes, that population could reasonably be described as being ambient, in the background, to the military activity.

If this is the primary purpose for the development of the LandScan data, it explains several things mentioned in an early article documenting the dataset (Dobson, et al. 2000). Reassurance is provided that grid cells with any sign of human activity would have at least some nonzero population assigned, communicating at the least the
possibility of some civilian presence. One portion of the “validation” used data for states in the southwestern United States similar to areas of the Middle East, which would be logical given the military’s interest in that area especially since the first Gulf War.

This may also help explain some of the limitations to the methodology used in the creation of the LandScan dataset. Probabilities for the assignment of population to the grid cells using the additional sources of data seemed to have been developed with little analytical foundation, almost as subjective guesses. But if the purpose of the dataset is to provide information on the potential risks of civilian casualties from military operations, accurate counts are hardly needed. Rough estimates as to whether the areas were densely populated, sparsely populated, or unpopulated would likely be sufficient. In this case, using best judgement to assign probabilities for the population allocation would seem reasonable. However this raises questions about the utility of the data for other applications, particularly those concerned with specific values at the grid cell level.

The Issue of Validation

Since no data are available on the ambiguously defined ambient population for any areas, much less for the grid cells, true validation of their procedure is not possible. The closest approximation of meaningful validation have been tests how effectively the procedure assigns population from larger areas (which might be the best data available for some countries) to the grid cells. This has been done by performing such allocations from large administrative units in developed countries that also have population data available for smaller areas to allow for some assessment of the accuracy of the allocation (Dobson, et al. 2000).

The reported evaluation considered 3 cases in different countries, but the most detail was provided on a test using states in the southwestern United States. This will be discussed further here. This region was chosen because its geography is similar to the Middle East. For the test, populations for states are allocated to the grid cells using their methodology. These grid cell populations are then aggregated to the county level and compared with the actual county populations as an assessment of the accuracy of the allocation. (It is not made clear why assessments were not made using more detailed data for smaller areas such as census tracts.)

The authors are satisfied with the results, pronouncing the confirmation of the validity of their methodology. They do admit to some inaccuracy in the county totals produced, with the simulated ambient population for Sacramento County being 49 percent higher than the census population. They suggest that this may in part be due to the actual ambient population being higher than the residential population, which they
say they expect should especially be the case for state capitals. No explanation is given for this rather bizarre suggestion.

The Impossibility of the Data Doing What It Claims for the U.S. and Likely Other Developed Countries

The prior issues raised concerning the limited documentation and lack of real validation of the LandScan model should at least give pause concerning the utility of the data as a measure of crime exposure. But even if these problems did not exist, a further issue relating to the census data being allocated to the grid cells means that the LandScan data cannot possibly be measuring what it is supposedly intended to measure for the United States and likely for other developed countries having spatially detailed census data.

The current website and a more recent publication acknowledge a limitation on the minimum size of the areas that can be used for the population data for the allocation to the grid cells. The website includes the following statement:

Very small administrative or enumeration areas equivalent to US census blocks or block groups [emphasis added] have unintended consequences for modeling an ambient population. Since the populations associated with census tables are places of residence, commercial and industrial areas may have zero or very low populations associated with them. Thus the output would be reflective of a residential only population distribution instead of an ambient population distribution (Oak Ridge National Laboratory 2019).

The paper by Rose and Bright (2014) included a similar statement. This statement, as far as it goes, is certainly correct. However, the explicit reference to census blocks and block groups suggests that the next largest census units are suitable for use in the allocation. This is backed up by a statement made by Dobson, et al. (2000) after discussing their validation exercise in the southwestern United States in which they allocated state populations to the grid cells. They assure readers that the actual LandScan data for the United States allocates the data not from the states but from census tracts.

Allocating the population from the census tracts to the grid cells makes it impossible for the LandScan data to provide grid cell populations that correspond to any notion of daytime/night-time or 24-hour ambient populations. The population allocated to grid cells within a census tract cannot exceed the population of that tract. The population allocated to the grid cells within a group of census tracts cannot exceed the population of those tracts. The central business district of a large urban area can have employment of 100 thousand or more, in some cases much more. This does not include the large numbers of additional people who will travel to and visit the central
business district during a workday. Since these people are not there during the entire 24-hour period, the ambient population, however, defined, will be substantially less. But any reasonable estimate of an ambient population will be far higher than the resident population of the census tracts encompassing the central business district that could potentially be allocated to those grid cells. Therefore, allocating populations from census tracts to the grid cells, no matter the method used, cannot produce reasonable estimates of central business district ambient population. The same will be true for other major concentrations of employment and activity, certainly regional shopping centers and large areas of suburban office development.

But the problems created by allocating the population from the census tracts to the grid cells is not limited to very large concentrations of nonresidential activity such as the central business district. Consider this hypothetical case of 2 adjacent census tracts. The first encompasses 2 of the grid cells and is entirely residential. The second has 4 grid cells; 2 are residential with the same population as the first tract; 2 are not not residential but commercial or industrial. The first tract would have its population allocated, presumably approximately equally, to the 2 grid cells. This would constitute an overestimate of the ambient populations for these cells as it would not be taking into account commuting and other travel out of the residential area to other areas during the day. The second tract with the nonresidential area would presumably have some of its population allocated to the nonresidential grid cells, reducing the population left to be allocated to the residential grid cells. Thus largely identical residential areas would have different ambient populations. In the first, all-residential tract, the estimates would be too high while in the second tract with the nonresidential areas, the estimates for the residential grid cells would likely be too low, as they would have been the sole source for the population allocated to the nonresidential grid cells.

The Use of Ambient Population in the Criminological Literature

The series of papers by Andresen (Andresen 2006, 2010, 2011; Andresen and Jenion 2010) brought the concept of ambient population and the use of the LandScan data to the criminological literature. In these papers, the ambient populations for areas have been used to compute crime rates, which are then compared with traditional crimes rates using residential population. All of the studies examined crime and use ambient population data for Vancouver, British Columbia, and in some cases for the surrounding area.

The first paper (Andresen 2006) provided a fairly detailed summary of the LandScan data based largely on Dobson, et al. (2000). Andresen accepted their presentation without question and was very positive. One of the next papers (Andresen 2010) did note that no validation had been carried out for the Vancouver area but said that was is beyond the scope of that paper. Andresen and Jenion (2010) did raise some
concerns that no independent validation of the data have been conducted. They also cited the problem that the algorithm used is proprietary so that the parameters used cannot be verified or compared to empirical estimates. (They seemed to assume that the parameters had been derived from empirical estimates, which may or may not have been correct.) And they further noted that estimates of ambient populations were not available to perform validation. They failed to question how empirical estimates of model parameters could have been made in the absence of any such data.

The information provided in the papers about the LandScan ambient population data actually provides further basis for questioning the reasonableness of that data. In the first paper (Andresen 2006) the analysis is carried out using census tracts, with the ambient population estimated for those areas. (Canadian census tracts are very similar to those used in the United States.) The summary statistics reported for the census (resident) population and the ambient population would seem to confirm that the grid cell ambient population has been estimated from the tract populations. The census maximum tract population was about 12 thousand while the maximum ambient population estimated from the data for the grid cells was only slightly higher at 15.7 thousand. It is impossible to believe that the daytime and then ambient populations in the census tracts encompassing the Vancouver central business district were not far higher than the residential populations. Andresen cited a comparison from a popular source contrasting extreme differences residential and ambient population (based on the presence of employment) in an area of downtown Houston to emphasize the importance of considering ambient population. But he proceeded to say that such radical differences are not present in Vancouver without citing any justification (unless he was basing this on the LandScan data). Even if Vancouver has substantially higher residential populations in the central business district tracts, employees commuting in and visitors coming to that area would have to be raising the ambient populations to far higher levels than the residential population.

One additional point with respect to the Andresen (2006) paper. In this paper and some of the subsequent papers, he estimates fairly standard models to predict crime rates using both resident population and ambient population and finds that the latter perform somewhat better. How might one account for this seeming improvement if the ambient population data are basically flawed? Presumably the ambient populations for the grid cells have been estimated by allocating the resident populations from the census tracts. The grid cell boundaries will generally not coincide with the tract boundaries, so many grid cells will have ambient populations derived from populations for 2 or more census tracts. Then Andresen estimates the tract ambient populations from the LandScan grid cells, now necessitating the allocation of grid cell populations to multiple tracts. The result of these conversions will be a smoothing of the census tract residential populations, reducing some of the random variation that occurs between tracts. And this reduction in the random variation of the tract crime rates when using
the ambient population data could have produced the somewhat better performance for those models.

Andresen (2010) examined the effects of using ambient populations rather than residential populations to calculate crime rates for municipalities in the Greater Vancouver area. A table in that paper lists the residential and ambient populations for each of the municipalities. They are virtually identical for Vancouver. This is completely implausible, as one would expect the city at the heart of the metropolitan area to have a much higher ambient population with commuting in from the other areas. It is also inconsistent with data presented in a prior table from another source showing that Vancouver has by far the lowest percent commuting out of the area.

Andresen and Jenion (2010) performed their analyses using enumeration districts. These are subdivisions of census tracts having the name also used by the United States Census for tract subdivisions before renaming them block groups. The means for census population and ambient population estimated for these are similar but the standard deviation is much higher variation ambient population. This consistent with the LandScan procedures allocating the tract populations unevenly to the grid cells. Looking at the maps showing census and ambient populations for the enumeration districts gives further pause. At least one of the enumeration districts in the central business district has a population in the highest class mapped. This ranges up to the maximum of 8,500. But other central business district enumeration districts actually have ambient populations less than their census populations.

Some final comments on the Andresen (2011) paper. This paper returns to using census tracts as the units of analysis. Summary statistics are again provided. The values for the census population are slightly different from those in the 2006 paper, to be expected given the former used data from the 1996 census and the latter from the 2001 census. However the maximum tract ambient population for this paper is 85,000, compared with less than 16,000 for the earlier paper. The discrepancy is not noted and it is impossible to know whether the difference results from major changes to the LandScan data (which are revised annually) or from errors in processing the data. A map of ambient population by census tract shows 3 clusters of tracts having the highest populations. Two are far from the central business district while 1 includes a very small portion of the central business district. The other central business district tracts have populations in the lowest or second lowest of the 5 population classes used for the map. Note that the tract with the highest population of 85,000 is not in the central business district. And that population is about one-eighth of the total population of Vancouver!

The papers by Andresen brought attention to the idea of ambient population to the criminological literature, but hardly in a positive way. Boivin (2013) did not use the LandScan data but did adopt the term ambient population from Andresen. Boivin provided a definition of ambient population different from the (vague) descriptions used by LandScan, describing ambient population as “an estimation of how many
people were present in a given area at any time of the day, any day of the year.” A vague definition is bad enough; a morphing of that into a different vague definition makes things even worse. Boivin then equated ambient population with the employment in a census tract (which is not consistent with either definition). Boivin draws a contrast to the work of Andresen and Jenion (2010):

…the authors [Andresen and Jenion 2010] aimed to calculate a rate based on an estimation of the actual number of individuals present in an area at any given time of the day – remaining residents and visitors altogether. In the current analysis, the goal was more about measuring the separate impact of residents and visitors on crime (Boivin 2013).

By visitors, Boivin was actually referring to employment in the tract without acknowledging the presence of additional visitors that could come to a tract.

Sheard, Malleson, and Birkin (2019) defined ambient population in the following manner:

The term ‘ambient population’ refers to the actual number of persons who are present within a particular area at any given time.

They measured ambient population using mobile phone data. Malleson and Andresen (2015) offered the following about ambient population:

The ambient population is highly dynamic and exhibits strong spatial and temporal fluctuations at various scales (e.g. hourly, daily, seasonal).

This is, of course, strongly at odds with the LandScan notion of ambient population as a 24-hour average as described by Andresen in the earlier papers. This paper used total Twitter messages over nearly a 2-year period as the measure of ambient population. And this is inconsistent with their statement that ambient population is highly dynamic with temporal variation.

Hipp, et al. (2019) also used Twitter data to estimate what they call “temporal ambient population,” population at different times of the day. Finally, Malleson and Andresen (2016) used multiple measure of population—residential census data, workday census data, Population 24/7 data from a project that attempts to redistribute census population to a set of grid cells using likely activities, mobile phone activity counts, and Twitter messages. Presumably some of these, all of these, or some combination of these constitute the ambient population?

At this point I am giving up on looking at other uses of ambient population in the criminological literature. It is a mess. The term “ambient population” has essentially
become meaningless. At most it can be seen to refer to some measure of the population/exposure to crime other than the residential population. Its continued and expanded use (and misuse) may have resulted from the persistent focus on crime rates, with the necessity of identifying some single measure of “population” to be used in the computing of those rates (as opposed to the use of count models to predict crime bringing the ability to simultaneously consider multiple measures of exposure).

Other Problems with the Andresen Ambient Population Papers

The papers by Andresen (2006, 2010, 2011; Andresen and Jenion 2010) using the LandScan ambient population data are of questionable value given the problems identified with that data. But even if those data were acceptable, the manner in which these papers draw inferences from the empirical results is fatally flawed, rendering the conclusions unsupported.

The fundamental problem can be seen even in the abstract of Andresen and Jenion (2010) where they make the following statement:

Calculated crime rates using the residential and ambient populations exhibit a weak statistical relationship. This provides a strong positive implication for the use of these data such that their utilization may give a more precise depiction of victimization, particularly when considering violent crime.

The conclusion that the ambient population data may be a superior measure of the risk of crime is just not supported by the evidence offered. That the crime rates calculated using the ambient population are dissimilar from the traditional rates using residential population does not imply that the ambient population crime rates are better. It just means that the ambient population and the ambient rates are different from the residential population and the residential rates. They could be better; they could be worse; they could be nonsensical.

This type of inference is employed within the paper. For example, after examining the relationship between ambient population and residential population across the enumeration districts and concluding that it is a weak relationship, they draw this conclusion:

At this point in the analysis, it is clear that obtaining the ambient population is worth the effort to supplement conventional calculations.

They do exactly the same thing after comparing crime rates calculated using ambient populations and residential populations:
The comparison of residential- and ambient-based violent crime rates further illustrates the utility of employing the ambient population in criminological research.

The same type of inference underlies the conclusions drawn in the other papers. More examples will not be offered here.

The problem is that the method of inference employed to draw the conclusions is a logical fallacy. The full argument is not explicitly presented as such (which would have made the logical fallacy more apparent). But it can be stated as follows. It begins with the assertion of the truth of the following conditional proposition:

If ambient population is a better measure of the risk of crime than residential population, then the relationship between ambient population and crime (such as crime rates) will be different from the relationship between residential population and crime.

There is nothing wrong with this. I would have no problem accepting the truth of this statement.

The papers then present empirical evidence comparing, for example, crime rates calculated using ambient population with crime rates calculated using residential population. They find these to be dissimilar. This is used to assert the truth of the following:

The relationship between ambient population and crime (such as crime rates) is different from the relationship between residential population and crime.

Again, I have no problem with their evidence and assertion of the truth of this statement (aside from the questions about the LandScan data).

But they then make the inference that the truth of the prior two statements implies the truth of the following statement:

Ambient population is a better measure of the risk of crime than residential population.

Drawing this conclusion from the prior statements involves the logical fallacy of affirming the consequent. This is the fallacy where the antecedent in an indicative conditional is claimed to be true because the consequent is true, an argument of the following form:

If A, then B
B

Therefore A

Thus it is the case that these papers provide absolutely no evidence to support the proposition that the LandScan ambient population is a superior measure of the risk of crime than the residential population.

A Further Note

It is not the intention here to go through these papers in detail to identify further flaws. Indeed, the problems with the LandScan ambient population combined with the logical fallacy in the arguments render the papers worthless. Nevertheless, while examining the inferences cited above in the Andresen and Jenion (2010) paper, I found a problem, repeated, which I have to comment on.

As mentioned above, Andresen and Jenion are looking at relationships involving ambient and residential populations to draw the conclusion that these are only weakly related. In looking at the relationship between the distributions of the ambient and residential populations across the enumeration districts, they begin by reporting the correlation coefficient $r = 0.54$ and say that this is not so high as to suggest one could substitute one for the other without affecting an analysis. I can accept that. But then they report the results of the bivariate regression of ambient population on residential population and make the preposterous claim that “the case for substitutability becomes even weaker,” since the adjusted $R^2 = 0.287$.

They do the same thing with crime rates, first giving the correlation and then doing the regression and reporting the adjusted $R^2$, once again suggesting that this is providing significant additional information.

Who reviewed this paper??!!!