Chapter 26:

Data Preparation for Crime Travel Demand Modeling

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Chapter 26:

Data Preparation for Crime Travel Demand Modeling

In this chapter, the data requirements for the crime travel demand model are discussed. At the minimum, there are four types of data that are needed for the crime travel demand module:

1. A zonal system;
2. Matched crime data listing both crime location and likely origin location. This can be, further, broken down by crime types, time of day, day of week, and other sub-sets of the total number of crimes;
3. Socioeconomic and land use data for the zones which are used as predictor variables; and
4. Network data on the road system and the transit system.

In addition, there can be supplementary data that help expand the predictive models. These include:

5. Policy-related data (e.g., strategic or planned interventions)
6. Crime data on the actual distribution of crimes by zones, which is used to correct the implied distribution from 2 above.

The following is a discussion of each of these requirements.

Choice of a Zonal System

The crime travel demand model is a zonal model. That is, it analyzes crime trips by zones. For all four stages, the estimates are for zones, not for individuals. Thus, at the trip generation stage, there are two zonal models - one predicting the number of crimes originating in each origin zone and one predicting the number of crime ending in each destination zone. At the trip distribution stage, there is a prediction of the number of crimes which originate in each origin zone that end up in each destination zone (the implicit number of *trips*). At the mode split stage, the trips for each origin-destination zone pair are, further, sub-divided into different travel modes. Finally, each origin-destination zone pair by travel mode is assigned a route. But, at all stages, the estimates are for zones.

Typical Zone Systems

This makes the choice of a zonal system very critical. In practice, three types of zone system have been used:

1. Census geography
2. Traffic analysis zones
3. Grid cells

Census geography follows the geography used by the U.S. Census Bureau (in the United States) or by other national census agencies. Traffic analysis zones are used by most transportation planning agencies for modeling transportation in a metropolitan area. They are typically super-sets of census geography (e.g., two census tracts combined). Finally, grid cells are uniform zones imposed on a metropolitan area. While they have desirable statistical properties, they are rarely used in practice.

Problems with Large Zones

In deciding on a choice of a zonal system, there are several important issues that must be balanced. The first problem one faces is that of zone size. Large zones can distort relationships. It can be shown that the size of a zone has an impact on the statistical relationships between the predictor variables and the dependent variables, which are the number of crime trips by either origin or destination zone. Typically, the larger the zone size, the stronger the relationship. The reason for this effect is complex and has to do with a number of factors, for example minimizing within-zone differences in travel behavior and, therefore, maximizing the between-zone variance relative to the within-zone variance (Langbein & Lichtman, 1978) or aggregating spatial autocorrelation to minimize adjacency effects (Anselin, 1995). But, the effect is well known. The cost of having this stronger statistical relationship is to produce a less precise estimate for the region since within-zone differences are minimized.

One can think of this in terms of an arbitrary point within a zone (e.g., the centroid of the zone though it could be any location within the zone that is taken as the focal point for estimation). All the data in the zone are assigned to that point. Thus, the number of crimes that originate within the zone or end within the zone are assigned to a single point. This means that whether a crime occurred at the edge of the zone or directly in the middle, it is assigned geographically to a single point. Similarly, any of the predictive socioeconomic or land use variables are also assigned to that point (e.g., median household income). Hence, any spatial differences within the zone are eliminated as all events and households are assumed to ‘live’ at that point. If there are two adjacent zones, for example, that differ in income levels, most likely there is a gradient of income from one to the other; however, putting the measurement of income

at a single point in each zone exacerbates the differences between the zones while ignoring the similarities (e.g., at the edges of the zones where the population on both sides are liable to be more similar). It should be clear that the larger the zone size, the greater the exaggeration between the zones. In other words, larger zones exacerbate differences between zones while minimizing similarities. The result is an oversimplification of the distribution of characteristics of those neighborhoods.

In addition, larger zones have too many trips that both originate and end in the same zone (intra-zonal or 'local' trips). Clearly, the larger the average size of a zone, the more likely that a trip will be entirely within the zone. Thus, there is a strong relationship between average zone size and the number of intra-zonal trips. This will be less useful since it minimizes the complexity of travel. The extreme would be to divide a metropolitan area into only a few zones (e.g., 4 or 5). The result would detect large scale travel patterns, but would lead to a majority of trips occurring within each zone. One would not be able to say very much about crime travel other than a few general patterns (e.g., crime trips from the central city to the suburbs).

On the other hand, if the zones are too small, there is a danger that there would be more cells in the trip distribution stage (see Chapter 28) than there are actual events. The result would be inadequate degrees of freedom in a model and unreliable coefficients. A zone model has to balance the need for increased precision with the ability to produce stable estimates.

Problems in Obtaining Data for Small Zones

In theory, the ideal zone size would be small, say on the order of a block or two. This would allow precision in estimates and the ability to examine the complexity of travel in a metropolitan area. The reason that this is not done very often, however, is the lack of data at the block or block group level. While crime data can be allocated to blocks or block groups, it is often difficult to obtain socioeconomic data at that level. In the United States, for example, while the U.S. Census Bureau will release data down to the block level, confidentiality requirements require that no data be able to identify individuals. Hence, there is very limited data at the block level, typically gender and race distribution. Block group data, on the other hand, is often easily available, including critical income factors.

The biggest problem with a block group zonal system is in obtaining employment data. The U.S. Census Bureau only collects a sample of employment data from the decennial census which they release in their Journey-to-work data set (U.S. Census Bureau, 2012). They release this data for fairly small geographical units (e.g., block groups) and also produce yearly estimates for larger geographical units. These data can be used to construct employment estimates for small geographical areas; however, it is current only in those years close to the

census year and becomes quickly outdated. The Bureau of Labor Statistics also collects employment information, but will not release it at such a small geography.

Thus, obtaining these data depends on local organizations, such as a Council of Government (COG) or a Metropolitan Planning Organization (MPO). Till now, these data have not typically been released at small geographies such as block groups, but, instead, at a larger geographical unit called a *traffic analysis zone* (TAZ). However, because of the widespread use of GIS and the increasing incorporation of high resolution aerial photography into GIS-based land information systems, this situation is changing. For example, at the Houston-Galveston Area Council, the MPO for the greater Houston area, employment estimates are made for as small a geography as a 1000 foot by 1000 foot grid cell, essentially a couple of city blocks. Thus, it is starting to become possible to obtain employment data at very small geographical levels. In the next few years, more and more data will be available for small geographical units and the size limitation mentioned above will slowly disappear.

There is a converse problem with size, however, that also occurs. If the zones are too small (e.g., if data could be obtained at a block face level), there will be too many cells with no crime events. The smaller the geographical unit, the more likely that there will be no events recorded. For example, to illustrate the crime travel demand model, I have used data from Baltimore County. The crime data were 41,974 incidents that occurred between 1993 and 1997 for which both a crime location and a crime origin were known. To model these incidents, traffic analysis zones (TAZ) were used. For Baltimore County, there were 325 destination TAZ's while for both Baltimore County and Baltimore City, there were 532 origin TAZ's. Taking the origin TAZ's, with 41,974 incidents the average number per TAZ was 78.9. However, in practice, 27 zones had no crimes originate from them (or approximately 5%). If a smaller geography was used (e.g., block groups), the number of zones with no crime originating in them would increase substantially, as would the percentage. At some point, if the geography becomes very small, a high proportion of the zones will have no crimes originating from them. This makes modeling very difficult as the average number of events will tend towards zero. While there are techniques for modeling a skewed distribution (which will be discussed in Chapter 27), the more skewed the distribution, the less accurate typically is the estimate. Extremely skewed distributions are more problematic for modeling than mildly skewed distributions as the variance terms become very complex to estimate (see Chapter 16).

Still, on average, a small zone system is preferable to a large one. There is so little data for very small geographies that the problem of zones being too small is an unlikely one, at least for the foreseeable future. Where possible, users should try to obtain data at the smallest geographical level for which data can be obtained.

Problems with Irregular Size and Shape

Another problem facing the choice of a zonal system is the irregular sizes and shapes of most zonal data. For example, the U.S. Census Bureau uses a unit called the *census tract* for the collection of census information. The census tract is supposed to be an area of approximately equal population (though it is rarely entirely equal). These units generally are wholly within jurisdictions (though there are exceptions) and they are made up of blocks and block groups (collections of blocks), but in turn are aggregated upward to form enumeration areas within each jurisdiction. This logic makes sense in terms of the mission of the U.S. Census Bureau, which is to take the census. The geography respects political jurisdictions (counties and cities), but is fine enough to help manage the data that is collected during the decennial census.

But, from a modeling viewpoint, this geography has problems. First, the area of census tracts typically increase from the central city outward to the far suburban edges of a metropolitan area. Because the logic of the census tract is to approximate an area of equal population, by necessity the tract area will increase with the lower densities in most suburban communities. Thus, any data assigned to a tract (or to a block or block group within a tract) will be less precise in the suburbs than in the central city. In a travel demand model, one can end up with absurdities whereby trips appear to originate at locations where there are no people simply because the centroid of the zone falls at a location where there are no households (e.g., in a reservoir). The uneven size of zones usually means that a travel model will be more precise in the center of a metropolitan area than in a suburb.

Second, because census tracts are often defined with respect to principal arterial roads (which form their edge), they often will have irregular shapes. This could add a potential source of error in that all events and household characteristics within a boundary are assigned to a single point in the zone. On the other hand, if the zones have been selected to represent a neighborhood which is relatively uniform, such irregularity may not be a problem. Nevertheless, if two zones have very different shapes (e.g., one is square while the other is pointed), allocation error (and, hence, modeling error) is liable to be greater in the one that is more irregular, all other things being equal, than in the one that is square. This is the so-called Modifiable Area Unit Problem (MAUP)(Wikipedia, 2012; Hipp, 2007; Wooldridge, 2002; Openshaw, 1984).

Again, ideally, a zone system should be a grid whereby each zone is a square of equal size; shape and area effects are constant for all zones. While geographers recognize the value of a grid cell for zonal allocation, in practice, it is rarely used. Among the transportation planning agencies in the country, very few use a grid system. Of the ones with which I am familiar, only

the Chicago Area Tranportation Survey (CATS) uses a grid system.[1] In Chapters 31 and 32, Richard Block and Dan Helms discuss applying the crime travel demand model to Chicago.

Therefore, to sum up, in practice, one has to balance four different criteria in selecting a zone system for a crime travel demand model:

1. Zone size (generally, smaller is better within limits)
2. Consistency of zone size (less variability is better)
3. Distortion due to shape (more regular is better)
4. Availability of data

Unfortunately, it is the fourth criterion - the availability of data that is usually the determining factor in the choice of a modeling zonal system. Hopefully, this will change in the future as more data at the smaller geographical level become available.

Trips from Outside the Study Area

One other problem confronts the choice of a zone system. Irrespective of which zone system is used (census geography, TAZ, grid cells), a decision has to be made about the extent of the area to be used in modeling. The choice of destination zones is made by the availability of crime data. Typically, data are collected by police departments for their jurisdiction. Unless data sets from several adjacent jurisdictions can be obtained and combined, the analyst typically will be restricted to modeling the jurisdiction for which the crime data has been collected. This is called the *Modeled Jurisdiction*.

Modeling the origin zones is a decision about which zones contribute to the crimes occurring in the modeled jurisdiction. That is, some of the origins of the crime trips occurring within the modeled jurisdictions may come from outside that jurisdiction. For example, in the case of Baltimore County, approximately 42% of the crimes occurring within that jurisdiction were committed by offenders who lived outside that jurisdiction, of which 38% originated from the City of Baltimore.

In such a case, it is very important to include zones beyond the modeled jurisdiction in the crime origin model. That is, to use Baltimore County as an example, if the predictive model for crime origins only included the 325 TAZ's within that jurisdiction, the model would not adequately assess the factors predicting crime origins.

---

[1] CATS was used as the prototype by the Federal Highway Administration for developing the original travel demand model. The grid was used because it minimized errors due to irregular size and shape. Nevertheless, that model has not been followed by planning agencies in the U.S.

But where does one draw the line? Eventually, because of limitations due to data or due to the need to restrict the analysis, a boundary has to be drawn around the study region. Some crimes will inevitably originate from outside that line. These are called *External Trips* and refer to the trips that originate from outside the study area. While there is no 'hard and fast' principle, generally transportation planners recommend that the study area include at least 95% of the trips that end in the modeled jurisdiction (Ortuzar & Willumsen, 2001). With such coverage, the 5% (or less) that are external trips will have little effect on the model parameters, and the amount of bias will be small (but will always exist unless 100% of the trips can be measured).

I will come back to this point in the next chapter. But, the critical point is that the zone system must incorporate a sizeable area in which at least 95% of the crimes originate from within. Going back to the Baltimore County example, adding in the City of Baltimore increased the percentage of trips originating within the study area from 58% (for just Baltimore County) to 96%, an acceptable level to 'draw a boundary' around the study region.

Small Area Limitations

A travel demand model is aimed at modeling travel patterns in a metropolitan-wide area. The model is particularly good at estimating travel for the region as a whole and for large sub-areas of the region. The model is not particularly good at estimating travel within small geographical areas. The problem of intra-zonal trips - trips in which the origin and the destination occur in the same zone, represent trips for which the model cannot describe the travel pattern. These are trips that the model detects are within a small area, but cannot estimate where these occur. Similarly, trips between adjacent zones are often imprecise in a travel demand model; the model can indicate the level of short trips, but the level of precision is low.

In other words, the crime travel demand model is good at capturing major travel patterns over a large area and not very good at localized travel. There are other modeling tools for small area travel analysis that provide much more detail about the neighborhoods and road system in which this travel occurs, such as microsimulation software of travel behavior in a neighborhood (Kitamura, Yoshii, & Yamamoto, 2009; Miller & Salvini, 1999).

Therefore, in order to apply a travel demand model to crime analysis, it is important to model a substantial part of a metropolitan area. The model will not be as accurate if a small city or area within a metropolitan area is chosen. In these chapters, crime travel in Baltimore County is used as an example case in order to illustrate the different components of the model. Baltimore County is a large jurisdiction covering approximately 640 square miles; it represents a sizeable part of the Baltimore metropolitan area. Combining the origin zones of Baltimore City with those of Baltimore County provides a very large proportion of the metropolitan area. In

other words, Baltimore County is large enough to model the crime destinations while the origin zones represent much of the metropolitan area.

On the other hand, if we attempted to apply the model to a small part of the region, for example the town of Towson, the model would be less precise and less accurate since that town represents a very small proportion of the overall region. In short, a crime travel demand model is useful for modeling either an entire metropolitan region or a sizeable part of a metropolitan region, but should not be considered for a small geographical area. It is a regional travel model, not a local model.

Calculation Limits for the Number of Zones

A final consideration has to do with the number of zones that can be modeled with the *CrimeStat* crime travel demand model. Depending on whether a computer is 64 bits or 32 bits and depending on the operating system, limits may be reached on the number of zones. For example, with a Windows 32 bit operating system, the routine can only access 4 Gb of RAM. If M is the number of origin zones and N is the number of destination zones, then a trip distribution matrix, which is subsequently used in the mode split and network assignment stages, involves N*M cells. Each digit requires 64 bits of RAM with 16 digits assigned per cell. There are also seven fields output. Thus, a trip distribution output file requires approximately M*N*64*16 bits of RAM.

To use an example, if the user has 1 Gb of RAM available, then approximately 8,388,608 grid cells could be handled (or a square matrix of 2,896 x 2,896). However, Windows requires some overhead as does *CrimeStat*. Thus, the actual number of grid cells that could be processed will be a little less.

One could, of course, add more RAM. In this case, the file size of the trip distribution matrix could be increased. However, there are limits to this. First, the calculations will slow down, at a rate that is exponential to the file size. At some point, the calculations will take so long as to be impractical.

Second, as mentioned a 32 bit operating system a 4 Gb limit. Thus, the maximum file size would be a square matrix of about 5,793 x 5,793. A 64 bit operating system, on the other hand, can access 32 Gb of RAM thus allowing about 268 million cells (or a square matrix of 16,384). Clearly, if the study area has many zones, then a 64 bit computer and operating system will be essential. But, even here there are limits. For example, in Chicago there are 21,068 blocks. Using these blocks as a zone model in the crime travel demand would be impossible even for a 64 bit computer since the matrix routines could not handle such a large matrix, even assuming that it is desirable to do so. Therefore, any zonal model that is selected must be

compatible with the calculation limits of the available RAM and the Windows operating system. In the case of Chicago, using block groups was an acceptable choice since there are only 2400.

Obtaining Crime Data

There are four types of data that need to be obtained.

Crime Data by Origins and Destinations

First, there is crime data. But, in order to estimate a crime trip, it is essential that these data have information on both crime origins as well as crime destinations. The most likely source of these data will be arrest records whereby both the crime location and the charged offender's residence are given. Only the police are liable to have these data. Thus, it will be necessary to obtain cooperation from the local police department for access to arrest records.

In the data, the residence location is taken as the origin while the crime location is taken as the destination of the trip. As mentioned in Chapter 13 on journey-to-crime, the "true" origin of the crime may not be known. First, the offender may not even have been living at the same residence as when arrested. Many offenders are highly transitory persons and a residence at the time of the arrest may not be the actual one from which the crime occurred. Second, the offender may not have traveled directly from home to the crime location, but may have committed the crime as part of his/her daily activities (intermediate trips). However, without any alternative data on the actual origins, there is little that can be done except assume that the residence when arrested is the origin. As long as this definition is kept, a consistent estimate can be obtained.[2]

In effect, one is asking the question, "What is the likelihood that an offender who lives in zone i will commit a crime in zone j at some point during a day?" It really does not matter whether the offender traveled from the home location to the crime location as opposed to going to the crime location from an intermediate location. The model is simply constructed with respect to residence location.

The data has to be organized so that the X and Y coordinates of both the residence location (the origin) and the crime location (the destination) are given. Figure 26.1 illustrates a

---

[2] If the actual origin was an intermediate location between the home and the crime location, then with a large sample of crimes and offenders the idiosyncrasies of one offender's crime travel pattern is not going to affect the coefficients of the prediction model to any great extent. If *all* offenders from a particular zone committed crimes from an intermediate location which was always the same, then that condition might affect the coefficients (assuming one could obtain the data). But, it is highly unlikely that all offenders will commit crimes in the same destination zone using the same intermediate zone as an origin.

typical data set. It will be necessary to geocode both locations in order to establish a 'crime trip', an assumed trip from a particular origin location to a particular destination location.

Figure 26.2 shows the location of 41,974 crimes committed in Baltimore County between 1993 and 1997 while Figure 26.3 shows the assumed origin location of the offenders who committed these 41,974 crimes. As seen, the origins are all over the region, but most (96%) are in either Baltimore County or Baltimore City. In other words, a 'crime trip' links the origin location of each crime with the actual destination where it occurred. If arrows were to be drawn from the origin to the destination, the entire map would be swamped with a series of lines.

Choosing a Zonal Model

The zonal framework used for the Baltimore County analysis was traffic analysis zones (TAZ). The reason for selecting theis was the availability of both population and employment data. The Baltimore Metropolitan Council is the Council of Governments and the Metropolitan Planning Organization for the greater Baltimore region. They use TAZ's for their transportation model. Since data were available by the TAZ's, it seemed like a plausible decision. But, as mentioned above, there are advantages and disadvantages to this decision. Approximately, 20% of all crime trips occur within the same zone (intra-zonal trips). Such a high proportion makes the overall model estimates prone to some error. Figure 26.4 shows the TAZ's for both Baltimore City and Baltimore County.

Note that there is a difference between the zones used for the origins and the zones used for the destinations. In the case of Baltimore County, there are 325 TAZ's that cover the County. However, as mentioned above, since many of the crimes occurring in Baltimore County originate in the City of Baltimore, the origin zones include those of the City as well as the County. Thus, there are 532 origin zones.

Assigning Crime Events to Zones

The next step involves assigning the crime origins and the crime destinations separately to the zonal model. That is, each crime event is assigned to zones twice, once for the origins and once for the destinations. Since an arrest record is an implicit crime trip, the residence location is assigned to a zone and the destination location is assigned to a zone. Then, the number of crimes originating from each zone are calculated by summing over all records to produce a distribution of crimes by origin zone. Similarly, the number of crimes ending in each zone are calculated by summing over all records to produce a distribution of crimes by destination zone. The result is two distributions of crimes by zone, one for origins and one for destinations.

Figure 26.1:

Crime Data Requirements

Minimum data requires origin and destination location

UCR	DATE	INCIDX	INCIDY	HOMEX	HOMEY
430	1/5/97	-76.8131	39.3822	-76.8131	39.3822
440	5/17/95	-76.4490	39.3355	-76.4489	39.3355
210		-76.4068	39.3388	-76.5281	39.3085
210		-76.4142	39.2801	-76.4142	39.2801
430		-76.5527	39.3908	-76.4410	39.3080
440		-76.7581	39.3131	-76.7709	39.3105
440	3/29/94	-76.5095	39.2735	-76.5095	39.2735
440	1/22/96	-76.7344	39.3212	-76.6899	39.3364
690	7/13/93	-76.4525	39.3012	-76.6050	39.3020
690	10/8/94	-76.5278	39.2584	-76.5051	39.3970
690	8/10/97	-76.7384	39.3275	-76.7384	39.3275
690	3/10/96	-76.7325	39.3018	-76.7325	39.3018

Figure 26.2:

Baltimore County Crime Locations: 1993-1997
Location of Crimes Committed by Offenders (N=41,974)

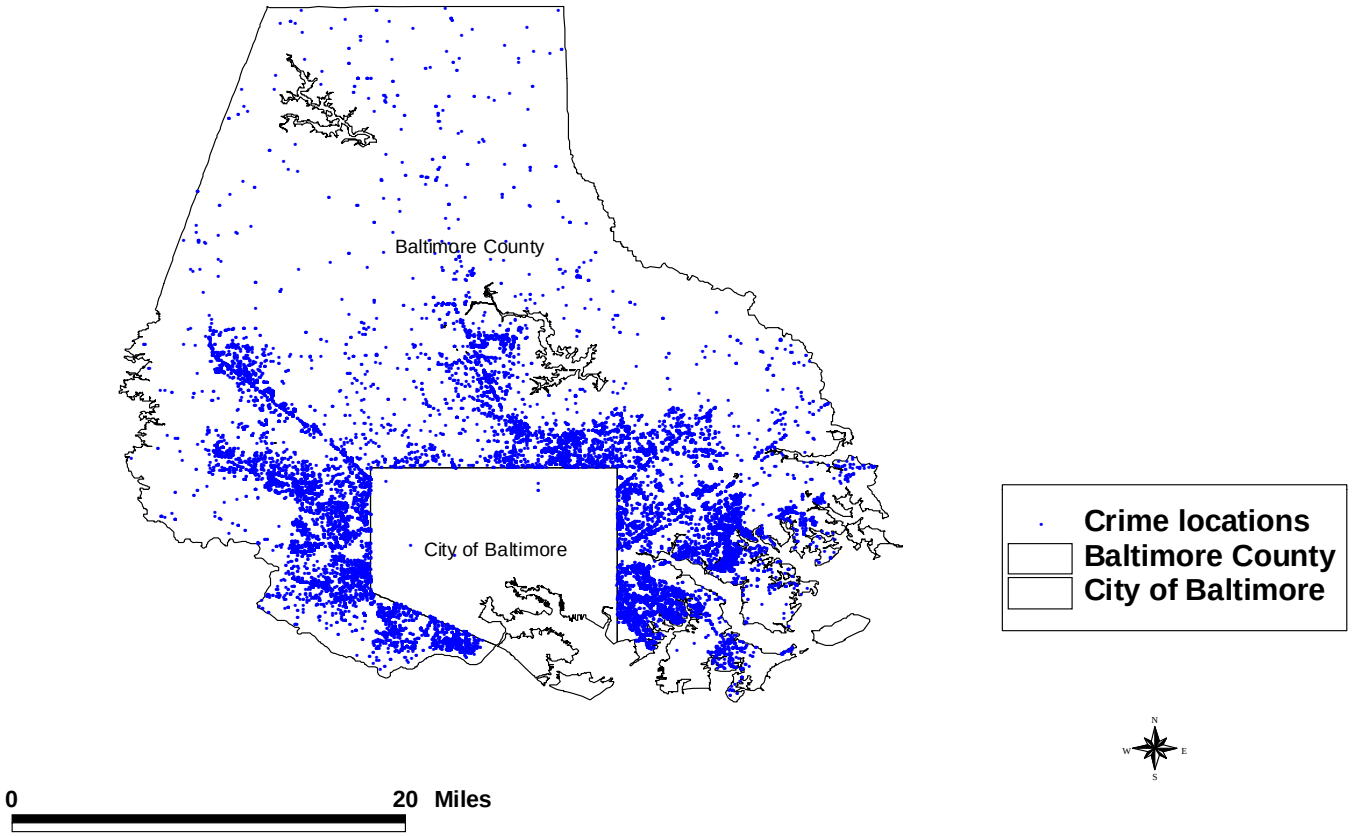


Figure 26.3:

Baltimore County Offender Residences: 1993-1997
Location of Offenders When Arrested (N=41,974)

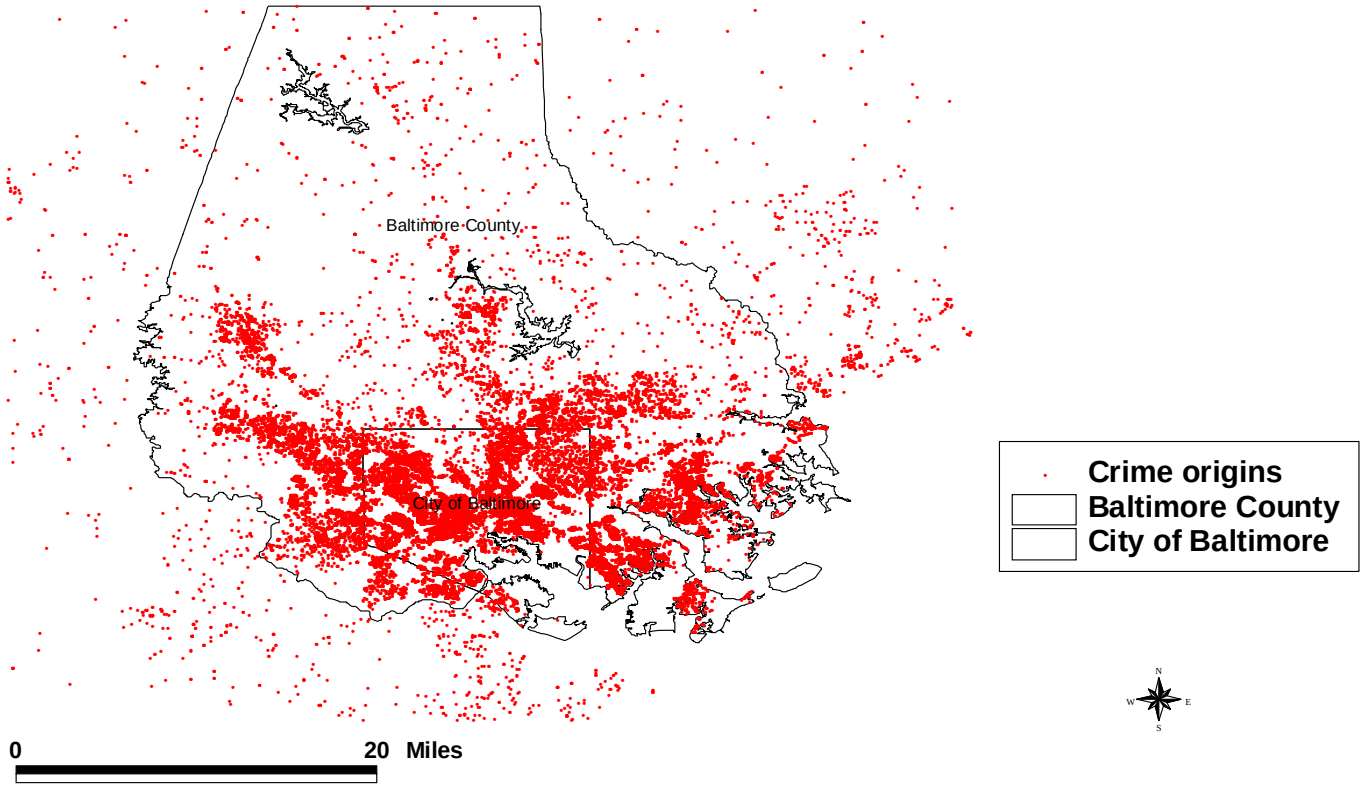
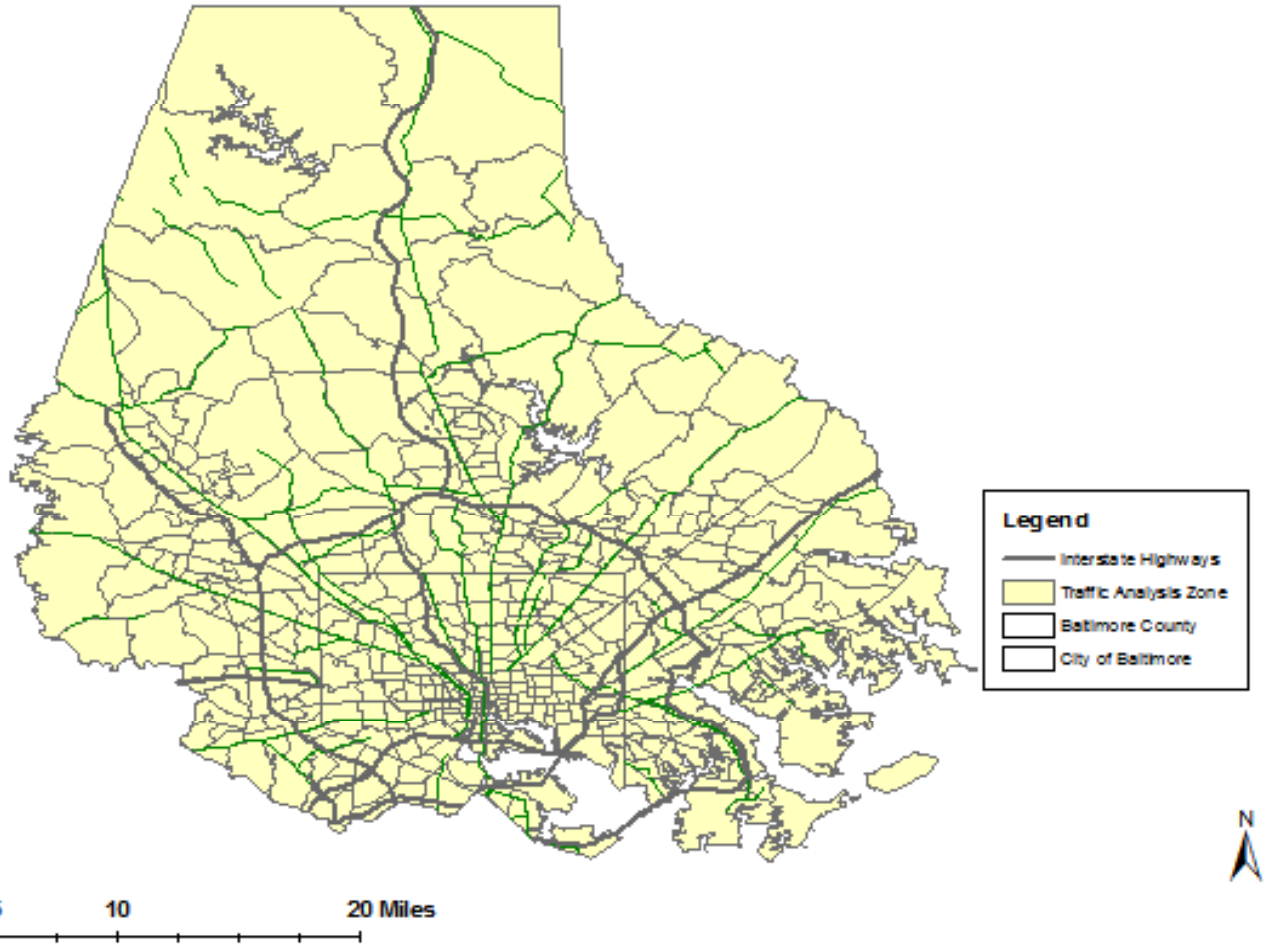


Figure 26.4:

Metropolitan Baltimore Traffic Analysis Zones: 1998



How does one assign crime events to a zone? There are two general ways to do this:

1. Nearest zone centroid - events are assigned to the zone centroid that is closest.

2. Point-in-polygon - events are assigned to the polygon within which it falls.

With the nearest zone centroid method, an incident is assigned to a zone to which it is closest whereas with the point-in-polygon method, an incident is assigned to a zone in which it falls within the boundary of that zone. Most GIS packages have a point-in-polygon routine and can implement that method.

In *CrimeStat*, on the Distance Analysis I page, there is an Assign Primary Points to Secondary Point routine that will make this assignment based on either method (see Chapter 6). In both cases, the incident file must be the primary file and the zonal file must be the secondary file. In the nearest zone centroid method, the routine will assign each event to the centroid to which it is closest. It will then sum the number of incidents assigned by zone and will add this as a new field to the secondary file (called *Freq*).

In the point-in-polygon method, the user must also provide the boundary file for the zones as an *ArcGIS* shape file. The routine will read the boundary file and will determine in which polygon an incident falls, and will then assign the incident to that zone. As with the nearest zone centroid method, it will then sum the number of incidents assigned by zone and will add this as a new field (*Freq*) to the secondary file. Chapter 6 presents details of these two routines, and is not repeated here.

There are advantages and disadvantages to each method. The nearest zone centroid has attributes that are probably closest to the location where the incident occurs. This is important in relating socioeconomic and land use characteristics to the events during the trip generation stage (see Chapter 27). Typically, social characteristics change gradually over an urban landscape so that an incident is probably closer to its nearest zone centroid than to any other zone centroid. In the case of the point-in-polygon method, incidents are not necessarily assigned to the nearest centroid since zonal polygons are frequently irregular in shape. Thus, to represent the underlying characteristics of the location in which the incident occurs by a point-in-polygon may end up assigning an incident to a zone that is quite different from where it should be located.

On the other hand, the main advantage of a point-in-polygon assignment is if the zone has a meaning in terms of containment or membership. For example, if a police reporting district (which could be a sub-set of a larger police precinct) is used as the zonal model, assigning incidents to the reporting district within which they fall will ensure that the incidents are assigned to the correct police precinct.

In other words, if it is important that events be assigned to the area to which they belong, then the point-in-polygon method is usually the best. On the other hand, if it is important that the incidents be assigned to the zone to which they are most similar, then the nearest centroid method is usually the best.

For Baltimore County, figure 26.5 shows the number of crimes by origin zone while Figure 26.6 shows the number of crimes by destination zone. In both cases, events were assigned by the nearest zone centroid method.

Adjusting Crime Events Estimated from Arrest Records for Accuracy

There is another subtlety that affects the assignment to a zone. The method that has been described assigns records in which there is both an origin and a destination location, such as an arrest record. The reason for doing this is that there is an implied trip between the origin and the destination, as was discussed above and in Chapter 25. However, there may be a difference between the distribution of crimes by destination from the arrest records and the actual distribution of crimes from all incidents. The reason is that arrest records represent only a sub-set of all the crime records and, often, a small sub-set. If there are any spatial differences in the arrest likelihood across a metropolitan area, it is possible that some areas will have a higher proportion of offenders being arrested than other areas. The result would be a discrepancy between the distribution of crimes by arrested individuals and the actual distribution of crimes. In other words, the distribution of crimes as identified by the arrest records could be a biased estimate of the actual distribution of crimes. The result could be that the origins of those offenders who were caught will be exaggerated relative to the origins of those offenders who were not caught, and the entire model could end up being biased.[3]

If there is a sizeable discrepancy between the distribution of crimes from the arrest records and the actual distribution of crimes, it is important to correct this. In the Assign Primary Points to Secondary Points routine on the Distance Analysis II page, it is possible to weight the assignment by another variable. This variable can reside on either the secondary (zone) file or on another file. A typical correction weight variable would be a proportion that adjusts the empirical distribution of crime destinations by the true distribution. Thus, a weight greater than 1.0 would increase the proportion whereas a weight smaller than 1.0 would decrease the proportion. A weight of 1.0 would maintain the same proportion.

---

[3] In the usual travel demand modeling conducted by transportation planners, the origins are assumed to be more accurate than the destinations. The origins are identified typically from census and other population enumerations whereas the destinations are estimated from surveys and employment databases. In the case of crime travel, however, the destinations are known with much greater accuracy since those locations are documented in police reports.

Figure 26.5:

Crimes Origins by TAZ
Number of Crimes Originating in TAZ
Baltimore County: 1993-1997

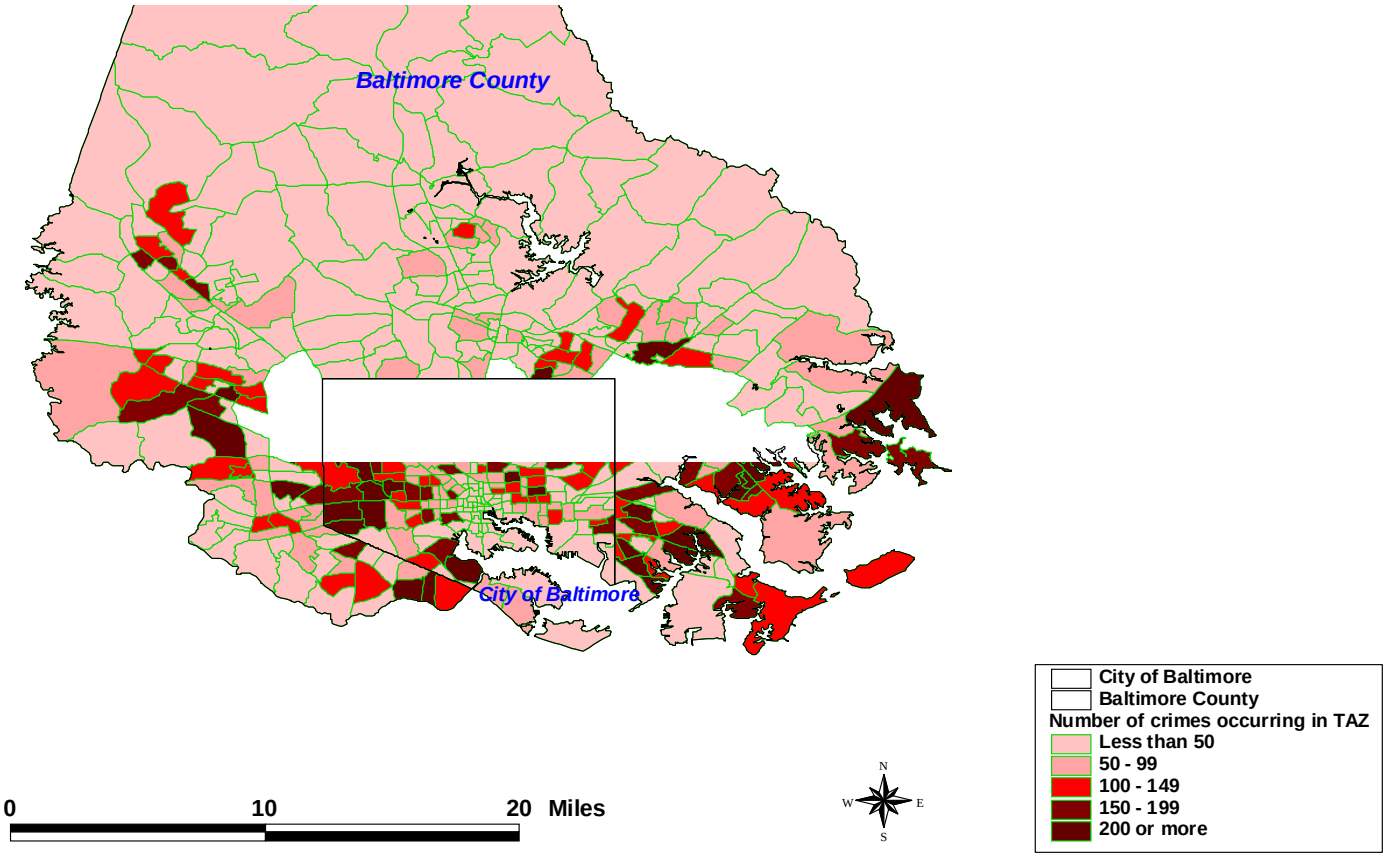
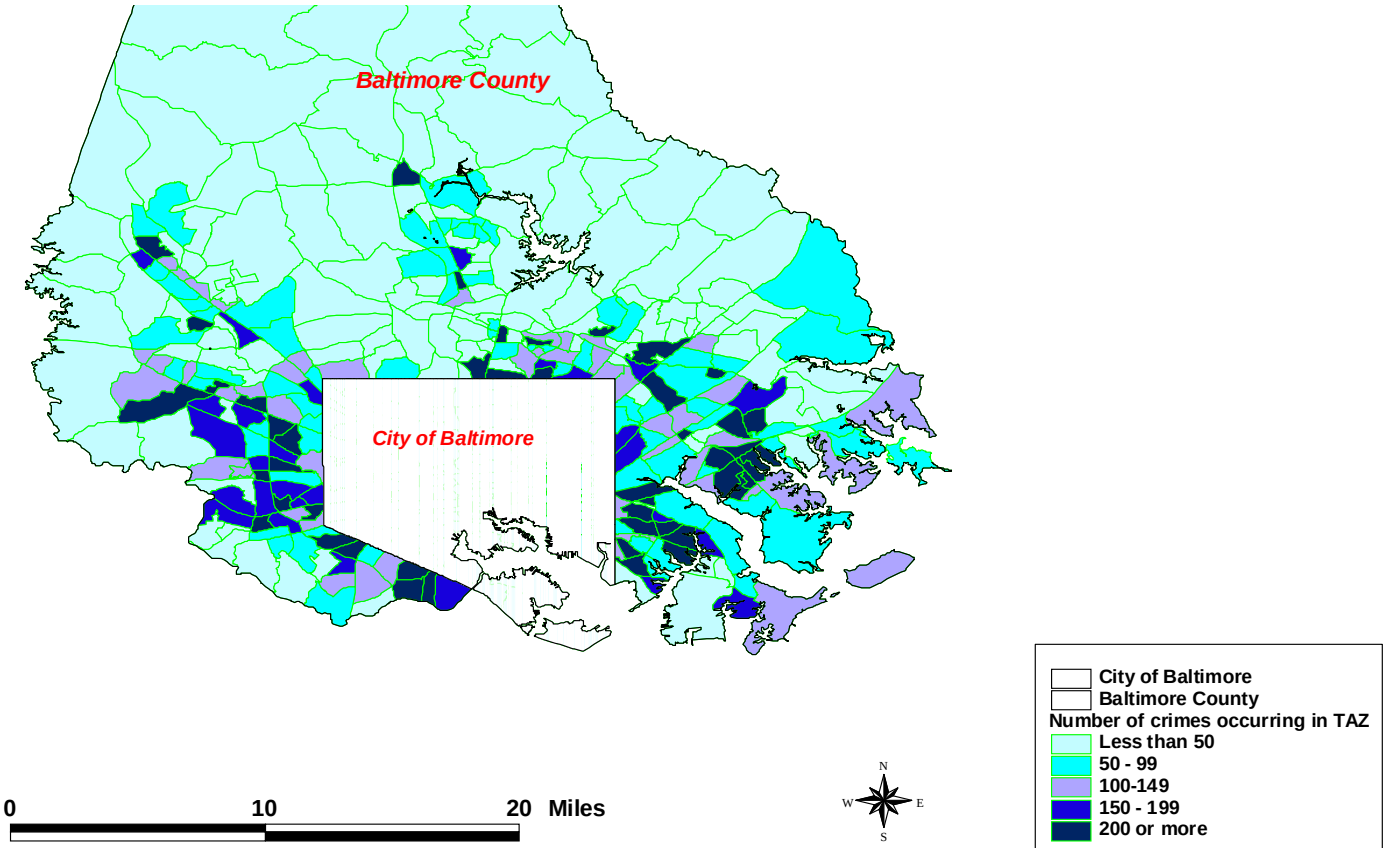


Figure 26.6:

Crimes Destinations by TAZ
Number of Crimes Occurring in TAZ
Baltimore County: 1993-1997



In order to do this, however, one has to convert the number of crime destinations into proportions. Let's take an example. Suppose the empirical and true distribution of crime destinations was as follows (Table 26.1):

Table 26.1:
Proportional Weighting Empirical Assignment of Crime Destinations

Zone	Empirical Distribution	True Distribution	Proportional Weight
101	.04	.05	1.25
102	.03	.025	0.83
103	.015	.015	1.00
etc.			

In the example, the actual (true) distribution of crimes for zone 101 is greater than what was measured in the incident-to-zone assignment by a factor of 1.25 to 1 (i.e., .05/.04). Thus, the weight assigned to zone 101 is 1.25. In zone 102, on the other hand, the actual distribution of crime destinations was smaller than what was estimated from the incident-to-zone assignment by a factor of 0.83. Thus, the weight assigned to zone 102 is 0.83. Finally, the proportion of crimes in the empirical and actual distributions for zone 103 is exactly the same. Thus, the weight assigned to zone 103 is 1.00.

The weight variable will be typically a column in the secondary file that corrects the empirical distribution. Naturally, the first time this is done, an analyst would probably not know the empirical distribution. Thus, it will be necessary to repeat the incident-to-zone assignment, the first time in order to count the empirical distribution while the second time to weight that count by the correction factor (which will have been added as a variable to the secondary - zonal, file). See Chapter 6 for a more complete discussion of weighting a primary points (incidents) to secondary points (zones) assignment.

Note, the adjustment of the empirical count (assignment) is done usually for the destination variable, not the origin variable. In the case of crime events, police will know the destination of the crime a lot more accurately than they will the origin since there is a crime record on file for the incident. Hence, any discrepancy between the empirical distribution of crimes and the actual distribution will only be known for crime locations (destinations). Therefore, in correcting the empirical distribution, we are assuming that we are also correcting the true distribution of origins, too. It should be obvious, though, that we really don't know. Unless one can obtain a "true" distribution of crime origins and, thereby, correct the origin distribution as well as the destination distribution, one has to assume that the adjustment in the destinations will

also correct the distribution of the origins during the balancing stage (see Chapter 27 on trip generation).

Obtaining Crime Data by Sub-Types

Till now, the discussion has focused on the total number of crimes that occur within a zone. Clearly, it is possible (and preferable) to break this down into distinct sub-groups. Thus, a separate distribution for robberies, burglaries, vehicle thefts, homicides, and other crime types can be complied. In each case, the separate distribution is being assembled in order to produce distinct models of crime travel by that type. The journey-to-crime literature has long illustrated the differences in travel distance by crime type and it would be expected that there are substantial differences in travel patterns as well. Most crime analysts and researchers will want to break down crimes into these distinct categories. Similarly, an analysis by time of day or day of week also would require breaking down crimes by these different temporal categories. In general, an analysis of all crimes is not very meaningful for most police departments. Instead, the focus has to be on crime types and, perhaps, times of day with other sub-sets also being important (e.g., method of operation, use of weapons).

The method used to assign these individual crimes to zones would be, however, exactly the same as for the total number of crimes that was illustrated above. As with the total number of crimes, there would be differential weighting of zones in order to correct any bias in the distribution of crimes calculated from the arrest records compared to the actual distribution of incidents as identified by total crime reports.

Adequate Sample Size

A problem with this approach arises, however. By breaking down crimes into distinct sub-groups (by crime type, time of day, day of week, method of operation, etc), smaller samples are produced. As the sample size decreases, the likelihood of modeling error increases. If the sample is too small, then any of the zonal estimates that are produced in the trip generation stage will be subject to considerable sampling error. Similarly, in subsequent stages (trip distribution and mode split), these small sample sizes are further broken down into cells with very small sample sizes, with most having zero incidents. In other words, sampling error becomes a problem if the total number of crimes is broken down into very small sub-sets, and the model becomes unreliable.

How would one know whether a model is unreliable or not? Probably the simplest way is to repeat the model on two different years worth of data. That is, the analyst constructs the travel demand model on one year's worth of data and then repeats it on another year. If the variables selected during the trip generation stage are the same and if their coefficients are approximately

equal, then the model would appear to be reasonably stable. On the other hand, if there are substantial differences in the selected variables and in their coefficients, most likely the data set was too small for the construction of a stable model. One could do formal tests on differences between the coefficients to see whether they are similar or different. But, a general review of the coefficients should indicate whether there is stability or variability. There is not a 'hard and fast' rule since any differences could be due to real changes in the environment creating crime. But, unless there is some obvious explanation for the differences, most likely they indicate that a model is too unreliable to be used from one year to the next (i.e., the sample size is probably too small).

Thus, there is a balance that has to be maintained between having a large enough sample to produce reasonably reliable trip generation and trip distribution coefficients, and breaking down the data into more meaningful categories for analysts and researchers. In general, I believe it is a good idea to model all crimes first before modeling specific sub-types. The reason is to establish baseline characteristics - variables and coefficients. It will become easier to understand how different crime sub-types vary once the overall distribution is known.

Developing a Predictive Model

The above discussion dealt with summarizing crime incidents by zones, both the location where the crimes occurred (the destinations) as well as the locations where the offender was living (the assumed origins). In order to develop a predictive model of crime origins and destinations, it is also necessary to put together a data set of predictive variables. Typically, these will be socioeconomic and land use variables, though other types of variables can be included.

Obtaining Socioeconomic Data

Population

The most common type of predictive data will be socioeconomic variables. Among these are population, employment, income levels, poverty data, and household characteristics. At the minimum, population will be an important variable. As mentioned in Chapter 25, the crime travel demand model is an aggregate (volume) model. That is, it counts the total number of crime trips (by origin and by destination). Since, the number of trips is generally a function of the total number of persons living in a zone, all other factors being equal, population inevitably will enter as either the most important or among the most important variables, as both an origin and a destination variable.

Population could be measured by sub-sets (or proxy) variables, too. For example, the number of households, the number of teenagers, and the number of married couples are also sub-

sets of the total population; the correlation among these variables is usually very high. Which variable is chosen will depend on what type of crime is being predicted. For the total number of crimes, probably the total population (or total number of households) should be used because it is a larger and more stable estimate of the total "at risk" population. For specific crimes, however, it may be desirable to choose a sub-set of population. For example, for car thefts, the distribution of males, ages 16-30, might be a more intuitive baseline variable as those age groups contribute disproportionately to vehicle thefts (as they do to most crime types). The disadvantage in using this variable may be the smaller sample sizes that are obtained for some zones. A good way to test this is to model it twice, once with total population and once with the sub-set variable. If the overall predictability of the model is about the same (or, better, if the sub-set variable predicts better than the total population), then the use of the sub-set population will be preferable to the total population. On the other hand, if there is not much difference, stick with total population as it is a larger, and more stable, variable.

Employment

A second variable that usually comes up is total employment. This is particularly valuable as a predictor of crime destinations since many crimes are attracted to employment areas (e.g., robberies, burglaries, vehicle thefts). Usually a distinction is made between *retail* and *non-retail* employment, though other distinctions can also be made (e.g., office employment, government employment, military employment). The reason is that retail employment is usually found in commercial areas (e.g., shopping malls, strip malls, retail centers). In the case of Baltimore County, for example, retail employment is the strongest predictor of crime destinations.

As an origin variable, too, employment could be important. In the three models that were compared for this version of *CrimeStat* (Baltimore County, Chicago, Las Vegas), employment was seen as a predictor variable for crime origins in several cases, too, usually as a negative predictor (i.e., less employment is associated with more crime). The reason may be less clear, but may have to do with the lack of opportunities in certain districts and neighborhoods.

Income levels

Another obvious variable is income measured in some way. The relationship between crime and low income has long been noted. There are several possible income-type variables that could be used in a model. The most obvious is the total income level of a zone. The U. S. Census Bureau has a total income variable that is part of their SF 3 release (U.S. Census Bureau, 2011a). This measures the total of all household incomes in the census. While this variable captures the total available income in the zone, it is not a very intuitive measure. Consequently, other measures are usually used, such as income per capita or median household income. Median

household income is usually a more typical measure since the average income per person can be affected by extreme values.

An important issue about income levels, no matter how measured, is that they inflate over time. That is, since income reflects monetary value at any one point, it does not have a fixed reference point. What this could mean in a model is that, over time, income levels will increase (in absolute terms) due simply to inflation. A model that established, for example, a negative relationship between income and crime (i.e., the higher the income of the zone, the less crime) for one year would end up predicting lower crime levels for another year simply due to inflation.

It is important to standardize income in order to prevent the impact of inflation affecting the model. There are two ways that this is usually done. First, one can standardize income by subtracting the mean and dividing by the standard deviation. That is,\

$$Z_i = \frac{I_i - \bar{I}}{SD_I} \quad (26.1)$$

where I_i is the income of each zone, i, $\bar{I}$ is the mean income of all zones, and Sd_I is the standard deviation of all zones. This is a classic standardized measure.

A second way to standardize income is to define *relative* income. That is, the income level of each zone is compared to the income level of the zone with the highest income. That is,

$$I_i = \frac{I_{max} - I_i}{I_{max}} \quad (26.2)$$

where I_{max} is the income level of the zone with the highest income. This index measures the income of a zone relative to the income of the highest income zone. The closer the income level of the zone is to the highest income zone, the smaller the index. Thus, this is an *income inequality index*, similar to the Gini index though more simply calculated. The zone with the highest income will have a value of 0 whereas the zone with the lowest income will have a positive value roughly reflecting the relative differences in income levels between the lowest and the highest.

Each of these measures will prevent a shift in the predicted values due to inflation, though they each measure slightly different attributes; the first measures just absolute income levels (standardized) while the second measures the degree of inequality.

Another type of income variable is the number of persons living under poverty. Again, the relationship between poverty and crime has long been noted (Bursik & Grasmick, 1993).. Thus, a variable that measures poverty directly could add sensitivity to a model that simple

income might not detect. The issue of measuring poverty, however, is a complex one. Different government agencies use different measures. For a discussion, see Citro and Michael (1995).

In general, typically the variables 'median household income' and the' number of persons (or households) living under the poverty line' do correlate quite well. Therefore, it is unlikely that both variables would be significant in a regression equation without, essentially, measuring the same thing. The same is true for education and income, which tend to correlate quite highly. Again, both variables in a regression equation would, essentially, be measuring the same thing. Thus, in a regression model, it is important to select only the strongest and most stable income variable in order to avoid duplicate measures (multicolinearity). I will return to this point in the next chapter.

Other socioeconomic variables

Other socioeconomic variables might be useful in a predictive model. Among these are race or ethnicity, vehicle ownership, number of single parent households, number of unemployed workers, number of persons living in large rental buildings, and others. Again, these variables might produce greater differentiation in a model. But, at the same time, they tend to overlap with income variables and may be measuring the same thing.

Obtaining Land Use Data

Aside from socioeconomic variables, there are land use variables that could be important in predicting both crime origins and destinations. Among these are parks, bars, pawn shops, check cashing businesses, the location of shopping malls, retail space, stadiums, train stations, intra-urban metro stations, bus stations, parking lots, hospitals, and adjacency to major freeways or arterial roads. There are a wide variety of land use variables that appear to be important in attracting crime as well as in providing an environment that may encourage people to commit crimes. A thorough elaboration of potential land use variables would help to identify particular attributes associates with crime and, thereby, increase the predictive ability of a model.

There are two ways to document these land uses. One is as a simple categorical ('dummy') variable whereby the field is given a '1' if that land use exists in the zone and a '0' otherwise (e.g., there is a park in the zone; a freeway runs through the zone; there is a stadium in the zone). The second is a count of the level of that land use variable (e.g., the number of bars; retail square footage; park acreage; number of parking stalls in a parking lot). The second variable is, clearly, more precise than the first, but is much harder to document. The availability of data will be a constraining factor in building up a set of land use variables that might predict crime origins or destinations.

Still, before an extensive data inventory is initiated, some cautionary words are in order. In the three studies illustrated in this version of *CrimeStat*, however, few land use variables survived once population, employment and income levels were included. The reason is that many land use variables correlate with these basic variables (e.g., the amount of retail space correlates with retail employment; bars correlate with low income). Thus, in spite of intuitively being related, it was found that most of the land use variables did not improve the models beyond the basic variables.

Special Generators

There are exceptions, however. Particularly, there are *special generators* that attract crimes out of proportion to the amount of employment at those locations. Among these are stadiums, major train stations, airports, and large parks. Because these are major regional facilities and, in the case of stadiums and parks, used only periodically, they may attract more crimes that would be expected on the basis of the level of employment at those locations. Traditional travel demand models have incorporated these as special variables because they can account for variability that is not general throughout the study area. In the next chapter, I will discuss this in more depth.

Spatial Location Variables

Centrality

In addition to socioeconomic and land use variables, spatial location variables *might* be relevant. There are two types of spatial location variables that might be relevant. The first is the *centrality* of the metropolitan area. In most American cities, the central downtown area has a uniqueness that is greater than that which is explained by any one variable. For example, not only is there a large amount of employment in most Central Business Districts (CBD), but there are amenities that are associated with a central location. Usually, there is a greater concentration of restaurants and stores in CBD's and other employment centers. Entertainment activities are often more concentrated in the CBD; this is not true in many large metropolitan areas (e.g., Los Angeles), but it is true in enough of them to make the CBD an entertainment center as well as an employment center. Similarly, transit lines tend to concentrate in the CBD.

In other words, the CBD is a unique place that affects crime trips. Some CBD's have a large number of crime incidents whereas others do not. Nevertheless, measuring it in a predictive equation *might* increase the predictability of a production or attraction model. A simple variable is the distance from some point within the CBD, for example distance from the City Hall. Zones that are close are liable to have a greater number of crime productions and crime attractions,

especially, than zones farther away. This type of spatial effect is very similar to the *first-order* effect described in Chapter 6.

The use of a distance from the CBD variable can usually strengthen a regression model. A study which illustrates how centrality predicts male-female differentials in motor vehicle crashes in Houston is by Levine (2011); male drivers are much more likely to get involved in crashes in the central city, particularly the CBD, than females whereas the differentials are much less in the suburbs. Another example is by Levine and Canter (2011) who showed that distance from the CBD predicted positively the number of DWI trips that ended in crashes that originated in each zone in Baltimore County (i.e., zones farther from the CBD produced more). Similarly, Levine (2007) found that distance from the CBD negatively predicted the number of bank robbery trips that originated in each zone (i.e., zones closer to the CBD produced more).

Local spatial autocorrelation

The second type of spatial effect is a localized similarity between adjacent zones. In other words, there frequently is spatial autocorrelation in crime productions or attractions between adjacent zones. These are the *second-order* spatial effects described in Chapter 6. Zones that have a lot of crimes occurring within them are frequently located next to zones that also have a lot of crimes occurring, and the converse.

If the user wants to incorporate local spatial autocorrelation explicitly in the trip generation stage, then the use of a Anselin's Local Moran, the Getis-Ord Local 'G' (see Chapter 9) or a simple adjacency measure (e.g., '1' if the average of adjacent zones is greater than the mean for all zones and '0' if it is not) may be sufficient in account for the localized spatial autocorrelation.

However, it should be noted that apparent second-order spatial effects may be simply by-products of first-order spatial effects. Because of the concentration of events in the central city, there are usually more local hot spots in the central city, too. Before arriving at a conclusion that there is definite local spatial autocorrelation, a user would be wise to incorporate a first-order global spatial autocorrelation variable, such as distance from the CBD. If there is additional variability after that in incorporated, then the local effect would most likely be rea.

Estimating spatial effects

The spatial regression models discussed in Chapter 19 explicitly incorporate spatial effects as a predictor variable. If a trip generation model includes both a first-order spatial effect (e.g., distance from the CBD) and a local spatial autocorrelation adjustment for each case (the Phi

coefficient to use the terminology of Chapter 19), then the model will handle both types of spatial autocorrelation.

An alternative is to ignore the spatial effects in the first stage – trip generation, since the second stage of the model - trip distribution, incorporates an explicit spatial component by weighting distance in estimating the interaction between zones. Thus, any spatial error produced during the trip generation stage is frequently compensated for during the trip distribution stage.

There are advantages and disadvantages to including first- or second-order spatial effects in a travel model. Since the trip distribution stage has an explicit spatial interaction term, any errors from the first stage (trip generation) are usually accounted for during the second stage. Thus, there is little advantage to be gained from including a second-order (spatial autocorrelation) variable. However, including a first-order variable can usually improve the predictability of the trip generation model.

Defining Policy or Intervention Variables

Aside from socioeconomic and land use variables, a model might include some policy or intervention variables. One of the best uses of a travel demand model is to model the likely effect of a change in one of the predictive variables. A simple one would be the likely effect of building a new facility, for example a shopping mall. In the estimation stage, if the analyst can show that shopping malls are associated with higher (or lower) numbers of crime occurring, then a theoretical mall could be placed in a zone and the model run with that as a new input for the zone (with every other variable being the same for all zones). Since the travel demand model is sequential, the impact of new crime trips being attracted to the zone can be followed through the different stages of the model.

There may be other policy or intervention *experiments* that can be conducted with a crime travel demand model. In each case, it is necessary to include the variable in the estimation model to establish a coefficient for it. Then, in the simulated experiment, the variable is re-arranged or allocated differentially and the model is recalculated. Again, the result can be used to estimate what the likely effects of the intervention could be on crime travel patterns.

Among the possible policy or interventions are the construction of a particular type of facility (as mentioned above with a new shopping mall), changing the level of policing in a zone, the creation of a drug treatment center, the establishment of a job retraining center, or the reduction in the number of adult book shops. There are a large number of possible interventions that might affect the level of crime - either produced (origins) or attracted (destinations). Further, not all of the interventions might reduce crime levels, but some could even increase it (e.g., add new shopping malls). Nevertheless, the ability to add interventions in the model makes it a useful

device to estimate the likely effects on crime levels without having to actually implement the changes.

In the three studies presented in this version of *CrimeStat*, there were no interventions that were estimated. Examples of simulated interventions can be seen in Levine and Canter (2011) who modeled selective police interventions to reduce DWI trips in Baltimore County that end in crashes from zones where a higher proportion of offenders resided and from zones where a higher proportion of crashes occurred and Levine (2007) who modeled both bank robbery trips in Baltimore County from residence to the bank and the escape route trip back to the residence. Still, this type of experiment or 'variable' is an important one and which could make the crime travel demand model a very powerful analysis tool.

Where to Obtain these Data?

Many of these data are easily found while other data are more difficult to locate. A lot of socioeconomic data is available in the decennial census and distributed by the U.S. Census Bureau. Data on population, households, and income levels can be obtained from the Census Bureau for geographies as small as blocks or block groups. One of the deficiencies of the census data, however, is the lack of information on employment.

An alternative is to obtain data from a Council of Governments or Metropolitan Planning Organization. A Council of Governments (COG) is a regional association of cities and counties that is involved in planning; sometimes it is called an Association of Governments. Virtually every metropolitan area in the United States has a COG that can be a source of information on both population, employment, and, occasionally, land use. Many COG's have a forecasting group that estimates both population and employment, sometimes for very small geographical units. The Houston-Galveston Area Council, for example, has an extensive database of all firms with 10 or more employees and updates this continually utilizing information on business permits, purchased lists from other organizations, and aerial photographs for identifying new commercial developments. They produce estimates of employment for small grid cells that are approximately 1000 feet on a side; however, these data are released only at the Traffic Analysis Zone (TAZ) level. For more information and a detailed list of local regional councils, see NARC (2012).

A Metropolitan Planning Organization (MPO) is a regional transportation planning agency. In many metropolitan areas (e.g., Los Angeles, Houston, Washington, DC), the MPO is part of the COG while in other metropolitan areas (e.g., San Francisco, Chicago), it is not. They will obtain both population and employment data for the TAZ's as part of their travel modeling functions. For more information and a detailed list of local MPOs, see AMPO (2012). In short, it is generally possible to obtain data on population and employment from either COGs or MPOs.

Land use data is more difficult to obtain. Simple information can often be obtained from Yellow Pages or online business directories, for example the location of bars and nightclubs. More detailed data may have to be obtained from particular cities and counties. Generally, larger cities have a planning department or a public works department who maintains some land use data. The quality of this information will vary, however, and may not be consistent across jurisdictions. In a large metropolitan area, it may be possible to obtain regional land use information from the COG, the MPO, regional utility companies, a database of business permits, tax assessors' offices, or even the Army Corps of Engineers.

The point that has to be realized is that a lot of effort is needed to put together a data base for modeling crime travel. Once developed, however, it can be used repeatedly as predictors for different types of crime and can be updated more easily. Like a GIS system, there is a substantial amount of effort 'up front' in order to build a model. But, once collected, the information can be very useful for a multitude of purposes.

Creating an Integrated Data Set

The information that has been collected - both data on crime origins and destinations as well as socioeconomic, land use and policy interventions, needs to be integrated into a single zonal model. That is, the data need to be allocated to zones, both origin zones and destination zones. The result will be ***two*** different data sets, one for crime origins and one for crime destinations. The origin data set will cover the origin zones while the destination data set will cover the destination zones. The same predictor variables, however, can be in both data sets as these variables could predict either origins or destinations, or both.

Allocating Data to Zones

There are two steps in assembling the data into two data sets. First, the data have to be allocated to the zonal system used. In some cases, these data may be easily available (e.g., obtaining population and employment data by TAZ's when the TAZ is the zonal unit used). In other cases, it may be necessary to allocate the data from one geographical zonal unit to another (e.g., from census block groups to TAZ's). GIS is a very powerful tool for allocating data from one "layer" to another. However, it has to be realized that errors will result from an allocation. For example, breaking up a larger zone into small sub-zones (e.g., breaking up a large census tract into four small grid cells) will lead to some error in the allocation. The GIS splitting routines usually assume that the data are split proportionately between the four 'pieces'. Thus, if employment from a census tract is allocated to two grid cells, one assumes that the workers are uniformly distributed within the census tract and the two grid cells will each capture a share equal to their area relative to the larger tract. This may or may not be true. Where it is not true, adjustments need to be made to ensure that zones represent relatively uniform populations.

The point is, there is error in allocating data from one type of unit to another, and the analyst has to be aware of these potential sources. It is generally better to obtain data at the smallest possible geographical unit in order to minimize the splitting problem described above. Aggregation usually causes less error than splitting. On the other hand, as mentioned at the beginning of this chapter, the larger the zonal unit that is used the greater the likelihood that there will be within-zone (intra-zonal) trips.

Combining Data into Origin and Destination Data Sets

The second step is the combining of all the data into two separate data sets, one for origins and one for destinations. All the data that are used for the origin model should be together while all the data that are used in the destination model should be together. Many variables will be in both data sets (e.g., population, employment, income) whereas some variables only make sense as an origin or a destination variable (e.g., residential areas as an origin variable for bank robberies; a rail station as a destination variable for larceny or robbery). Since the origin zones will usually be more numerous than the destination zones (because they include the destinations and those from surrounding jurisdictions), the data have to be consistent across all zones that are used.

For use in the *CrimeStat* crime travel demand module, these data sets should be in one of the acceptable formats (Excel, dbf, dat, or ODBC-compliant). I have found that building the data first in a spreadsheet (e.g., Excel) is easier to do because variables can be more easily added. Once constructed, the spreadsheet is converted into a dbf file for use by *CrimeStat*.

Obtaining Network Data

The final type of data that needs to be obtained is a network. This is important for the third and fourth stages in the crime travel demand model - mode split and network assignment. In the mode split routine, trips from each origin zone to each destination zone are divided into different travel modes. For driving travel modes, travel has to go along a road network. For walking or biking, there may be additional segments that are not in the road network (e.g., bike paths, short cuts for pedestrians); these can usually be added to the road network to make a more realistic representation. However, for transit modes, the trips have to go along a transit route. In the network assignment routine, all zone-to-zone trips by each travel mode are assigned to particular routes. For this, a network is needed, one for each mode.

In both these cases, travel occurs along a network. That is, the distance (or travel time or travel cost) from one location to another is calculated using the network, rather than as direct or indirect distance. A network is a collection of segments that are interconnected. Travel can only occur on the segments. Each segment has two or more nodes and one or more connecting lines.

Travel is from segment to segment. Hence, the *end nodes* have a special status as the connectors which allow travel from one segment to another.

In Chapter 30, a more extensive discussion of the shortest cost/path algorithm used for network travel is explained. But, essentially, a 'trip' goes from the origin location to the closest location on the network. It then proceeds along the network, taking the shortest path, until it reaches a node closest to the destination. It then travels from that node to the final destination. Thus, the *representation* of the network is very critical. It has to be accurate and reasonably comprehensive.

There are three types of basic networks that need to be considered:

1. Road network (with additional walking or biking segments)
2. Bus network
3. Train network (if appropriate).

In addition, there can be specialized bicycle networks that are distinct from the road network. However, most transportation agencies model bike trips using the road network. I will discuss each of these.

Road Network

In a GIS system, there are typically two types of road networks that are used:

1. A bi-directional (or linear) network
2. A single-directional network.

Bi-directional road network

In a bi-directional network, travel can occur in both directions along a segment. A typical example is the TIGER system created by the U.S. Census Bureau (2011b). In this system, each segment typically represent the travel along a road from one intersection to another (i.e., a block in length), though there are exceptions. Travel can occur in both directions in the network unless there are special codes added to indicate a one-way street. The TIGER system, in particular, has a number of attributes associated with it - sides (left side, right side), address ranges (on both sides), census and political designators (again, by sides), and other attributes. This type of network is very common in GIS systems and is widely used in police departments. Because of the address ranges and because it is easily available from the U.S. Census Bureau or companies who improve the TIGER system, this type of network forms the basis of most geo-coding systems.

There are problems with a bi-directional network, however. Among these are the inabilities to distinguish direction and one-way streets. From a network modeling perspective, travel can occur in either direction. It is possible to put a field in the data base that identifies whether the street is one-way or not and to indicate the direction of travel. But, this has to be added by the user since the TIGER system does not specify that information.

A second problem is the lack of information about travel time or cost on the network. The only metric in the TIGER system are address ranges and, implicitly, distance. However, since travel varies substantially by type of road (larger functional classes have higher speeds) and by time of day due to differing levels of congestion, such a system lacks very important information for modeling travel. The TIGER (or similar) system does have functional class codes that distinguish different levels of road capacity (e.g., Interstate highways, state highways, principal arterial roads, collector roads, etc). It is possible to assign arbitrary average speeds to each of these classes (e.g., 45 miles per hour to an interstate highway; 30 miles per hour to a principal arterial; 20 miles an hour to a collector road; and so forth). By doing so, a reasonable approximation to actual travel can be obtained. However, there is still not a sensitivity to travel time by time of day. For example, in an urban area, travel at the peak afternoon 'rush hour' (e.g., 3:30 PM - 7 PM) will be, on average, a lot slower than at off-peak hours.

This brings up a third problem, namely that there is no interaction between the direction of travel and the travel time. On most principal arterial roads, travel is unequal in speed at any one time. For example, in many metropolitan areas, travel towards the downtown area is much slower in the morning than in the opposite direction, whereas the reverse is true in the afternoon. A bi-directional network cannot distinguish this and the analysts have to add multiple fields to the attribute file in order to make these distinctions (e.g., PM peak from node A to node B direction; PM peak from node B to node A direction; etc).

A fourth problem may or may not exist with a bi-directional network. These networks were designed to allow the U.S. Census Bureau to carry out the decennial census. Thus, a lot of attention has given to accuracy of streets and address ranges. Much less attention has been paid to the connectivity of the streets. A lot of the digitizing that goes into the network has been done by local governments, and the quality of this digitizing varies considerably. Some jurisdictions have very precise networks that are updated frequently while other jurisdictions have poorly defined networks that are often out of date. Drivers may know that they can travel from point A to point B via road C, but the network may not have been sufficiently updated to allow that trip to occur in a representation. In some cases, gaps between segments have been noted; the gaps may be very small, but they would prevent a model from 'traveling' from one segment to the next.

Single-directional road network

A single-directional network, on the other hand, separates travel in each direction. For example, if there are two nodes that connect a segment (node A and node B), then there are typically two segments for travel in each direction (from node A to node B, and from node B to node A). In this representation, a one-way street is simply a segment that does not have a reciprocal pair (i.e., there is only a node A to node B segment, and not the reverse).

Most transportation agencies use single-directional road networks for their travel demand modeling. The reason is that multiple attributes can be assigned to each direction separately, a feature that simplifies the building of a realistic network. Thus, speeds for different time periods can be assigned as separate fields on each segment (or, what is usually, done there are separate networks for the different travel periods that are modeled). Travel volumes can be assigned to each segment which, in turn, allows the creation of a *vehicle miles traveled* (VMT) field (length x volume). VMT, in turn, can be combined with travel speed to produce an estimate of travel time (e.g., VMT divided by speed - in miles per hour, times 60 to produce minutes traveled). Further, one-way streets are automatically handled since each direction is a separate segment (i.e., there just won't be a reciprocal pair in the opposite direction).

In short, a single-directional network allows more flexibility in the creation of a network and the ability to distinguish travel in different directions as well as travel time by direction and time of day. It is not surprising, therefore, that most travel demand models use a single-directional representation. Note, one can add these attributes to a single-directional network, but this requires many additional fields.

A further strength of a single-directional network is that it is usually quite up-to-date and connectivity has been ensured. Most transportation agencies spend a lot of time cleaning and updating the network. While there are always errors in a network representation, the accuracy of most modeling networks is very good.

There is a downside to single-directional networks, however. Typically, most single-directional networks model only the larger roadways, those that contribute to regional travel. Thus, all freeways, principal arterial roads, minor arterial roads, and some collector roads are included. However, most neighborhood streets are not included. The reason this is done is because the travel demand model is aimed at estimating regional and sub-regional travel patterns. Very localized travel is not of importance (and, in fact, is typically intra-zonal in nature). The result is a very efficient network because it is a lot smaller. But, there may be some error by using a 'skeleton' network. In particular, local travel might be distorted with such a simplified network. For example, if a neighborhood is bounded by four arterial roads, but with no internal streets, according to the model a crime event that originates from within the neighborhood (i.e.,

the offender lives inside the neighborhood) could take any of the four arterial roads to leave the network. In reality, the offender will probably take a particular route rather than necessarily the arterial that is closest to the offender's address. This can be handled, but it requires additional coding.[4]

As an example, for Baltimore County and the City of Baltimore, Figure 26.7 shows the 49,015 segments in the TIGER representation of these two jurisdictions while Figure 26.8 shows the 11,045 segments that are used by the Baltimore Metropolitan Council in their travel demand modeling.[5] Further, since most of the streets in the modeling representation are two-way streets, in effect, there are only about 5,000-6,000 actual streets. In other words, the TIGER network is 4.4 times larger than the modeling network. This makes calculation a lot slower than with a simplified network.[6] As we shall see in Chapter 30, the accuracy of a network is essential for a more realistic modeling of actual travel routes by offenders.

Bus Network

A bus network, on the other hand, is a specialized road network that follows the actual routes used by buses. The general road network is useful for modeling driving, walking and bicycle trips. But, it cannot be used for bus trips. The reason is simply that buses don't use every street but only the larger arterial roads. Further, travel along many bus routes is variable. That is, a full route might be used during the peak rush hours, but a shortened route might be used during the off-peak hours. Similarly, the frequency of buses (what is called *headway* by transit agencies) varies by time of day; again, in the rush hours, buses are more frequent (though slower) than during the off-peak hours.

A bus network, therefore, is essential for modeling bus trips during both the mode split stage (when trips between zones are split into separate travel modes) and during the actual network assignment.

---

4 For example, transportation modelers often put in *centroid connectors*. These are pseudo-segments that connect a zone centroid with an arterial. It is possible to add pseudo-roads to the modeling network to force travel to follow a particular route. But, it does take a lot of editing to do this.

5 The modeling network was obtained from the Baltimore Metropolitan Council and, with their permission, is illustrated here.

6 As an example of the efficiency of a modeling network compared to a TIGER network, the network assignment routine took six times longer to run with the TIGER network for Baltimore City and Baltimore County than with the modeling network. See Chapter 30 on network assignment for more information about the rules for network travel.

Figure 26.7:

TIGER Street Network
49,015 Road Segments

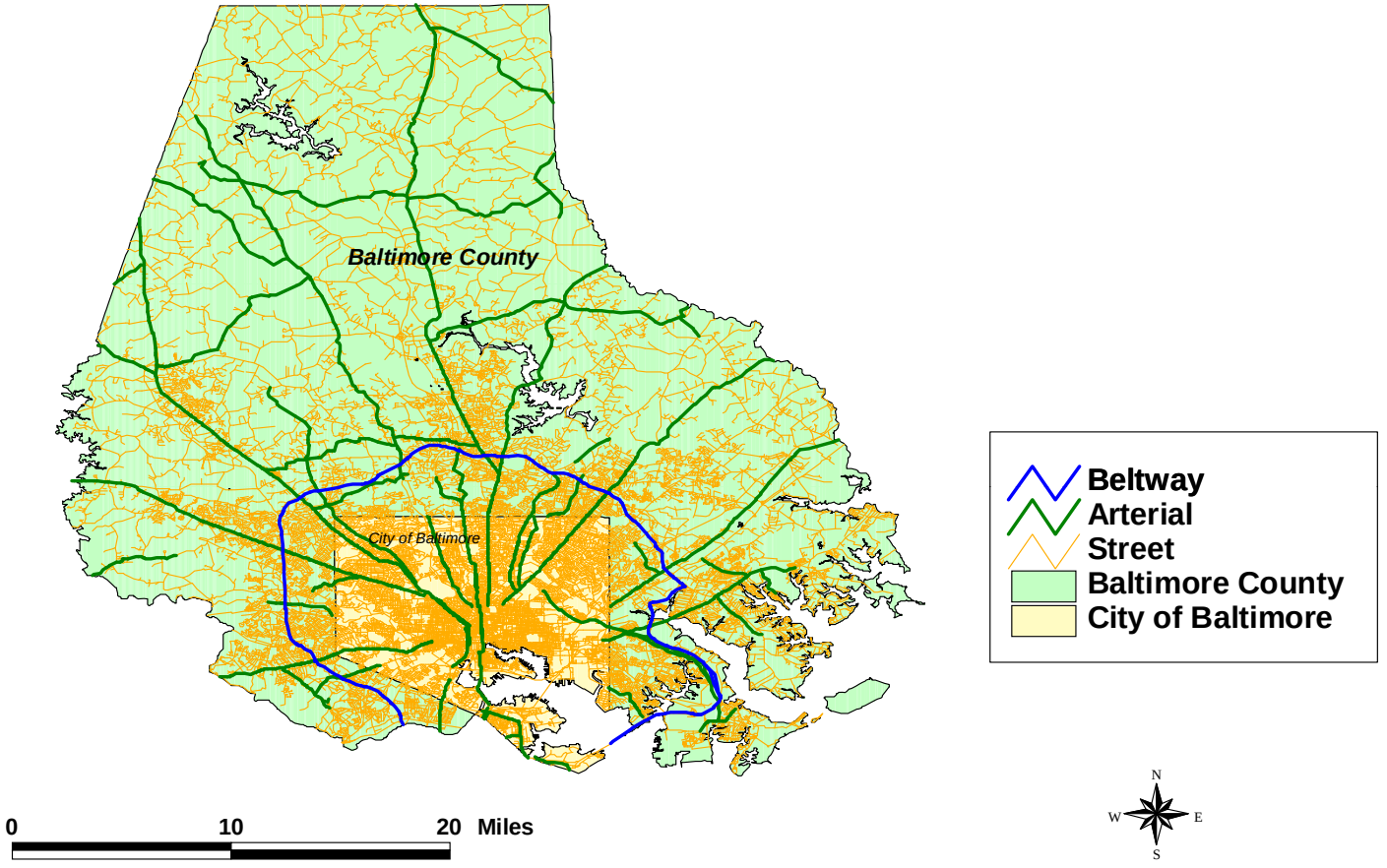
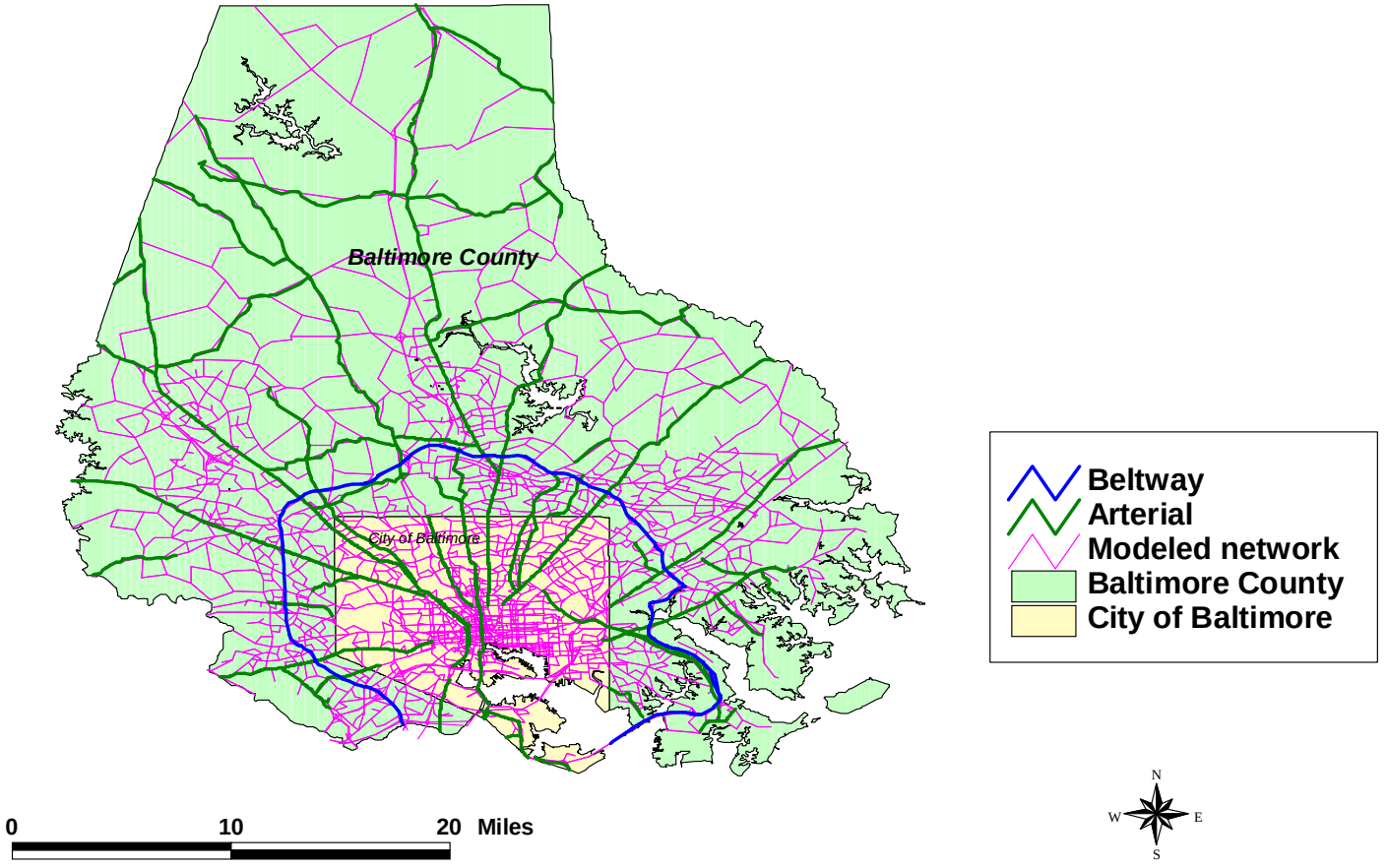


Figure 26.8:

Modeled Street Network
11,045 Road Segments



There are two components of a bus network that are required in the network, one of which is essential and the other is more optional. The first is a representation of the segments used in a bus network. Essentially, this is a network that shows where the buses travel. Bus travel can only occur along this network. As with road travel, the bus network can be represented either as a bi-directional or as a single-directional network though, again, most transportation modelers and transit agencies represent bus routes as single directions.

The second component is the location where access to the buses is allowed (i.e., the bus stops). Without explicitly indicating where there are loading and unloading points, a network routine would simply find the shortest distance from the origin to the bus route and 'add' the trip at that location. In practice, for most transit agencies, the degree of error in allowing direct access anywhere on the route is small since most bus routes stop very frequently (every couple of blocks). Thus, it may not be that important to actually code the bus stops since the amount of modeling error will be insignificant. However, for express buses and for those routes where there is a sizeable distance between bus stops, it is important to code the actual bus stops. In Chapter 30, there is a more extensive discussion of coding bus routes. Figure 26.9 illustrates the bus network for Baltimore County and Baltimore City.

Train network

In those metropolitan areas that have intra-urban train travel, it is important to also obtain a rail network. An offender cannot travel on a train except by using the existing rail system. Further, unlike the bus network, it is impossible to 'enter' the train except at explicit station locations. Thus, it is critical to obtain both the network and the station locations. Figure 26.10 illustrates the intra-urban rail system in Baltimore County and Baltimore City.

Where to Obtain Network Data?

There are many more choices in obtaining network data than with socioeconomic or land use data. Road networks can be obtained from the U.S. Census Bureau (for the TIGER system) or from vendors who improve on the TIGER system. For a modeling network, however, about the only choice is the Metropolitan Planning Organization (MPO). Since MPO's model regional travel on a continuous basis, most agencies in a metropolitan area will defer to them for that activity. Transit networks can also be obtained from MPOs though the transit agencies will have their own networks that are usually more comprehensive than those of the MPO. As with all data, the MPO might charge for the data set, though policies vary widely.

Figure 26.9:

Baltimore Bus Network
Bus Routes

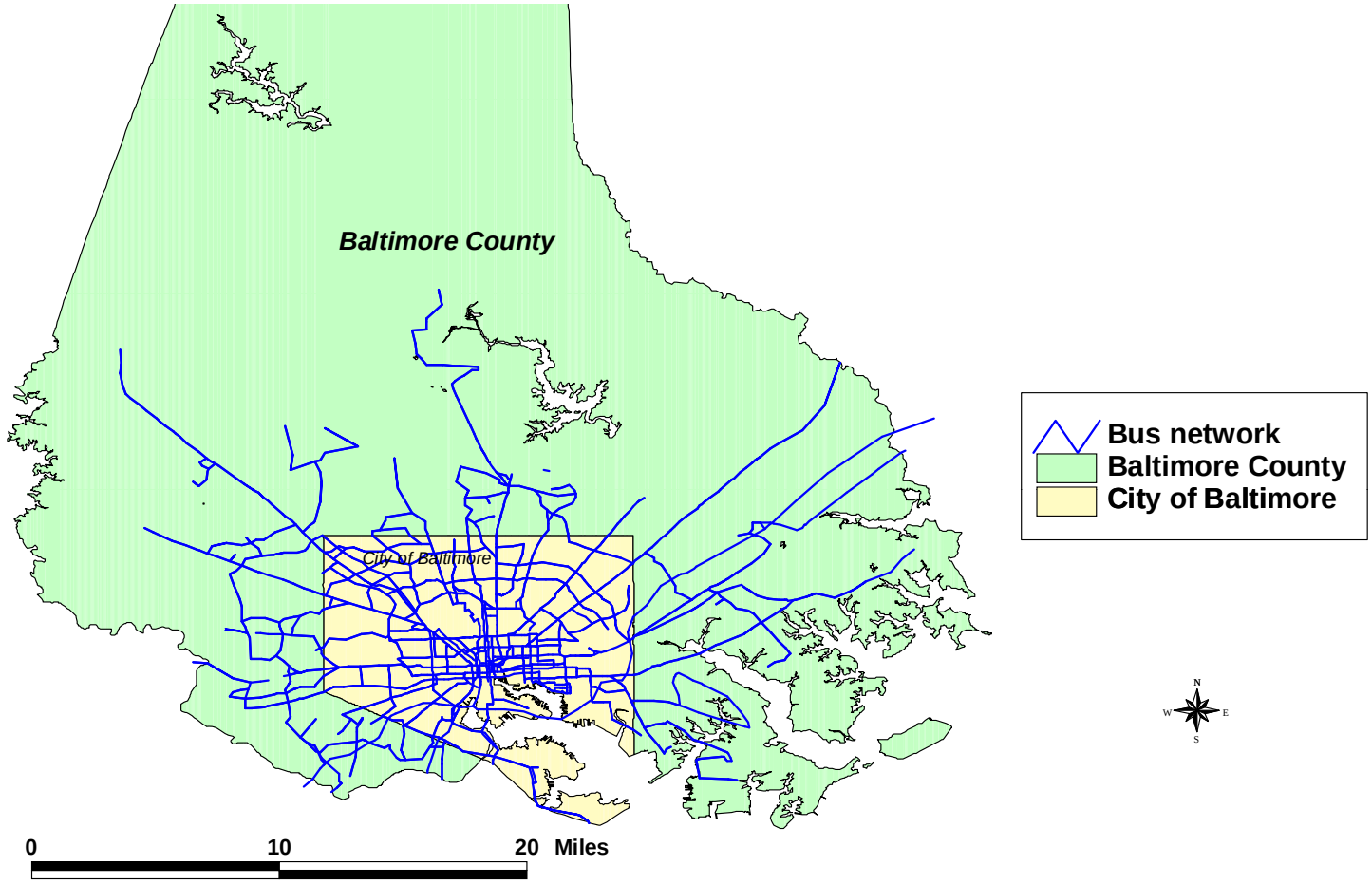
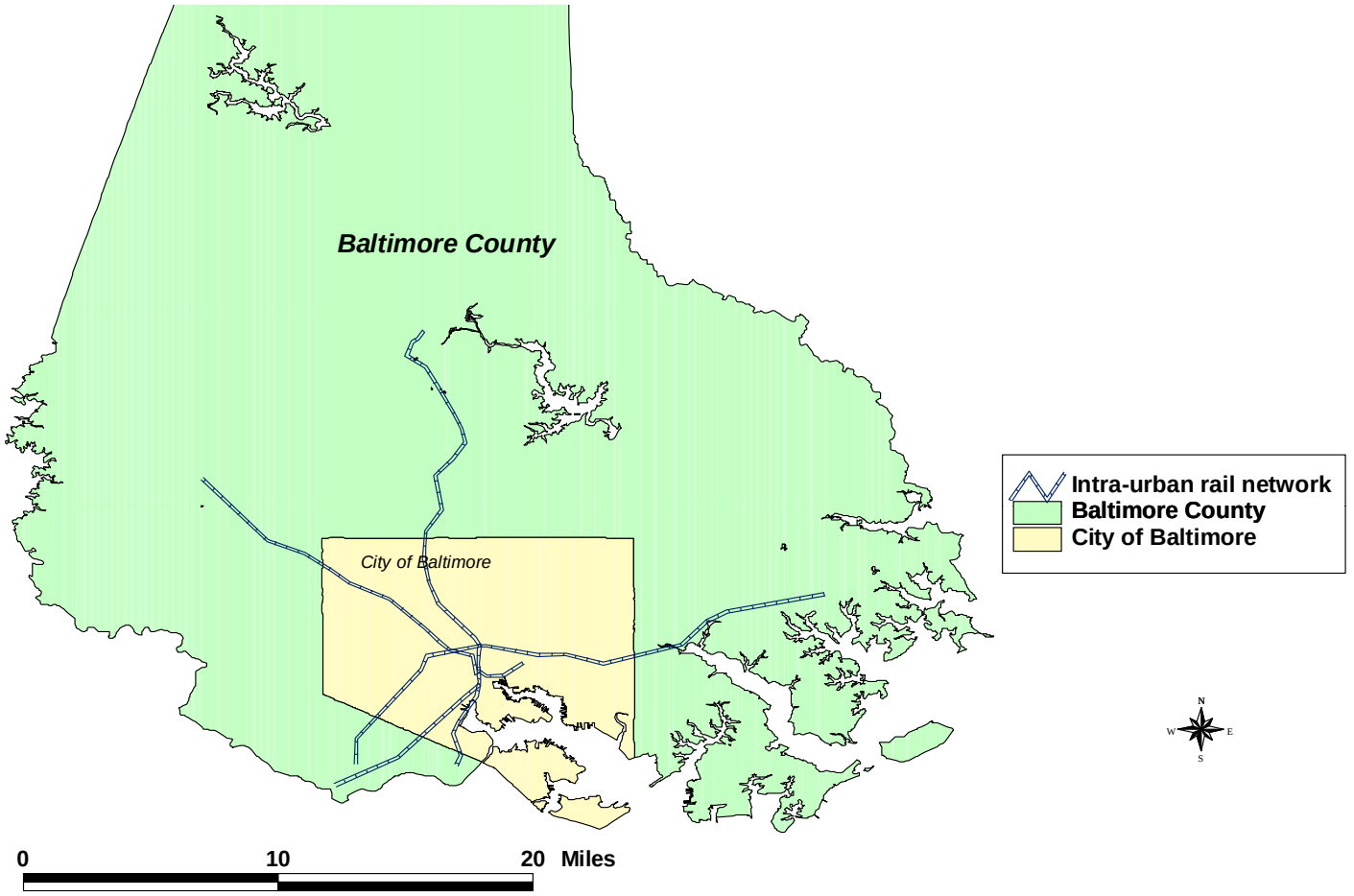


Figure 26.10:
Baltimore Intra-Urban Rail Network



Conclusion

In summary, a quite extensive collection of data is needed to run the crime travel demand model. Crime data, socioeconomic data, land use data, policy intervention scenarios, and network data must be obtained and prepared prior to running the models. Further, in practice, a lot of editing and 'cleaning' of data will be required during the modeling phase in order to improve the predictions.

Nevertheless, once the data are obtained, the model can be developed quite quickly. In the next chapter, we will examine the first stage of the crime travel demand model - trip generation.

References

AMPO (2012). *AMPO: Highlights & What's New*. Association of Metropolitan Planning Organizations: Washington, DC. http://www.ampo.org/. Accessed May 7, 2012.

Anselin, Luc (1995). Local indicators of spatial association - LISA. *Geographical Analysis*. 27, No. 2 (April), 93-115.

Bursik, R. J., Jr. & Grasmick, H. G. (1993). Economic deprivation and neighborhood crime rates, 1960-1980. *Law and Society Review*, 27, 263-268.

Citro, C. F. & Michael, R. T. (eds) (1995). *Measuring Poverty : A New Approach*. Panel on Poverty and Family Assistance, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council: Washington, DC.http://www.census.gov/hhes/www/img/povmeas/ack.pdf. Accessed May 7, 2012.

Hipp, J. R. (2007). Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point. *American Sociological Review* 72:659-680.

Kitamura, R., Yoshii, T., & Yamamoto, T. (2009). The Expanding Sphere of Travel Behaviour Research: Selected Papers from the 11th International Conference on Travel Behaviour Research. Emerald Group Publishing, Ltd: Bingley, U.K. http://books.google.com/books?id=fFqEnNOWKw8C&pg=PA375&lpg=PA375&dq=microsimulation+of+travel+behavior&source=bl&ots=ArxmN7EIZl&sig=rIUukRBjCApH22qDQ0UXp5dUOGs&hl=en&sa=X&ei=jRmkT_3aFIOi8ATImsS5CQ&ved=0CGQQ6AEwCA#v=onepage&q=microsimulation%20of%20travel%20behavior&f=false. Accessed May 4, 2012.

Langbein, L. I. & Lichtman, A. J. (1978). *Ecological Inference*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-010. Beverly Hills and London: Sage Publications.

Levine, N. (2011). Spatial variation in motor vehicle crashes by gender in the Houston Metropolitan Area. *Proceedings of the 4th International Conference on Women's Issues in Transportation. Volume II: Technical Papers*, Transportation Research Board: Washington, DC. 12-25. http://onlinepubs.trb.org/onlinepubs/conf/cp46v2.pdf. Accessed May 7, 2012.

Levine, N. (2007), "Crime travel demand and bank robberies: Using CrimeStat III to model bank robbery trips". *Social Science Computer Review*, 25(2), 239-258.

Levine, N. & Canter, P. (2011). Linking origins with destinations for DWI motor vehicle crashes: An application of crime travel demand modeling. *Crime Mapping*, 3, 7-41.

References (continued)

Miller, E. J. & Salvini, P. A. (1999). Activity-based travel behavior modeling in a microsimulation framework. Paper presented at IATBR Conference, Austin, TX. December. http://www.civ.utoronto.ca/sect/traeng/ilute/downloads/conference_papers/miller-salvini_iatbr-97.pdf. Accessed May 4, 2012.

NARC (2012). *Welcome to NARC*. National Association of Regional Councils: Washington, DC. http://www.narc.org/. Accessed May 7, 2012.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem.* Norwich: Geo Books. ISBN 0-86094-134-5.

Ortuzar, J. D. & Willumsen, L. G. (2001). *Modeling Transport* (3rd edition). J. Wiley & Sons: New York.

U.S. Census Bureau (2012). *Commuting (Journey to Work)*. U.S. Census Bureau: Washington, DC. http://www.census.gov/hhes/commuting. Accessed May 7, 2012.

U.S. Census Bureau (2011a). *Summary File 3 (SF3)*. U.S. Census Bureau: Washington, DC. http://www.census.gov/census2000/sumfile3.html. Accessed May 7, 2012.

U.S. Census Bureau (2011b). *Tiger Products*. U.S. Census Bureau: Washington, DC. http://www.census.gov/geo/www/tiger/. Accessed May 8, 2012.

Wikipedia (2012). Modifiable Area Unit Problem. Wikipedia. http://en.wikipedia.org/wiki/Modifiable_areal_unit_problem. Accessed May 7, 2012.

Wooldredge, J. (2002). Examining the (Ir)Relevance of Aggregation Bias for Multilevel Studies of Neighborhoods and Crime with an Example Comparing Census Tracts to Official Neighborhoods in Cincinnati. *Criminology* 40:681-710.