

Chapter 11:
Head-Bang Interpolation

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Interpolation II Tab	11.1
Head-Bang Technique	11.1
Median Smoother	11.3
Data Elements	11.3
Zone Structure	11.3
Local Neighborhood	11.3
Triplets	11.4
Screens	11.5
Decision Rules for Head-Bang Statistic	11.5
Example to Illustrate Decision Rules	11.6
Rates and Counts	11.9
Need to Define Intensity Variable	11.9
Smoothed median for count variable	11.10
Smoothed median for rate variable	11.10
Setup	11.10
Output	11.12
Example 1: Using the Head-Bang for Mapping Houston Burglaries	11.13
Example 2: Using the Head-Bang for Mapping Houston Burglary Rates	11.13
Example 3: Using the Head-Bang Routine to Create Burglary Rates from Separate	
Counts of Burglaries and Households	11.18
Uses of the Head-Bang Routine	11.21
Limitations of the Head-Bang Routine	11.21
Interpolated Head-Bang Statistic	11.23
Method of Interpolation	11.23
Choice of Bandwidth	11.24
Adaptive bandwidth	11.24
Fixed bandwidth	11.24
Output (areal) Units	11.24
Calculate Densities or Probabilities	11.24
Absolute densities	11.24
Relative densities	11.25
Probabilities	11.25
Output	11.25
Example: Using the Interpolated Head-Bang to Visualize Houston Burglaries	11.25
Advantages and Disadvantages of the Interpolated Head-Bang	11.28
References	11.29

Chapter 11:

Head-Bang Interpolation

Interpolation II Tab

The interpolation II tab includes the Head-Bang routine and the Interpolated Head-Bang routine. Figure 11.1 show the graphical interface for the Interpolation II page, which includes the Head-Bang and the Interpolated Head-Bang routines.

Head-Bang Technique

The Head-Bang statistic is a weighted two-dimensional smoothing algorithm that is applied to zonal (polygon) data, such as census tracts, traffic analysis zones, or zip codes. It was developed at the National Cancer Institute to smooth out ‘peaks’ and ‘valleys’ in health data that occur because of small numbers of events (Pickle, Mungiole, Jones, Gretchen, & White, 1996; Mungiole, Pickle, & Simonson, 2002; Pickle & Su, 2002).

For example, with lung cancer rates (lung cancer cases relative to population), counties with small populations could show extremely high lung cancer rates with only an increase of a couple of cases in a year or, conversely, very low rates if there was a decrease in a couple of cases. On the other hand, counties with large populations will show stable estimates because their numbers are larger; their confidence intervals will be small because changes from one year to the next will have little effect on their rates.

However, unlike other smoothing techniques, such as kernel density interpolation (discussed in Chapter 10), the Head-Bang is designed to remove small-scale local variations within a data set while preserving regional trends. It is particularly useful where there are large differences in the population sizes of the different zones, which can lead to huge variability in the rates over the study area.

The aim of the Head-Bang statistic, therefore, is to smooth out the values for smaller geographical zones while generally keeping the values for larger geographical zones. It is a variance reduction technique. The methodology is based on the idea of a median-based Head-Banging smoother proposed by Tukey and Tukey (1981) and later implemented by Hansen (1991) in two dimensions. Mean smoothing functions tend to over-smooth in the presence of edges while median smoothing functions tend to preserve the edges.

Figure 11.1:

Interpolation II Screen

CrimeStat IV

Spatial Modeling II | Crime Travel Demand | Options

Data Setup | Spatial Description | Hot Spot Analysis | Spatial Modeling I

Interpolation I | Interpolation II | Space-time analysis | Journey-to-Crime | Bayesian Journey-to-Crime Estimation

Head-Bang

Rate Count Create Rate

ID:

Numerator of Rate Denominator of Rate

Baseline unit of rate:

Use weight variable:

Number of neighbors:

Save Head-Bang

Interpolated Head-Bang

Save Interpolated Head-Bang

Compute Quit Help

Median Smoother

The Head-Bang routine applies the concept of a median smoothing function to a three-dimensional plane (an area where there is information about an attribute that occurs at each location). It is a moving average that is placed over each location (zone) but which includes information on the values of the neighboring locations. But, unlike kernel density estimation which interpolates the data to a grid, the Head-Bang preserves the zonal organization.

The Head-Bang algorithm used in *CrimeStat* is a simplification of the methodology proposed by Mungiole, Pickle and Simonson (2002) but similar to that used by Pickle and Su (2002).¹ Consider a set of zones with a variable being displayed. In a raw map, the value of a variable assigned to any one zone is independent of the values for nearby zones. However, in a Head-Bang smoothing, the value of any one zone becomes a function of the values of nearby zones. It is a moving average that provides an estimate for the value of the zone, similar to kernel density interpolation (discussed in Chapter 10) but it preserves the geographical arrangements of the zones. It is useful for eliminating extreme values in a distribution and adjusting the values of zones to be similar to their neighbors.

Data Elements

There are three elements to the data that are relevant to the Head-Bang. First, there is a zone structure. All the data are represented by zones. Second, for each zone, there is a variable of interest that will be smoothed, Z_i . This variable could be a count (e.g., the number of crimes), a rate (e.g., the number of crimes relative to the population), or even a continuous variable (e.g., median household income). Third, each zone has a weight variable, W_i (e.g., population). The values of the variable of interest, Z_i , are then estimated using the weights, W_i .

Zone Structure

The user has to choose an appropriate zone organization. A trade-off has to be made between the geographical size of the zone and the value of the weighting variable (typically population). Ideally, the zones should be as small as possible in terms of area (e.g., census tracts or even blocks) so that local variations in rates can be seen. On the other hand, a small geographical zone can have a small population, which creates volatility in the rate from one year to the next. While the technique can smooth the rates of zones with small populations, huge variability in the rates over time might be seen. Thus, choosing larger population zones would

¹ The Head-Bang statistic is sometimes written as Head Bang or even Headbang. We prefer to use the term with the hyphen (Head-Bang).

produce more stable rates. The technique has been used on zones as large as counties for national comparisons (Pickle & Su, 2002). With large zones, local variations cannot be seen, though for national comparisons that is less critical.

Local Neighborhood

The procedure works through a local neighborhood around each zone. A neighborhood can be defined in different ways. The *CrimeStat* routine uses the nearest neighbor routine (discussed in Chapter 6) to identify the K nearest neighbors where K is defined by the user. Thus, distance is the criterion for identifying a zone as being a nearest neighbor, consistent with the approach of Mungiole, Pickle and Simonson (2002). An alternative definition is that of contiguity (or adjacency) so that the neighbors share a common border. In this version of *CrimeStat*, this definition is not implemented.

For each zone in turn (the central zone), a set of neighbors is defined. Mungiole and Pickle (1999) found that 6 neighbors generally produced small errors between the actual values and the smoothed values, and that increasing the number did not reduce the error substantially. On the other hand, increasing the number of neighbors smoothed the data too much. They also found that choosing fewer than 6 neighbors could sometimes produce unstable results.

Triplets

In the original formulation of the Head-Bang technique (Hansen, 1991; Mungiole, Pickle & Simonson, 2002), the Head-Bang was applied to *triplets* around the central point (zone), the zone which is to be smoothed. Since the aim of the technique is to smooth zones that have similar underlying rates while highlighting regional differences in rates, the idea was to create a separation or cleavage in the neighborhood around the central zone. Hence, Hansen (1991) proposed the concept of a *triplet*.

A triplet is two other points (endpoints) that forms a straight line with the central point.² The line should separate neighboring zones with higher rates from those with lower rates. However, since the three points will not usually be in a perfect straight line, the angle formed between the central point and the two endpoints must be greater than a certain threshold. They typically used a minimum angle of 135° separation.

2 They call this *collinear*. However, the term collinear has different meanings including variables that are high correlated with each other. To avoid confusion, this chapter does not use that term.

Screens

The two endpoints are then assigned to two groups (called *screens*). Of the two endpoints, the one with the higher value is assigned to a *high screen* while the one with the lower value is assigned to a *low screen*. After all neighborhood points have been assigned to each screen, the median of each screen is taken and the value of the central point compared to these.

Pickle and Su (2002), however, simplified the procedure by simply dividing the values of the neighbors into two screens irrespective of whether they formed a triplet with the central point or not. The results are virtually identical to the initial Head-Bang results. Consequently, that procedure is adopted in the *CrimeStat* version. The values of the neighbors are sorted from high to low irrespective of whether they form triplets or not and divided into the high screen and 'low screen groupings at the middle record. If the number of neighbors is even, then the two groups are of equal size and mutually exclusive; on the other hand, if the number of neighbors is odd, then the middle record is counted twice, once with the high screen and once with the low screen.

For each screen, the median value is calculated. The median of the high screen is called the *high median* and the median of the low screen is called the *low median*.

Decision Rules for Head-Bang Statistic

The value of the central zone is then compared to these two medians. The decision rules are as follows:

1. First, if the value of the central zone falls between the two medians (low and high), then the central zone retains its value.
2. Second, if the value of the central zone is *either* higher than the high median *or* lower than the low median, then its weight determines whether it is adjusted. It is compared to the screen to which it is closest (high or low). If it has a weight that is greater than all the weights of its closest screen, then it maintains its value.

For example, if the central zone has a value greater than the high median but also greater has a weight greater than any of the high screen zones, then it still maintains its value.

3. However, if its weight is smaller than the weights of *any* zone in its closest screen, then the central zone takes the value of the median for the screen to which it is closest.

For example, if the central zone is closer to the high median than to the low median but has a weight that is smaller than one or more zones in the high screen, then it takes the high median as its value. Similarly, if the central zone is closer to the low median than to the high median but has a weight that is smaller than one or more zones in the low screen, then it takes the low median as its value.

4. After all points (zones) have been assigned estimates for the variable of interest, Z_i , the process is repeated 9 more times to ensure that the final smoothing is stable.

The logic ensures that if a central zone is large in size relative to its neighbors (i.e., has a greater weight), then its observed rate is most likely an accurate indicator of risk. However, if the zone is smaller in size than its neighbors, then its value is adjusted to be like its neighbors. In this case, extreme rates, either high or low, are reduced to moderate levels (smoothed). ‘Peaks’ or ‘valleys’ are minimized while the values of real edges in the data are maintained.

Example to Illustrate Decision Rules

A simple example will illustrate this process. Suppose the intensity variable is a rate (as opposed to a count; see below). For each point, the eight nearest neighbors are examined. Taking one zone (“A”), suppose the eight nearest neighbors of zone A have the following values (Table 11.1).³ Note that the value at the central point (zone A) is not included in this list. These eight are the nearest neighbors only.

Table 11.1:
Example: Nearest Neighbors of Zone “A”

<u>Neighbor</u>	<u>Rate</u>	<u>Weight</u>
B	10	1000
C	15	3000
D	12	4000
E	7	1500
F	14	2300
G	16	1200
H	10	2000
I	12	2500

³ Mungiole and Pickle (1999) found that 6 neighbors generally produced small errors between the observed and smoothed values. However, sometimes adding more neighbors can improve the stability. The example above uses 8 neighbors.

Next, the 8 neighbors are sorted from the lowest rate to the highest (Table 11.2). The record number (neighbor) and weight value are sorted along with the rate.

**Table 11.2:
Sorted Nearest Neighbors of Zone "A"**

<u>Neighbor</u>	<u>Rate</u>	<u>Weight</u>
E	7	1500
B	10	1000
H	10	2000
D	12	4000
I	12	2500
F	14	2300
C	15	3000
G	16	1200

Third, a cumulative sum of the weights is calculated starting with the lowest rate (Table 11.3). Fourth, the neighbors are then divided into two groups at the median. Since the number of records is even, then the low screen records are E, B, H, and D while the high screen records are I, F, C and G. The weighted medians of the low screen and high screen are calculated. Since these are rates, the low screen median is calculated from the first four records while the high screen median is calculated from the second four records.

**Table 11.3:
Cumulative Weights for Nearest Neighbors of Zone "A"**

<u>Neighbor</u>	<u>Rate</u>	<u>Weight</u>	<u>Cumulative Weight</u>
E	7	1500	1500
B	10	1000	2500
H	10	2000	4500
D	12	4000	8500
I	12	2500	11000
F	14	2300	13300
C	15	3000	16300
G	16	1200	17500

Using record E as an example, the calculations are as follows (assume the baseline is ‘per 10,000’). The rate is multiplied by the weight and divided by the baseline (for example, $7 * 1500/10000 = 1.05$). This is called the *score*; it is an estimate of the count of events in that zone. Table 11.4 shows the scores for each record and the cumulative score. The cumulative score of each screen is divided in half to obtain the median.

For the low screen, the median score is $8.85/2 = 4.425$. This falls between records H and D. To estimate the rate associated with this median score, the interval in scores between records H and D is interpolated, and then converted to rates. The interval between records H and D is 4.80 ($8.85-4.05$). The low screen median score, 4.425, is $(4.425-4.05)/4.80 = 0.0781$ of the distance for that interval. For the rates, the interval between records H and D is 2 ($12-10$). Thus, 0.0781 of that interval is 0.1563. This is added to the rate of record H to yield a low median rate of 10.1563.

For the high screen, the median score is $12.64/2 = 6.32$. This falls between records F and C. To estimate the rate associated with this median score, the interval in scores between records F and C is interpolated, and then converted to rates. The interval between records F and C is 4.50 ($10.72-6.22$). The low screen median score, 6.32, is $(6.32-6.22)/4.50 = 0.0222$ of the distance in that interval. The interval between the rates of records F and C is 1 ($15-14$). Thus, 0.0222 of that interval is 0.0222. This is added to the rate of record F to yield a high median rate of 14.0222.

Table 11.4:
Cumulative Scores by Screens for Nearest Neighbors of Zone “A”

<i>Low screen</i>				
<u>Neighbor</u>	<u>Rate</u>	<u>Weight</u>	<u>“Score”</u> Rate*Weight/Baseline	<u>Cumulative</u> Score
E	7	1500	1.05	1.05
B	10	1000	1.00	2.05
H	10	2000	2.00	4.05
D	12	4000	4.80	8.85
<i>High screen</i>				
<u>Neighbor</u>	<u>Rate</u>	<u>Weight</u>	<u>“Score”</u> Rate * Weight/Baseline	<u>Cumulative</u> Score
I	12	2500	3.00	3.00
F	14	2300	3.22	6.22
C	15	3000	4.50	10.72
G	16	1200	1.92	12.64

Finally, the rate associated with the central zone (zone A in our example) is compared to these two medians. If its rate falls between these medians, then it keeps its value. For example, if the rate of zone A is 13, then that falls between the two medians (10.1563 and 14.0222).

On the other hand, if its rate falls outside this range (either lower than the low median or higher than the high median), its value is determined by its weight relative to the screen to which it is closest. For example, suppose zone A has a rate of 15 with a weight of 1700. In this case, its rate is higher than the high median (14.0222) but its weight is smaller than three of the weights in the high screen. Therefore, it takes the high median as its new smoothed value. Relative to its neighbors, it is smaller than three of them so that its value is probably too high.

But, suppose it has a rate of 15 and a weight of 3000? Even though its rate is higher than the high median, its weight is also higher than the four neighbors making up the high screen. Consequently, it keeps its value. Relative to its neighbors, it is a large zone and its value is probably accurate.

For counts (discussed below), the comparison is simpler because all weights are equal. Consequently, the count of the central zone is compared directly to the two medians. If it falls between the medians, it keeps its value. If it falls outside the two medians, then it takes the one to which it is closest (the high median if it has a higher value or the low median if it is lower).

Rates and Counts

The original Head-Bang statistic was applied to rates (e.g., number of lung cancer cases relative to population). In the *CrimeStat* implementation, the routine can be applied to counts (volumes) or rates or can even be used to estimate a rate from counts. Counts have no weights (i.e., they are self-weighted). In the case of rates, though, they should be weighted (e.g., by population). The most plausible weighting variable for a rate is the same baseline variable used in the denominator of the rate (e.g., population, number of households) because the rate variance is proportional to 1/baseline (Pickle and Su, 2002).

Need to Define Intensity variable

Whether a count or a rate variable is to be analyzed, an *Intensity* variable must be defined on the Primary file page (see Chapter 3). The Intensity variable is the variable that will be smoothed. If it is not defined, then the Head-Bang routine will not be available. Note that a separate weight variable on the Primary file page should also be used to weight the data if a rate is being analyzed. But, this can only work in conjunction with a defined intensity variable. In other words, be sure that an intensity variable is defined to use the Head-Bang routine.

Smoothed median for count variable

With a count variable, there is only a count (the number of events). There is no weighting of the count since it is self-weighting (i.e., the number equals its weight). In *CrimeStat*, the count variable is defined as the Intensity variable on the Primary file page. For a count variable:

1. If the value of the central zone falls between the two medians (low median and high median), then the central zone retains its value.
2. If the value of the central zone is higher than the high median, then it takes the high median as its smoothed value;
3. If the value of the central zone is lower than the low median, then it takes the low median as its smoothed value.

Smoothed median for rate variable

With a rate, there is both a value (the rate) and a weight. The rate variable is defined as the Intensity variable on the Primary file page. However, there is a separate weight that must be applied to this rate to distinguish a large zone from a small zone. In *CrimeStat*, the weight variable is always defined on the Primary file page as the Weight variable.

Depending on whether the rate is input as part of the original data set or created out of two variables from the original data set, it will be defined slightly differently. If the rate is part of the original data set, then it is defined as the intensity variable on the Primary file page. However, if the rate is created out of two variables from the Primary file data set, it is defined on the Head-Bang interface under 'Create rate'.

Setup

For either a rate or a count, the statistic requires an *intensity* variable be defined in the Primary file. The user must specify whether the variable to be smoothed is a rate variable, a count variable, or two variables that are to be combined into a rate. If a weight is to be used (for either a rate or the creating of a rate from two count variables), then it must be defined as an Intensity on the Primary file page. Note that if the intensity variable is a rate, it should also be weighted. A typical weighting variable is the population size of the zone.

The user has to complete the following steps to run the routine:

1. **Define input file** and coordinates on the Primary file page
2. **Define an intensity variable**, $Z(\text{intensity})$, on the Primary file page.
3. **OPTIONAL: Define a weighting variable** in the weight field on the Primary file page (for rates and for the creating of rates from two count variables)
4. **Define an ID variable** to identify each zone.
5. **Select data type:**
 - A. **Rate:** the variable to be smoothed is a rate variable which calculates the number of events (the numerator) relative to a baseline variable (the denominator).
 - a. The baseline units should be defined, which is an assumed multiplier in powers of 10. The default is 'per 100' (percentages) but other choices are 0 (no multiplier used), 'per 10' (rate is multiplied by 10), 'per 1000', 'per 10,000', 'per 100,000', and 'per 1,000,000'. This is not used in the calculation but for reference.
 - b. If a weight is to be used, the 'Use weight variable' box should be checked.
 - B. **Count:** the variable to be smoothed is a raw count of the number of events. There is no baseline used.
 - C. **Create Rate:** A rate is to be calculated by dividing a count variable by a baseline variable.
 - a. The user must define the count (numerator) and baseline (denominator) variables.
 - b. The baseline scale *must* be defined, which is an assumed multiplier in powers of 10. The default is 'per 100' (percentages) but other choices are 1 (no multiplier used),

'per 10' (rate is multiplied by 10), 'per 1000', 'per 10,000', 'per 100,000', and 'per 1,000,000'. This is used in the calculation of the rate.

- c. If a weight is to be used, the 'Use weight variable' box should be checked.
6. **Select number of neighbors.** In CrimeStat, the number of neighbors can run from 4 through 40. The default is 6. If the number of neighbors selected is even, the routine divides the data set into two equal-sized groups. If the number of neighbors selected is odd, then the middle zone is used in calculating both the low median and the high median. It is recommended that an even number of neighbors be used (e.g., 4, 6, 8, 10).
7. **Select output file.** The output can be saved as a dbase 'dbf' file. If the output file is a rate, then the prefix RateHB is used. If the output is a count, then the prefix VolHB is used. If the output is a created rate, then the prefix CrateHB is used.
8. **Run the routine** by clicking 'Compute'.

Output

The Head-Bang routine creates a 'dbf' file with the following variables:

1. The ID field
2. The X coordinate
3. The Y coordinate
4. The smoothed intensity variable (called 'Z_MEDIAN'). Note that this is not a Z score but a smoothed intensity (Z) value
5. The weight applied to the smoothed intensity variable. This will be automatically 1 if no weighting is applied.

The 'dbf' file can then be linked to the input 'dbf' file by using the ID field as a matching variable. This would be done if the user wants to map the smoothed intensity variable.

Example 1: Using the Head-Bang Routine for Mapping Houston Burglaries

Figure 11.2 shows a map of Houston burglaries by traffic analysis zones. The mapped variable is the number of burglaries committed in 2006. On the Head-Bang interface, the 'Count' box was checked, indicating that the number of burglaries will be estimated. The number of neighbors was left at the default 6. The output 'dbf' file was then linked to the input 'dbf' file using the ID field to allow the smoothed intensity values to be mapped. The variable is mapped with 5 equal-size intervals. Quintiles could also have been used.

Figure 11.3 show a smoothed map of the number of burglaries conducted by the Head-Bang routine. To be consistent with Figure 11.2, 5 equal-size intervals were also used. Comparing the two maps, it can be seen that there are fewer zones in the Head-Bang map that are in the lowest interval/bin (in yellow). The actual count was 498 zones with scores of less than 10 in Figure 11.3 compared to 528 zones in Figure 11.2. Also, there are fewer zones in the highest interval/bin (in black) as well. The actual count was 181 zones with scores of 40 or more in Figure 11.3 compared to 215 zones in Figure 11.2. In other words, the Head-Bang routine has assigned many of the highest or lowest values to the median values of their neighbors.

Example 2: Using the Head-Bang Routine for Mapping Burglary Rates

The second example shows how the Head-Bang routine can smooth rates. In the Houston burglary data base, a rate variable was created that divided the number of burglaries in 2006 by the number of households in 2006. This variable was then multiplied by 1000 to minimize the effects of decimal place (the baseline unit). Figure 11.4 show the raw burglary rate (burglaries per 1,000 households) for the City of Houston in 2006.

The Head-Bang routine was set up to estimate a rate for this variable (Burglaries per 1000 Households). On the Primary file page, the intensity variable was defined as the calculated rate (burglaries per 1,000 households) because the Head-Bang will smooth the rate. Also, a weight variable was selected on the Primary file page. In this example, the weight variable was the number of households. With any rate, there is always the potential of a small zone producing a very high rate. Consequently, the estimates were weighted to ensure that the values of each zone are proportional to their size. Zones with larger numbers of households will keep their values whereas zones with small numbers of households will most likely change their values to be closer to their neighbors.

On the Head-Bang interface, the 'Rate' box was checked (Figure 11.5). The ID variable was selected (which is also TAZ03). The baseline number of units was set to 'per 1000'; this is for information purposes, only, and will not affect the calculation.

Figure 11.2:

Burglaries in Houston: 2006

Number of Burglaries by Traffic Analysis Zones

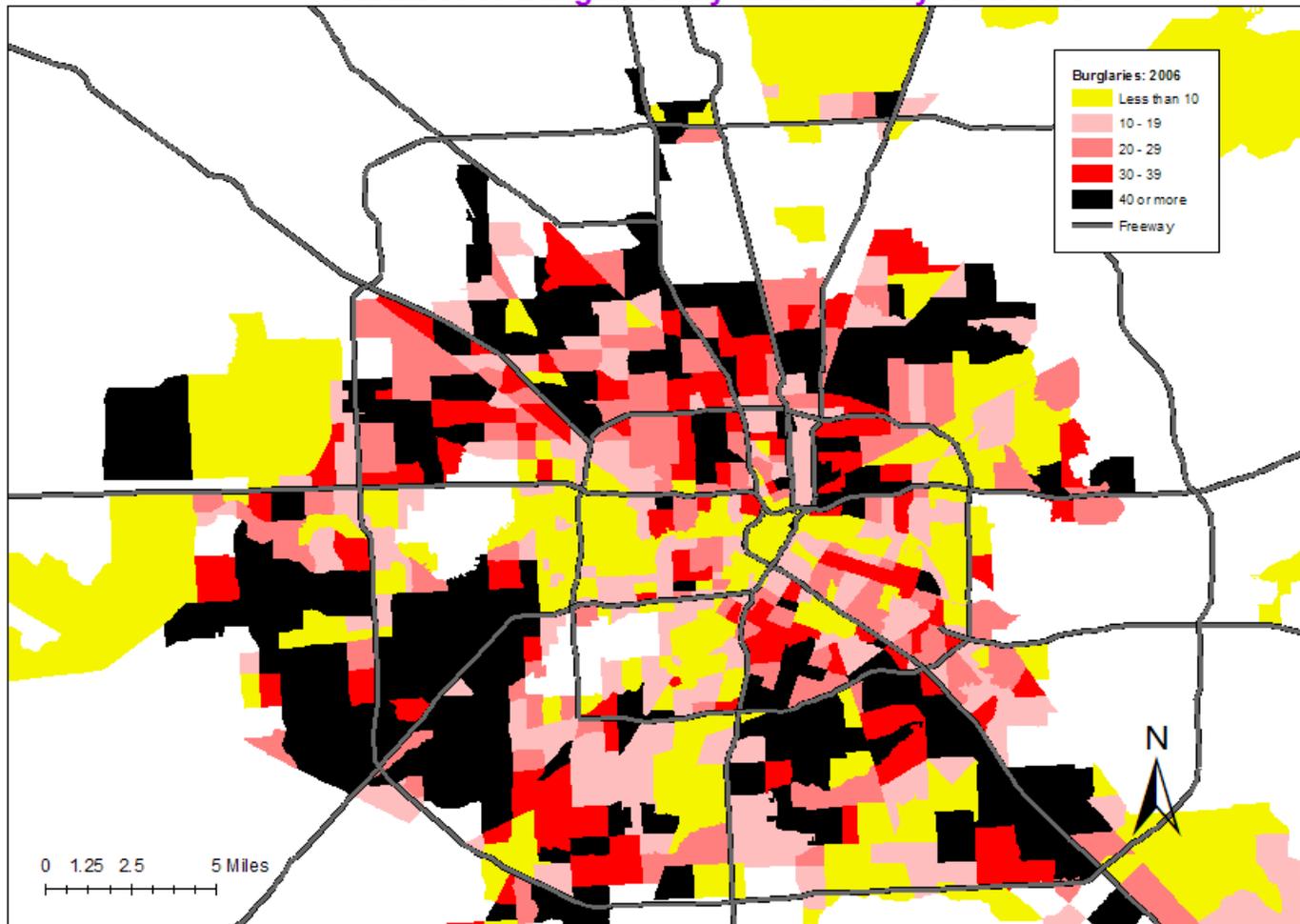


Figure 11.3:

Head-Bang Estimate of Burglaries in Houston: 2006

Smoothed Number of Burglaries by Traffic Analysis Zones

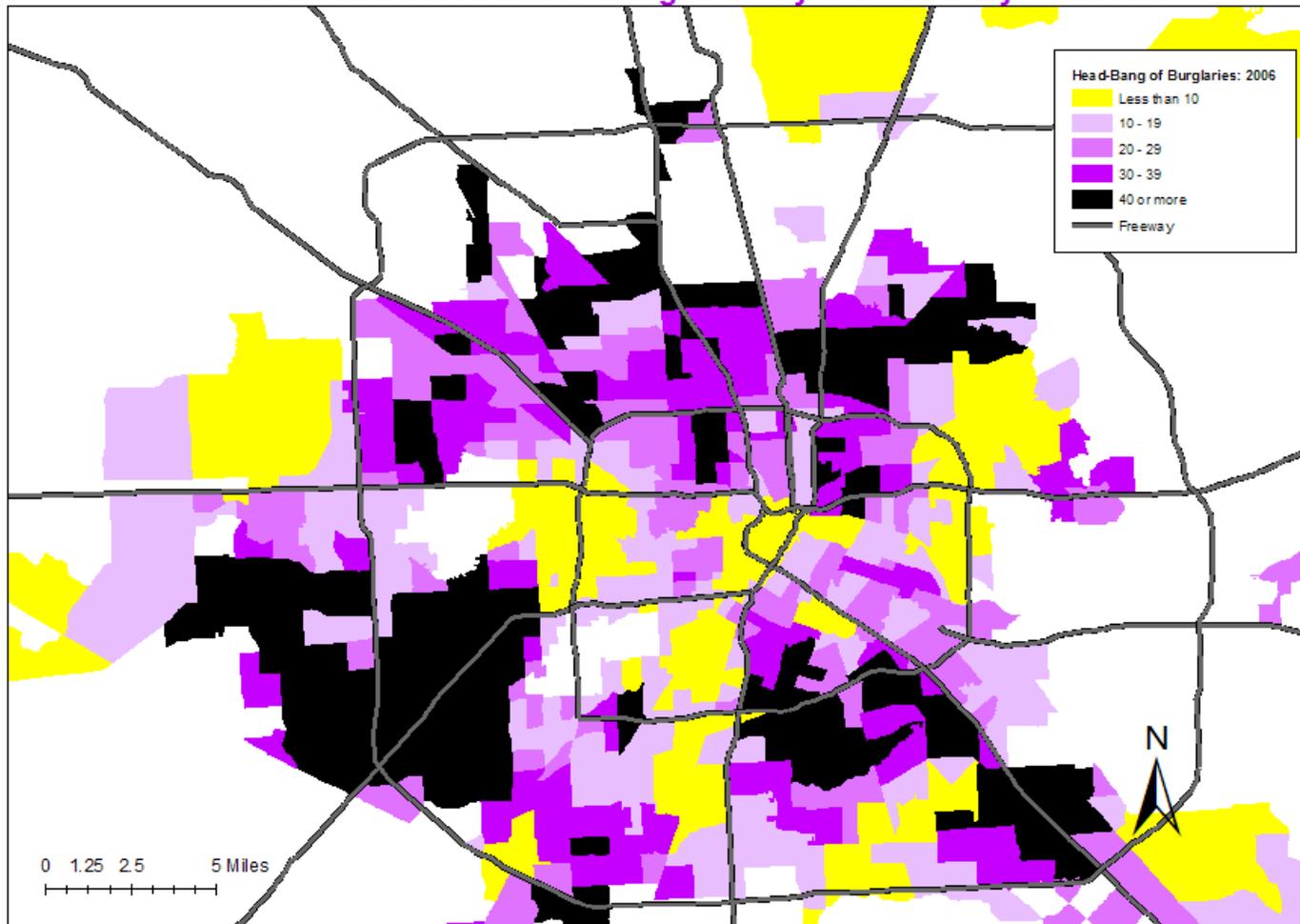


Figure 11.4:
Burglary Rate in Houston: 2006
Number of Burglaries per 1000 Households

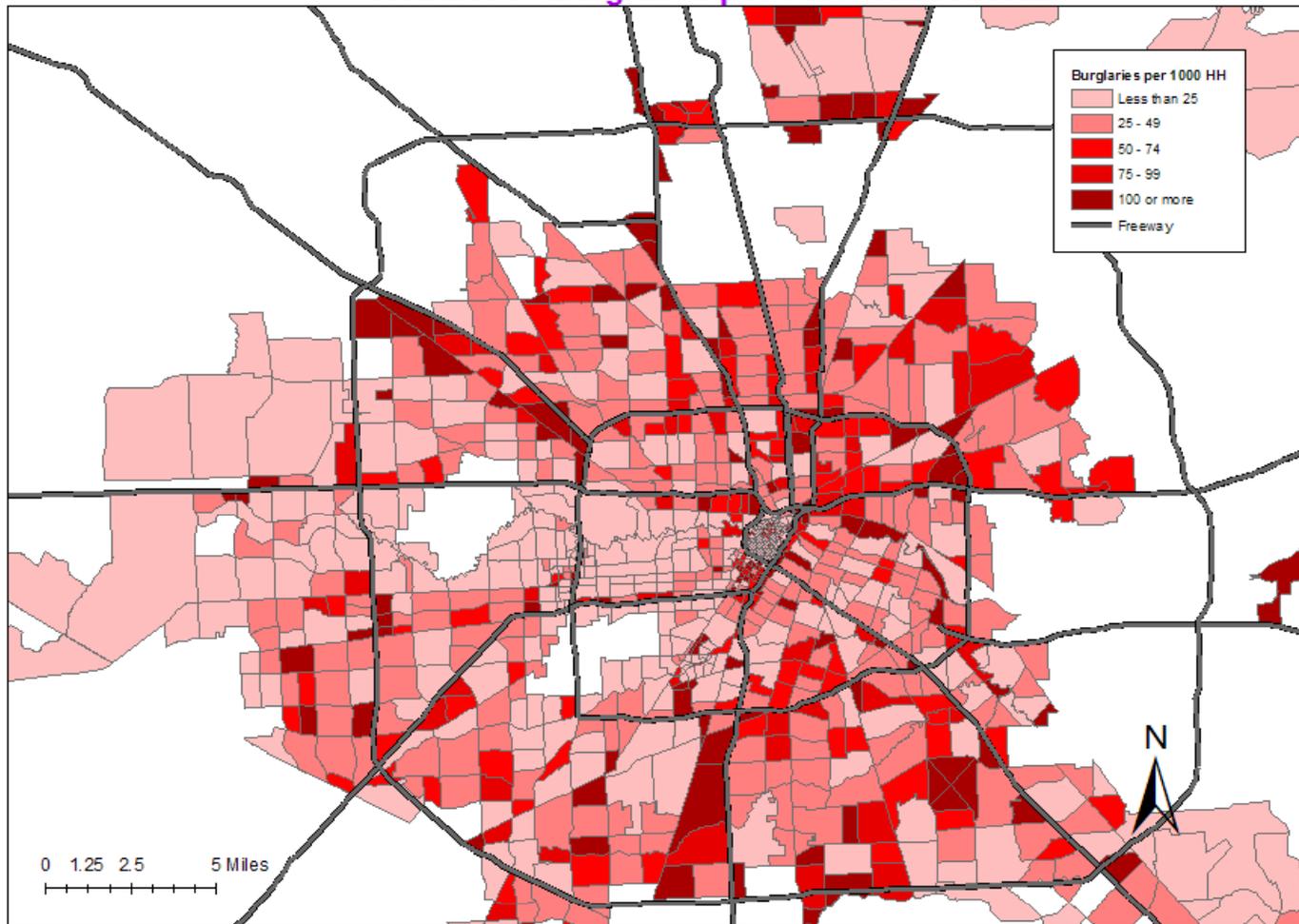


Figure 11.5:

Defining Rates with Head-Bang Routine

The screenshot shows the 'CrimeStat IV' software window with the 'Spatial Modeling II' tab selected. The 'Options' sub-tab is active, and the 'Spatial Modeling I' section is expanded to show the 'Head-Bang' routine configuration. The 'Head-Bang' checkbox is checked. The 'Rate' radio button is selected, and the 'ID' is set to 'TAZ'. The 'Baseline unit of rate' is set to 'per 1000'. The 'Use weight variable' checkbox is checked. The 'Number of neighbors' is set to 6. The 'Save Head-Bang' button is visible. The 'Interpolated Head-Bang' checkbox is unchecked, and the 'Save Interpolated Head-Bang' button is visible. The 'Compute', 'Quit', and 'Help' buttons are at the bottom of the window.

Option	Value
Head-Bang	Checked
Rate	Selected
Count	Not Selected
Create Rate	Not Selected
ID	TAZ
Baseline unit of rate	per 1000
Use weight variable	Checked
Number of neighbors	6
Numerator of Rate	per 100
Denominator of Rate	8

On the Primary file page, the number of households was chosen as the weight variable and the 'Use weight variable' box was checked under the Head-Bang routine.. The number of neighbors was left at the default 6. Finally, an output 'dbf' file was defined in the 'Save Head-Bang' dialogue.

The output 'dbf' file was linked to the input 'dbf' file using the ID field to allow the smoothed rates to be mapped. Figure 11.6 show the result of smoothing the burglary rate. As can be seen, the rates are more moderate than with the raw numbers (comparing Figure 11.4 with Figure 11.6). There are fewer zones in the highest rate category (100 or more burglaries per 1,000 households) for the Head-Bang estimate compared to the raw data (64 compared to 185) but there are also more zones in the lowest rate category (0-24 burglaries per 1,000 households) for the Head-Bang compared to the raw data (585 compared to 520). In short, the Head-Bang routine reduced the rates throughout the map.

Example 3: Using the Head-Bang Routine to Create Burglary Rates from Separate Counts of Burglaries and Households

The third example illustrates using the Head-Bang routine to create smoothed rates. In the Houston burglary data set, there are two variables that can be used to create a rate. First, there is the number of burglaries per traffic analysis zone. Second, there is the number of households that live in each zone. By dividing the number of burglaries by the number of households, an exposure index can be calculated. Of course, this index is not perfect because some of the burglaries occur on commercial properties, rather than residential units. But, without separating residential from non-residential burglaries, this index can be considered a rough exposure measure.

On the Head-Bang interface, the 'Create Rate' box is checked (Figure 11.7). The ID variable is selected (which is TAZ03 in the example). The numerator variable is selected which, in the example is the number of burglaries (BURG2006). Next, the denominator variable is selected. In the example, the denominator variable is the number of households (HH2006). The baseline units must be chosen and, unlike the rate routine, are used in the calculations. For the example, the rate is 'per 1,000' which means that the routine will calculate the rate (burglaries divided by households) but then will multiply by 1,000. On the Head-Bang page, the 'Use weight variable' box under the 'Create rate' column is checked.

Next, the number of neighbors are chosen, both for the numerator and for the denominator. One has to be careful about the denominator especially if some zones have very few households. The result would be an extreme rate estimate. To avoid this, it is recommended that a larger number of neighbors be used for the denominator than for the numerator. In the

Figure 11.6:

Head-Bang of Burglary Rate in Houston: 2006

Smoothed Estimate of Number of Burglaries per 1000 Households

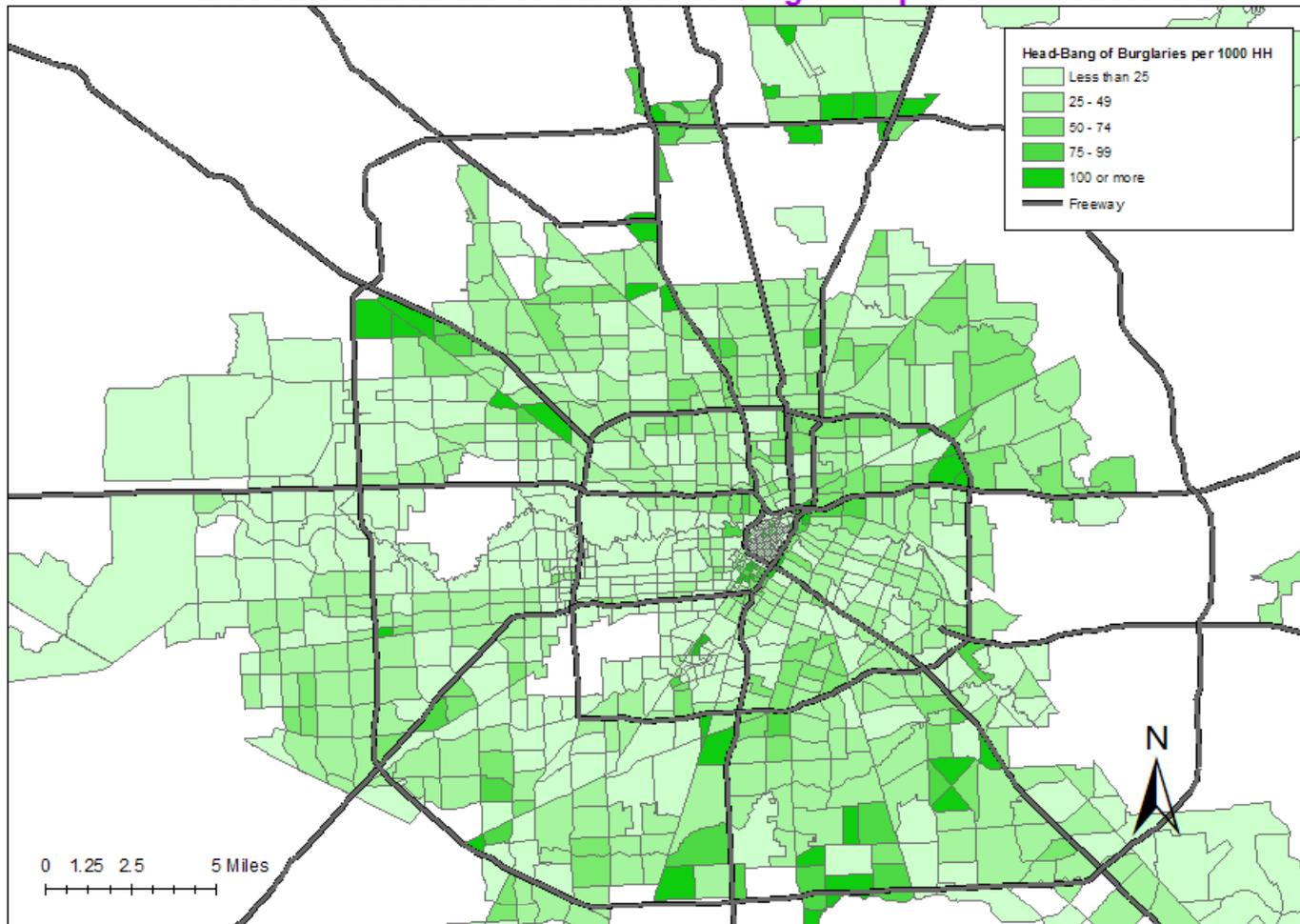


Figure 11.7:

Creating Rates with Head-Bang Routine

The screenshot shows the 'CrimeStat IV' software window with the 'Spatial Modeling II' tab selected. The 'Options' sub-tab is active, and the 'Spatial Modeling I' section is expanded to show the 'Head-Bang' routine. The 'Head-Bang' checkbox is checked. The 'Create Rate' radio button is selected. The 'ID' is set to 'TAZ03'. The 'Numerator of Rate' is 'BURG2006' and the 'Denominator of Rate' is 'HH2006'. Both have a 'per 1000' baseline unit. The 'Use weight variable' checkbox is checked. The 'Number of neighbors' is set to 6. There are buttons for 'Save Head-Bang', 'Interpolated Head-Bang', and 'Save Interpolated Head-Bang'. At the bottom are 'Compute', 'Quit', and 'Help' buttons.

Option	Value
Head-Bang	Checked
Rate	Selected
Count	Not Selected
Create Rate	Selected
ID	TAZ03
Numerator of Rate	BURG2006
Denominator of Rate	HH2006
Baseline unit of rate	per 1000
Use weight variable	Checked
Number of neighbors	6

example, the default of 6 neighbors is chosen for the numerator variable (burglaries) while 8 neighbors are chosen for the denominator variable (households).

Finally, a 'dbf' output file was defined and the routine was run. The output 'dbf' file was then linked to the input 'dbf' file using the ID field to allow the smoothed rates to be mapped. Figure 11.8 show the results. Compared to the raw burglary rate (Figure 11.2), there are fewer zones in the highest category (36 compared to 185) but also more zones in the lowest category (607 compared to 520). Like the rate smoother, the rate that is created has reduced the rates throughout the map.

Uses of the Head-Bang Routine

The Head-Bang routine is useful for several purposes. First, it eliminates extreme measures, particularly very high ones ('peaks'). For a rate, in particular, it will produce more stable estimates. For zones with a small population, a few events can cause dramatic increases in the rate. The Head-Bang smoother will eliminate those extreme fluctuations. The use of population weights for estimating rates ensures that unusually high or low proportions that are reliable due to large populations are not modified whereas values based on small populations are modified to be more like those of the surrounding counties. Similarly, for counts (counts), the method will produce values that are more moderate.

Limitations of the Head-Bang Routine

On the other hand, the Head-Bang methodology does distort data. Because the extreme values are eliminated, the routine aims for more moderate estimates. However, those extremes may be real. Consequently, the Head-Bang routine should not be used to interpret the results for any one zone but more for the general pattern within the area.

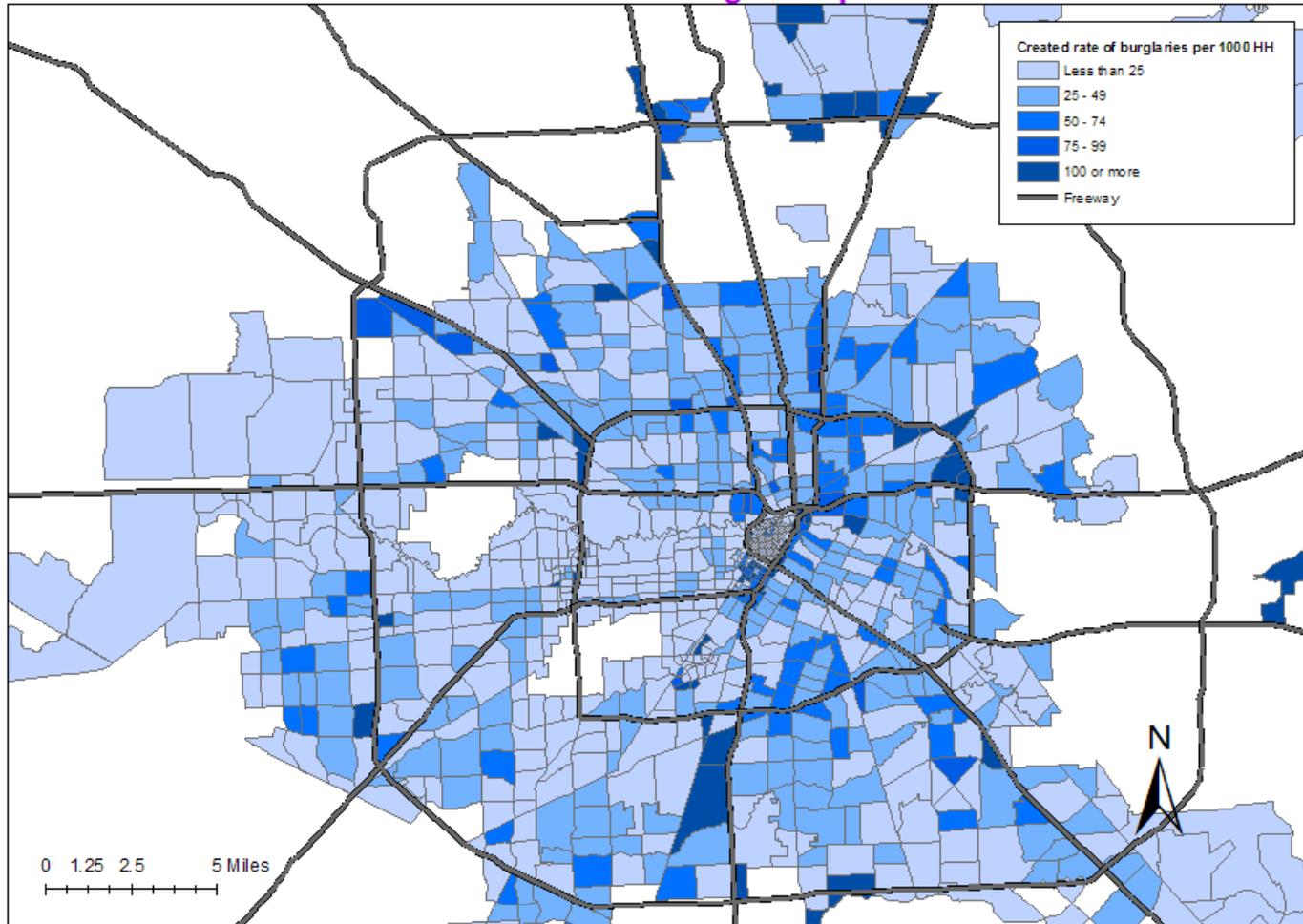
Also, there is often a trade-off that the user will have to make between the geographical size of the zone (smaller is better) with the stability of the estimates (larger population is better). Choosing zones that are very small (e.g., blocks or block groups) can produce unreliable estimates while choosing zones that are too large (e.g., districts or even counties) can eliminate meaningful local variations. The choice of zones is critical.

However, if used carefully, Head-Bang smoothing is a powerful tool for examining risk within a study area and for identifying changes in risk over time.

Figure 11.8:

Head-Bang of Burglary Rate in Houston: 2006

Created Rate of Number of Burglaries per 1000 Households



Interpolated Head-Bang Statistic

The Head-Bang calculations can be interpolated to a grid. If the user checks this box, then the routine will also interpolate the calculations to a grid using kernel density estimation. An output file from the Head-Bang routine is required. Also, a reference file is required to be defined on the Reference File page.

Essentially, the routine takes a Head-Bang output and interpolates it to a grid using a kernel density function. The same results can be obtained by inputting the Head-Bang output on the Primary file page and using the single kernel density routine on the Interpolations I page. The user must then define the parameters of the interpolation. However, there is no intensity variable in the Interpolated Head-Bang because the intensity has already been incorporated in the Head-Bang output. Also, there is no weighting of the Interpolated Head-Bang estimate.

The user must then define the parameters of the interpolation. Chapter 10 discussed these in more detail and provided guidelines, which will not be repeated here.

Method of Interpolation

There are five types of kernel distributions to interpolate the Head-Bang to the grid:

1. The **normal** kernel overlays a three-dimensional normal distribution over each point that then extends over the area defined by the reference file. This is the default kernel function. However, the normal kernel tends to over-smooth. One of the other kernel functions may produce a more differentiated map;
2. The **uniform** kernel overlays a uniform function (disk) over each point that only extends for a limited distance;
3. The **quartic** kernel overlays a quartic function (inverse sphere) over each point that only extends for a limited distance;
4. The **triangular** kernel overlays a three-dimensional triangle (cone) over each point that only extends for a limited distance; and
5. The **negative exponential** kernel overlays a three dimensional negative exponential function over each point that only extends for a limited distance.

The different kernel functions produce similar results though the normal is generally smoother for any given bandwidth.

Choice of Bandwidth

The kernels are applied to a limited search distance, called 'bandwidth'. For the normal kernel, bandwidth is the standard deviation of the normal distribution. For the uniform, quartic, triangular and negative exponential kernels, bandwidth is the radius of a circle defined by the surface. For all types, a larger bandwidth will produce smoother density estimates and both adaptive and fixed bandwidth intervals can be selected.

Adaptive bandwidth

An adaptive bandwidth distance is identified by the minimum number of other points found within a circle drawn around a single point. A circle is placed around each point, in turn, and the radius is increased until the minimum sample size is reached. Thus, each point has a different bandwidth interval. The user can modify the minimum sample size. The default is 100 points. If there is a small sample size (e.g., less than 500), then a smaller minimum sample size would be more appropriate).

Fixed bandwidth

A fixed bandwidth distance is a fixed search radius for each point. The user must define the radius and its distance units (miles, nautical miles, feet, kilometers, meters.)

Output (areal) units

Specify the areal density units as points per square mile, per squared nautical miles, per square feet, per square kilometers, or per square meters. The default is points per square mile.

Calculate Densities or Probabilities

The density estimate for each cell can be calculated in one of three ways:

Absolute densities

This is the number of points per grid cell and is scaled so that the sum of all grid cells equals the sample size. This is the default.

Relative densities

For each grid cell, this is the absolute density divided by the grid cell area and is expressed in the output units (e.g., points per square mile).

Probabilities

This is the proportion of all incidents that occur in each grid cell. The sum of all grid cells equals 1. Select whether absolute densities, relative densities, or probabilities are to be output for each cell. The default is absolute densities.

Output

The results can be output as a *Surfer for Windows* file (for both an external or generated reference file) or as an *ArcGIS* '.shp', *MapInfo* '.mif', *ArcGIS Spatial Analyst* '.asc', or ASCII grid 'grd' file (only if the reference file is generated by *CrimeStat*.) The output file is saved as IHB<root name> with the root name being provided by the user.

Example: Using the Interpolated Head-Bang to Visualize Houston Burglaries

The Houston burglary data set was, first, smoothed using the Head-Bang routine (Figure 11.3 above) and, second, interpolated to a grid using the Interpolated Head-Bang routine. The kernel chosen was the default normal distribution but with a fixed bandwidth of 1 mile. Figure 11.9 show the results of the interpolation.

To compare this to an interpolation of the original data, the number of burglaries in each zone was interpolated using the single kernel density routine. The kernel used was also the normal distribution with a fixed bandwidth of 1 mile. Figure 11.10 show the results of interpolating the raw burglary numbers.

A comparison of these two figures shows that they both capture the areas with the highest burglary density. However, the Interpolated Head-Bang produced fewer high density cells which, in turn, allowed the moderately high cells to stand out. For example, in southwest Houston, the Interpolated Head-Bang showed two small areas of moderately high density of burglaries whereas the raw interpolation merged these together.

Figure 11.9:
Interpolated Head-Bang Estimate of Burglaries in Houston: 2006
One Mile Bandwidth for Traffic Analysis Zones

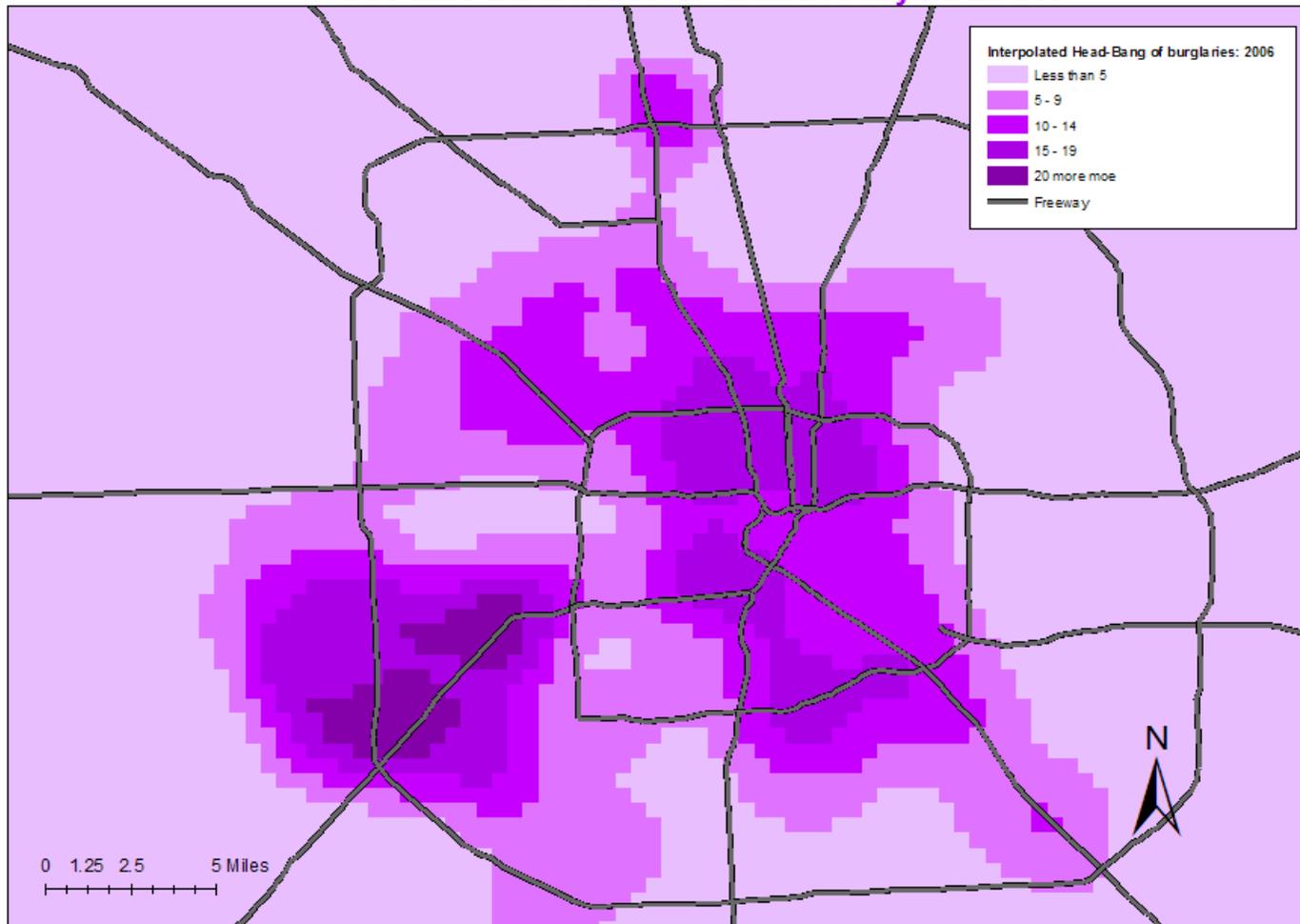
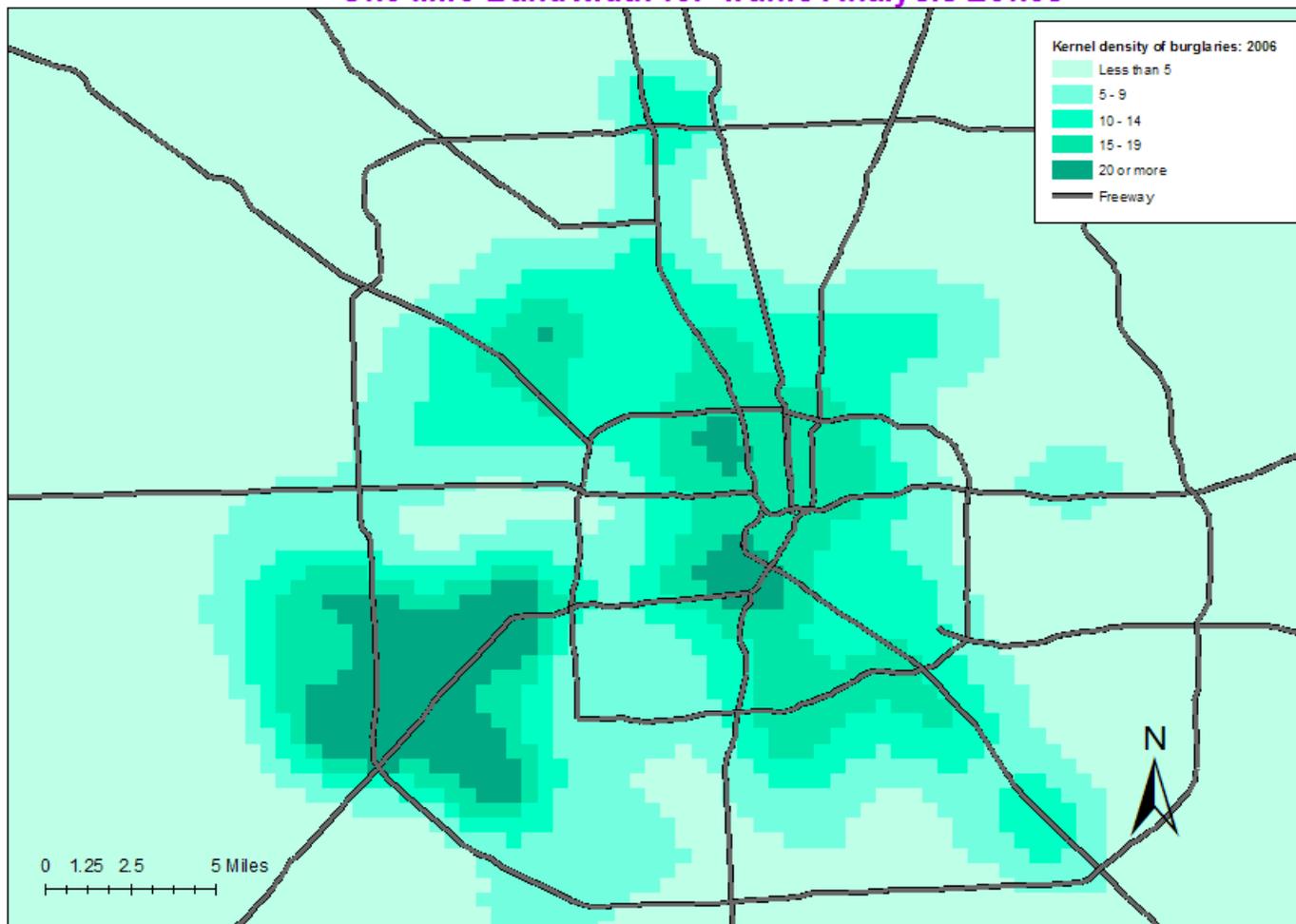


Figure 11.10:
Kernel Density Interpolation of Burglaries in Houston: 2006
One Mile Bandwidth for Traffic Analysis Zones



Advantages and Disadvantages of the Interpolated Head-Bang

The Interpolated Head-Bang routine has advantages and disadvantages over the regular Head-Bang. Its advantages are that it, like the Head-Bang, captures the strongest tendencies by eliminating 'peaks' and 'valleys'. But, it allows a smoother representation of the data. Zones are usually of unequal size with those in the center of a metropolitan area being much smaller than those in the periphery (the MAUP problem; see Wikipedia, 2012; Hipp, 2007; Wooldridge, 2002; Openshaw, 1984).). There is a visual distortion that occurs with large areas simply due to the larger area that they cover. The Head-bang will mute the effect of extreme low or high values in the periphery, but it will not eliminate the visual distortion that one sees in looking a map.

On the other hand, the interpolated Head-Bang routine does this by smoothing the data even more than the Head-Bang itself. There is a danger that it could over-smooth the data. The user has to determine whether the elimination of areas with very high or very low density values is real and not just due to small number of events.

For law enforcement applications, this may or may not be an advantage. Some hot spots, for example, are small areas where there are a many crime events. Smoothing the data may eliminate the visibility of these. On the other hand, large hot spots will generally survive the smoothing process because the number of events is large and will usually spread to adjacent grid cells. As usual, the user has to be aware of the advantages and disadvantages in order to decide whether a particular tool, such as the interpolated Head-Bang, is useful or not.

References

Fink, A. M. (1988). How to polish off Median Polish. *SIAM J. Sci. and Stat. Comput.*, 9(5), 932-940.

Hansen, K.M., Simonson, K. H. and Statistical Methodology and Applications Branch, NCI (2010). *Head-Bang PC Software (version 3.0)*. Surveillance Research, Center Control and Population Sciences, National Cancer Institute. <http://surveillance.cancer.gov/headbang/>

Hansen, K.M., Simonson, K. H. and Statistical Methodology and Applications Branch, NCI (2010). *Head-Bang PC Software (version 3.0)*. Surveillance Research, Center Control and Population Sciences, National Cancer Institute. <http://surveillance.cancer.gov/headbang/>

Hansen, K. M. (1991). Head-banging: robust smoothing in the plane, *IEEE Transactions on Geoscience and Remote Sensing*, 29, 369-378.

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.

Mungiole, M., Pickle, L. W. & Simonson, K. H. (2002). Application of a weighted Head-Banging algorithm to Mortality data maps", *Statistics in Medicine*, 18, 3201-3209.

Mungiole, M. & Pickle, L. W. (1999). "Determining the optimal degree of smoothing using the weighted head-banging algorithm on mapped mortality data", In ASC '99 - Leading Survey & Statistical Computing into the New Millennium, Proceedings of the ASC International Conference, September. Available at <http://srab.cancer.gov/headbang>.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Geo Books. [ISBN 0-86094-134-5](https://www.isbn-international.org/product/0-86094-134-5).

Pickle, L. W. & Su, Y. (2002). Within-State geographic patterns of health insurance coverage and health risk factors in the United States, *American Journal of Preventive Medicine*, 22 (2), 75-83.

Pickle, L. W., Mungiole, M., Jones, G. K., & White, A. A. (1996). *Atlas of United States Mortality*. National Center for Health Statistics: Hyattsville, MD.

Tukey, P. A. & Tukey, J. W. (1981). Graphical display of data sets in 3 or more dimensions, Barnett, V. (ed.), *Interpreting Multivariate Data*, Wiley, New York.

References (continued)

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

Wooldridge, 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.