Determining the Optimal Degree of Smoothing Using the Weighted Head-banging Algorithm on Mapped Mortality Data

Michael Mungiole and Linda W. Pickle

Abstract

We have used the weighted head-banging smoother on mapped mortality data to remove background noise while retaining edge effects. The mean squared error (MSE) between the raw and smoothed data was used as the measure of the degree of smoothing for each combination of parameters as applied to several data sets to which different levels of random noise had been added. The minimum number of nearest neighbors required for adequate smoothing was approximately six. The smoothed map results varied only slightly when the number of nearest neighbors was further increased, however, selecting too few nearest neighbors sometimes caused odd behavior.

Keywords: nonlinear smoothers, geographic analysis

Introduction

Use of statistical smoothing techniques has been a useful method when attempting to determine the spatial pattern, if any, that exists with mapped data. Previously, Kafadar (1994) has shown that head-banging (Hansen, 1991) has demonstrated good results when using the criterion mean squared error (MSE) between raw data and smoothed raw data with added noise. Her results included comparison of MSE results for several linear and nonlinear smoothers, including head-banging. While these results were encouraging for head-banging, the algorithm did not take into account data reliability. We wanted to also consider data reliability in using this algorithm, hence, it was modified to include weights. This work (Mungiole et. al., in press) has shown that weighted head-banging algorithm is an appropriate method for smoothing mortality data. The algorithm exhibits a number of characteristics that are advantageous for mortality and mapped two-dimensional data. These characteristics include the ability to retain (i.e., features are relatively unchanged when compared to presmoothed data) edges and perimeter values, retain spikes when they constitute reliable data, and remove noise for data that is comprised of small numbers (unreliable data). When the weighted head-banging algorithm was used to obtain the smooth maps for the National Center for Health Statistics' mortality atlas (Pickle et al., 1996), we used a consistent set of parameters to obtain each smoothed map. These parameters were based on considering several different data sets having various degrees of noise and spatial correlation.

It is evident to those involved in smoothing two-dimensional data that the degree of smoothing is somewhat subjective for any method employed. Researchers have used various methods to determine the appropriate parameter value(s) when smoothing two-dimensional data. Although we have previously used a consistent set of parameters in smoothing mapped mortality data, it seemed appropriate to further investigate this issue by manipulating selected parameters in the head-banging algorithm. Hence, this study was undertaken to more closely determine the optimal set of parameters to use with the weighted head-banging algorithm. The algorithm was applied to two different causes of mortality that had different inherent levels of noise and spatial correlation.

Methods

When smoothing one-dimensional data, the selection of neighboring points to consider while smoothing is reasonably straightforward. Normally one would select an equal number of points on either side of the one being smoothed. This becomes somewhat more complicated when attempting to smooth two-dimensional data because the overall area to be smoothed is not necessarily a uniform lattice structure and the relative positions of neighboring points need to be considered. To take this issue into account, the head-banging algorithm for two-dimensional data is based on using a set of triples (three "nearly" collinear points), with the center point for each triple representing the one being smoothed. The degree of collinearity is specified by selecting the minimum central angle (θ^*) that is acceptable for each triple.

For the smoothing process, the larger of the two endpoints for each triple is placed in a high endpoint grouping. Similarly, the smaller endpoint for each triple is placed in a low endpoint grouping. Medians for the high and low endpoint groupings are then determined. These two medians and the value being smoothed, along with the associated weights for each of these three values, are then compared to determine an overall median which represents the smoothed value. A more detailed explanation of the smoothing process and how the weights are considered is given in Mungiole et. al. (in press). After each smoothed value is determined, all values are updated simultaneously. This represents a single iteration of the smoothing process which continues for the number of iterations specified. Along with the number of iterations and θ^*, the number of nearest neighbors (NN) and the maximum number of triples (NTRIP) that may be used to smooth each point are also specified. These four parameters represent the total number of parameters that can be specified by the user for the smoothing process.

For two-dimensional data values along the perimeter, the actual number of triples that are considered in smoothing a perimeter value may be less than the number that are used to smooth an interior value. This is because fewer triples may exist at or near the perimeter that meet the collinearity requirement. Hence, the actual number of triples used to smooth perimeter values are likely to be less than the maximum number specified (NTRIP). In addition, if no triples are found that meet the smoothing criteria, then one or more triples are extrapolated from the nearest neighbors (Hansen, 1991). If the number of triples used is less than NTRIP, there would likely be a disproportionately heavy influence on the perimeter values being smoothed from the few nearest neighbors that are used for extrapolation.

While it is apparent that several other smoothing methods exist, our success with weighted head-banging for mortality data led us to further investigate this method to more closely define the optimal values for the parameters that determine the degree of smoothing. To consider a reasonable number of combinations of the four parameters (number of iterations, θ^*, NN, and NTRIP) that can be manipulated, the effort was reduced to selecting various values only for NN and NTRIP. Manipulating these two parameters while keeping the other two constant was based on results of Hansen (1991) and our unpublished research work. Specifically, these two parameters were selected because they have been determined to be the most important parameters as far as influencing the degree of smoothing. For this study, we set NTRIP=2/3 NN (rounded to the nearest integer) when NN was the parameter being manipulated. To more fully investigate the manipulation of smoothing parameters, NTRIP was also varied for a particular mortality cause to determine how this parameter influenced the degree of smoothing. For this part of the analysis, NN was held constant at a sufficiently large value. We selected $\theta^*=135°$ since this value gives good results across several different data sets. Finally, we have found that 10 iterations are sufficient for smoothing nearly all types of data, such that no additional changes occur to the data values by the time this iteration is reached.

The process of selecting a criterion required that there be some appreciable change in the criterion value as NN and NTRIP were manipulated. After considering several possibilities, we selected MSE as the criterion because it was successfully used by Kafadar (1994) in distinguishing the quality of different smoothers. Another reason is that MSE is a relatively logical and simple statistic to consider in most evaluations, including determining the appropriate degree of smoothing. With the selection of this criterion, it was hoped to determine if MSE provided measures of the minimum and most appropriate degrees of smoothing. The specific equation used to determine MSE is as follows:

$$\mathrm{MSE} = \sum_{I=1}^{798} (s_{i,NN} - r_i)^2 / 797$$

where r_i = actual rate for HSA i
$s_{i,0} = r_i + x\%$ noise (unsmoothed)
$s_{i,NN} = r_i + x\%$ noise (smoothed) given NN>0

The data used in this study were heart disease and HIV mortality for white males, ages 35-44. These observed data were used as input to a mixed effects hierarchical model, as explained in the mortality atlas (Pickle et. al., 1997). The output from this modeling process were predicted age-specific mortality data to which noise was added and was considered as the presmoothed data in the smoothing process. Noise was obtained by calculating various levels of Gaussian noise having a mean of zero and the standard error representing selected percentages of the mean mortality rate. The results were then smoothed and the smoothed data were compared to the predicted data which were used as the standard in the MSE calculation.

The unit of area considered in this study is the health service area (HSA). This is an aggregation of counties based on where people receive their hospitalization (Makuc et. al., 1991). Because of this criterion, some HSAs may be represented by a single city comprised of one or more counties. To determine the contribution of individual HSAs to the total MSE, an influence measure for each HSA was obtained. This was obtained by using the following equation:

$$\text{Influence value}_{[i]} = \mathrm{MSE}/\mathrm{MSE}_{[i]}$$

where i represents the particular HSA. This equation is simply a ratio of the total MSE divided by the MSE when a single HSA is removed from the calculation. These methods were employed for both the heart disease and HIV mortality data.

Three different levels of noise were selected for each of the two mortality causes. For heart disease, levels were selected for which the standard error of the noise represented 10, 20, and 30% of the mean mortality rate for this cause. For comparability of results, it was desired to obtain maximum MSEs that were approximately equal to those obtained for heart disease. Because the distribution for HIV, for example, was highly skewed to the right and the rates (in general) were lower than those for heart disease, it required Gaussian noise with standard errors equal to 47, 57, and 67% of the mean rates added to the predicted rates to provide maximum MSEs similar in value to those obtained for heart disease. For any given value of added noise, the amount added to each HSA was equal to the Gaussian random value divided by the respective HSA's weight. This resulted in the less reliable HSAs generally having a larger amount of added noise. When considering the different noise levels across mortality causes, the requirement to have similar maximum MSEs resulted in a greater amount of noise being added to the causes whose mean rates were lower. This reflected the situation in observed data where there is reduced reliability for causes with lower rates due to sparse death data.

In the MSE calculations, a value was also obtained for the case of NN=0 which is the MSE between the predicted and presmoothed (predicted plus noise) data. This is meaningless in terms of degree of smoothing but represents the maximum MSE that can occur for a particular level of noise.

Results and Discussion

Figure 1 contains the mapped data for heart disease. The predicted age-specific mortality rates (Figure 1a) are modified by having 30% noise added to these rates. These presmoothed data (Figure 1b) are then smoothed using the weighted head-banging algorithm using selected values of NN and NTRIP (Figure 1c-h). In proceeding from Figure 1c to 1h, there is an increase in the values of NN and NTRIP which result in an increase in smoothing. Figure 2 shows results for HIV mortality rates, with the subfigures analogous to those presented for heart disease in Figure 1. For this mortality cause, 65% noise was added to the predicted rates prior to smoothing. Both of these figures indicate that there are moderate changes in the mapped data when proceeding from the presmoothed (predicted plus noise) to the smoothed cases for NN=4 and 6. Above NN=6, there are minor changes in the mapped data with increasing degree of smoothing.

A plot of the MSE vs. NN for heart disease (Figure 3) indicates that there is an "elbow" in the curve (for larger levels of noise) that occurs at approximately NN=6. Above this parameter value the curve plateaus, implying that one should select a value for NN > 6 when smoothing these data. For HIV(Figure 4), there is an odd result for the MSE vs. NN plot at NN=6 which prevents determining where the expected elbow in the curve is located. Specifically, for six nearest neighbors, there is a steep spike that is evident in the solid curves for all three selected noise levels.

To investigate this anomaly further, an influence analysis was conducted to determine the influence of each HSA on the total MSE. Performing this calculation for the HIV smoothed data with 65% noise gives results that are shown in Figure 5. This figure indicates that there is a single HSA whose influence value is approximately two. Thus, when this HSA is removed from the MSE calculation the net MSE is approximately cut in half, i.e., this particular HSA contributes one half of the total MSE for HIV when smoothed using NN=6. The only other HSA that appears to have a substantial influence has a value of .07 which results in contributing about 6% to the total MSE.

When the HSA that contributes half of the total MSE is removed from the calculation of MSE, the results are as indicated in the dashed curve in Figure 4. The values of these curves at NN=6 are consistent with the results obtained in Figure 5 in that the MSE is approximately twice as large when all HSAs are considered in the analysis. Note that the dashed curve shows a fairly consistent increase in MSE as the degree of smoothing decreases.

Upon closer inspection of the presmoothed and smoothed data, it was determined that when six nearest neighbors are selected, the mortality rate for HSA number 83 (comprised of Nassau and Suffolk counties, NY) assumes the very high mortality rate of HSA number 94 (New York City). The large weight associated with the rate for HSA number 94 is an important factor contributing to this undesirable effect. When values of NN≠6 are used, the mortality rate for HSA number 83 is smoothed to values much closer to its predicted rate.

Results of the influence analysis for HIV suggested that we further investigate the heart disease data to determine if a similar effect occurred for one or more HSAs. The MSE vs. NN plots in Figure 3 do not indicate any expected anomalies similar to the one found for the HIV data. To check for this possibility, we selected several values of NN for 30% noise and found that the only instance where there was a disproportionate contribution of an individual HSA occurred for NN=4. These results are indicated in Figure 6. While it should be noted that the vertical scale in this figure is quite different from the influence values for HIV (Figure 5), there is a single HSA which appears to be an outlier, having an influence value slightly larger than 1.08. As was the case for HSA number 83 for HIV, this HSA (number 516) was also one that is along the perimeter of the United States. Specifically, it is along the Mexican border of Texas.

When HSA number 516 is removed from the MSE calculation, the results show a slight decrease in the total MSE when fewer than six nearest neighbors are used in the smoothing process (Figure 3, dashed curve). For this case, the presmoothed rate for HSA number 516 was abnormally high because a large amount of random noise added to its predicted rate and this high value was retained after smoothing when NN=4. When six or more nearest neighbors were used in the head-banging algorithm, the rate for HSA number 516 was reduced to a value close to its predicted rate.

For both causes of death, the anomalous HSAs along the U.S. perimeter indicated that the smoothing process for these HSAs encountered edge effects. For the case of HIV data being smoothed with NN=6, triples may have been extrapolated and the New York City rate was applied to HSA number 83. For heart disease data for NN=4, the number of triples used to smooth HSA number 516 was less than NTRIP. In both cases, the result was an abnormally high rate that had a strong impact on increasing the total MSE.

It should be indicated that the mapped HIV results (Figure 2) for the anomalous smoothed HSA at NN=6 is not apparent. This is due to several factors. The choropleth maps only have a very limited number of classes that can be shown and still be understood from a cognitive standpoint. A second reason is that for data classed by quantiles (which was done in this study), the highest and lowest classes often have large ranges to include the outliers. This particular HSA was in the highest class for both the presmoothed data (Figure 2b) and the smoothed case with NN=6 (Figure 2d). Finally, and probably most importantly, the area of the HSA is reasonably small and has little visual effect on the map reader.

While the MSE results and a qualitative analysis of the maps may seem sufficient for determining appropriate levels of smoothing, we also considered the degree of spatial correlation when NN was varied. These results are considered for the case of white male (ages 35-44) heart disease for different values of NN using the Moran statistic as a measure of spatial correlation. As previously indicated, the maps (Figure 1) and MSE plot (Figure 3) indicate that one should select a value of NN=6 or higher to obtain an appropriate amount of smoothing to the predicted rates with added noise. Figure 7 is a plot of NN vs. the Moran statistic for the case with 30% added noise. A smoothed depiction of this curve results in an elbow occurring at the location where NN is approximately 9. At increased levels of smoothing, there is little additional increase in the spatial correlation. This indicates that NN=9 is a good fit to this predicted data set.

This same analysis was performed for four other data sets (HIV, prostate cancer, pneumonia and influenza, and suicide). There is a consistency across data sets in the appropriate degree of smoothing when using spatial correlation as a criterion in that the results (data not shown) indicate that the elbow in the NN vs. Moran statistic curve occurs when NN is between 7 and 9. Hence, for all five causes of mortality considered, there is little additional increase in spatial correlation above this small range of values. It is likely that the predicted (prior to adding noise) rate for each cause has some influence on the variability of the value of NN where the elbow occurs.

All of the previous results presented were for cases in which NN was manipulated while NTRIP maintained a value equal to 2/3 NN. Figure 8 shows a series of maps that exhibit the influence of NTRIP on the amount of smoothing for the case of suicide for white females, ages 65-74. For these data, a constant value of NN=12 was used while NTRIP was varied between 1 and 12 and 117% noise was added to the predicted rates. The results indicate apparent changes in the smoothed maps for NTRIP<3 but there is little effect on map appearance for higher values of NTRIP. Hence, this parameter is quite robust in the effect it has on MSE.

While there is a subtle increase in the MSE for smoothed results when NTRIP is reduced below 3, there is an outlier HSA that would not be apparent in the plot of MSE vs. NTRIP (Figure 9) but there is some evidence of this when one considers the ranges in the mortality rates. A single outlier is the reason that the MSE for NTRIP=1 and 2 are higher than for NTRIP>2. The map legend (Figure 8) indicates a maximum mortality rate of 233 deaths per 100,000 population which occurs after noise is added to the predicted rates. Upon smoothing using NTRIP =1 and 2, this maximum rate is reduced to 105 (which occurs for a single HSA) and it is not until NTRIP is increased to 3 that the rate is reduced to 19.5, a value slightly higher than the maximum predicted rate of 17.6.

When the degree of spatial correlation and the manipulation of NTRIP are considered together for this data set, the results are consistent with the maps and the MSE plot. In Figure 10, the plot of NTRIP vs. Moran statistic indicates there is virtually no spatial correlation (Moran statistic=0.02) after noise is added to the predicted rates and it is not until NTRIP=2 that the Moran statistic reaches a value comparable to that for the predicted rates in which the Moran statistic was 0.1172. Above NTRIP=4, there is little additional spatial correlation obtained with increased levels of smoothing. But if one had used as few as two maximum number of triples, the spatial correlation would have been at a level approximately half the value at which it stabilizes for NTRIP>3.

Conclusions

The individual cases where HSAs that are along the perimeter of the United States can readily take on the value of a neighbor that may be vastly different in magnitude suggest that there should be a minimum number of nearest neighbors considered when smoothing. Based on the results of the two data sets considered in this study, it appears that at least six nearest neighbors should be selected when using the weighted head-banging algorithm. This is especially important for areas along the perimeter where there aren't even six nearest neighbors that would meet the minimal triple angle requirement and/or the extrapolated values used to obtain additional triples prior to smoothing are disproportionately influenced by a single value.

Using MSE as a measure of the appropriate degree of smoothing provides results that are consistent with the mapped data. After accounting for the potential anomaly of a perimeter value changing drastically, there is little difference in MSE when the number of nearest neighbors varies among values greater than six. Similarly, for the mapped data, there are some discernible differences between NN=4 and NN=6 but little changes occur when larger values of NN are selected for smoothing.

Including the use of the Moran statistic in determining the optimal degree of smoothing adds an additional dimension to this process. When considering how this statistic varies with the smoothing parameters, it seems apparent that it provides consistent results in determining an appropriate amount that a mortality data set needs to be smoothed. Changes in the Moran statistic as smoothing parameters are varied, when considered along with the value of this statistic for the predicted rates, provide a fairly reliable measure of the appropriate smoothing parameter values that should be used.

Finally, it is recommended that one perform an influence analysis when comparing the results between the presmoothed and smoothed data. This simple analysis can easily identify potential areas that contribute a disproportional amount to the MSE and whose values change radically after being smoothed.

About the Authors

Michael Mungiole is a Mathematical Statisticians with the National Center for Health Statistics, CDC. Linda Pickle is a Mathematical Statistician with the National Cancer Institute, NIH.
Michael can be contacted at National Center for Health Statistics, 6525 Belcrest Road, Room 915, Hyattsville, MD 20782, USA; tel. 301-436-7904 x145; fax 301-436-7955; e-mail mim4@cdc.gov.

References

Hansen, KM. Head-banging: Robust smoothing in the plane. *IEEE Trans on Geoscience and Remote Sensing*. 29:369-378. 1991.

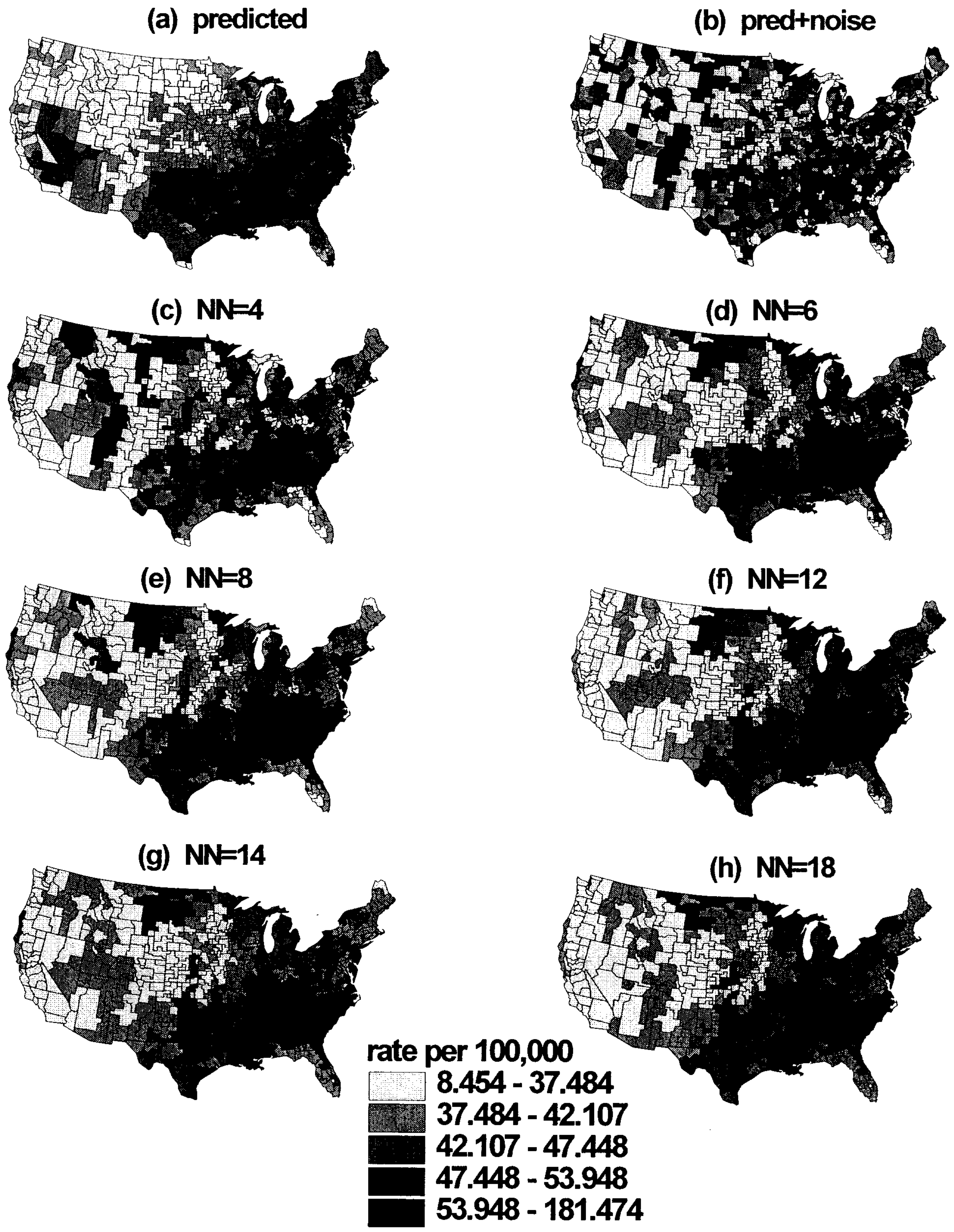
Kafadar, K. Choosing among two-dimensional smoothers in practice. *J Comp Simul.* 18:419-439. 1994.

Makuc, DM., Haglund, B, Ingram, DD, et. al. *Health service areas for the United States*, Vital Health Statistics 2(112), National Center for Health Statistics, Hyattsville, MD, 1991.

Mungiole, M, Pickle, LW, Simonson, KH. Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine* (in press).

Pickle, LW, Mungiole, M, Jones, GK, and White, AA. *Atlas of United States Mortality.* Hyattsville, MD: National Center for Health Statistics, 1996.

List of Figures



(a) predicted

(b) pred+noise

(c) NN=4

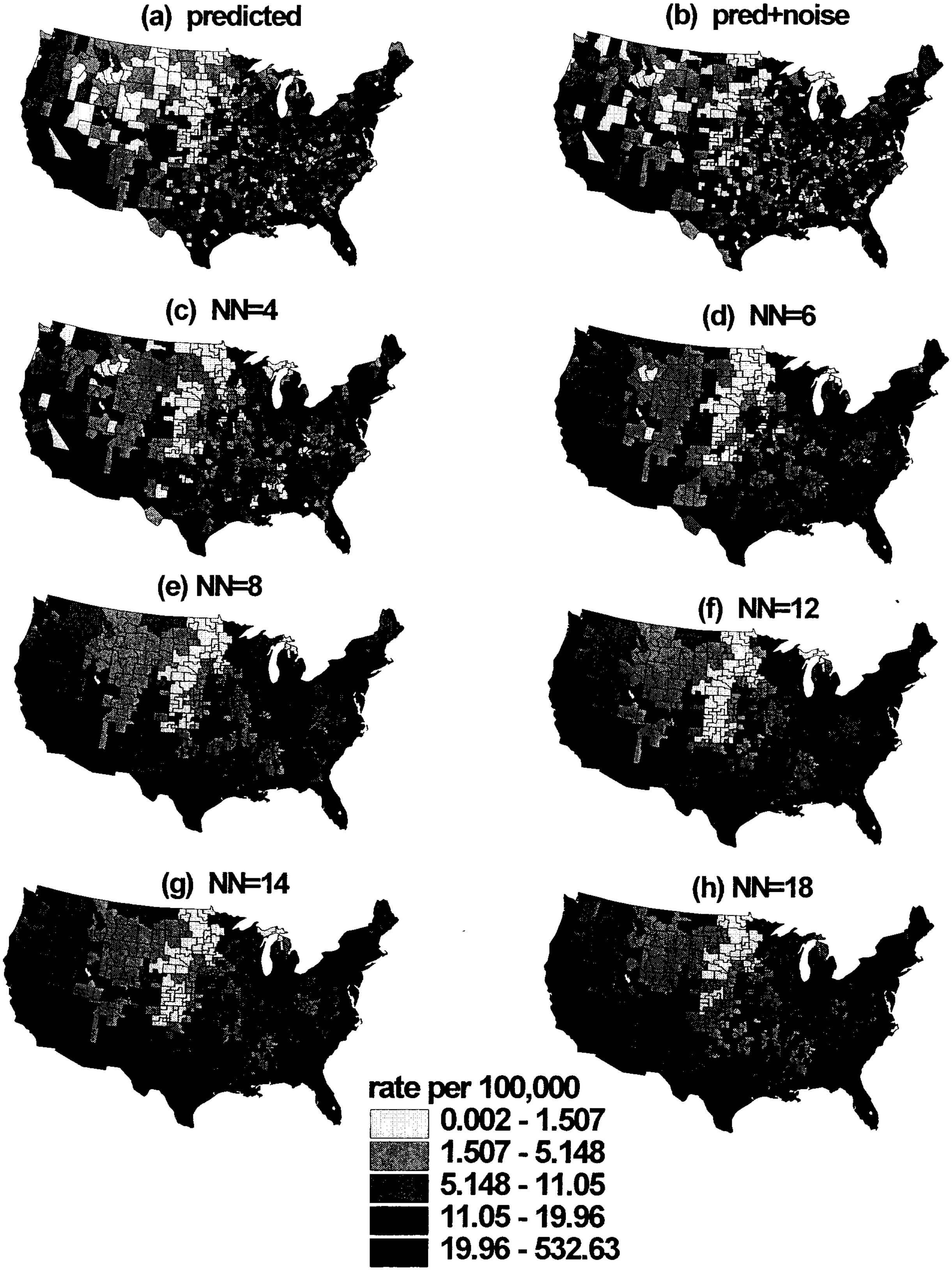
(d) NN=6

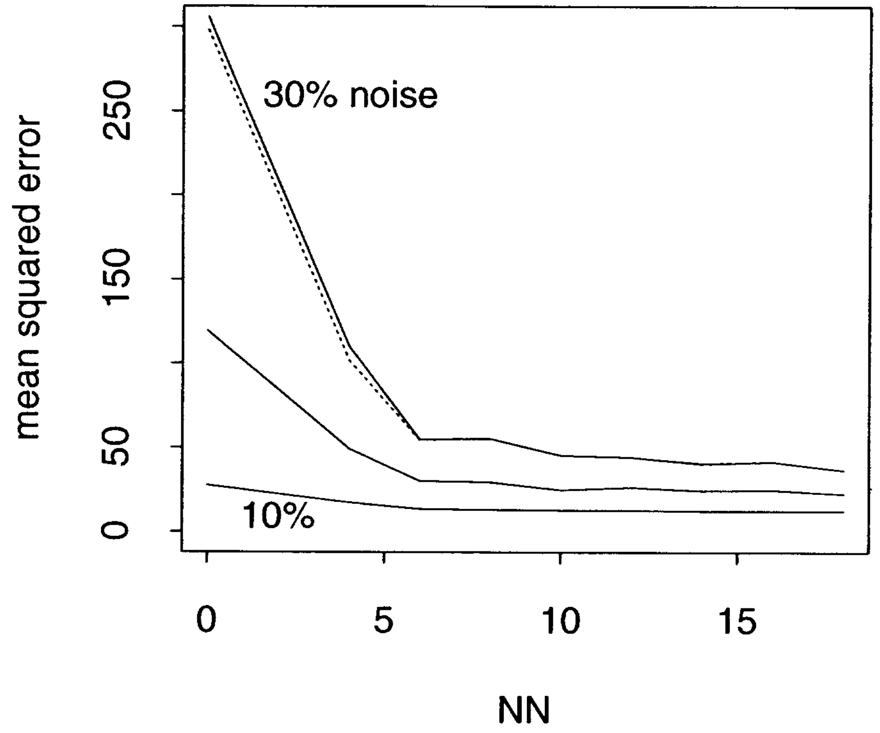
(e) NN=8

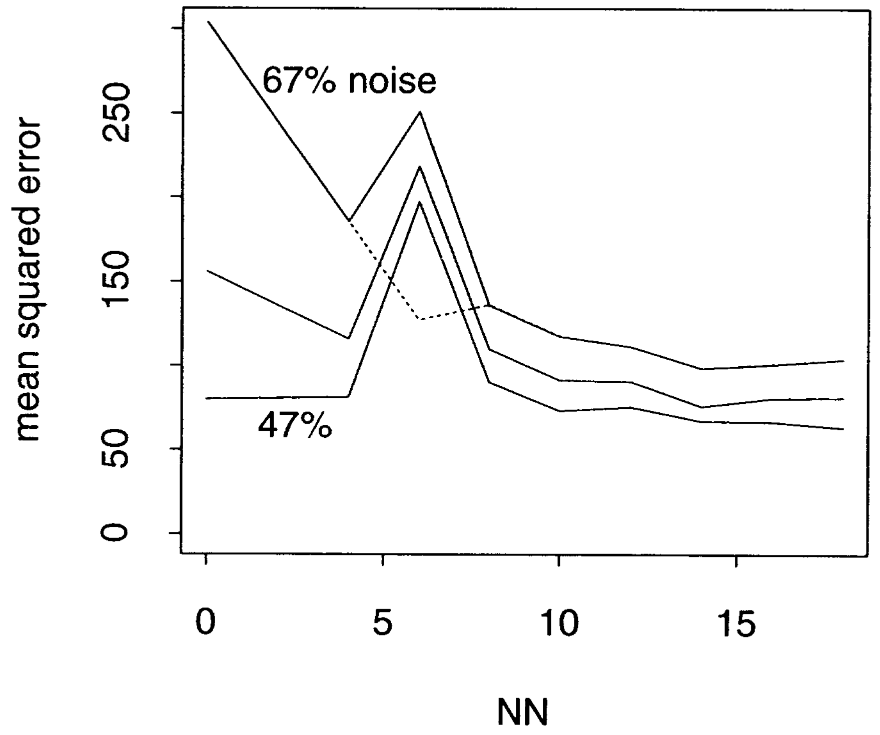
(f) NN=12

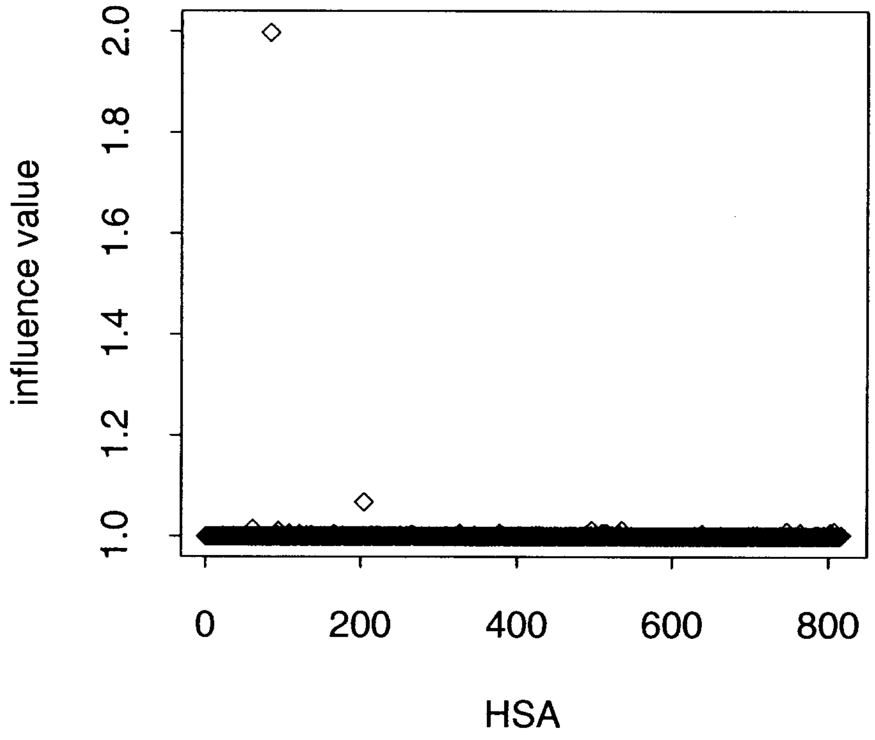
(g) NN=14

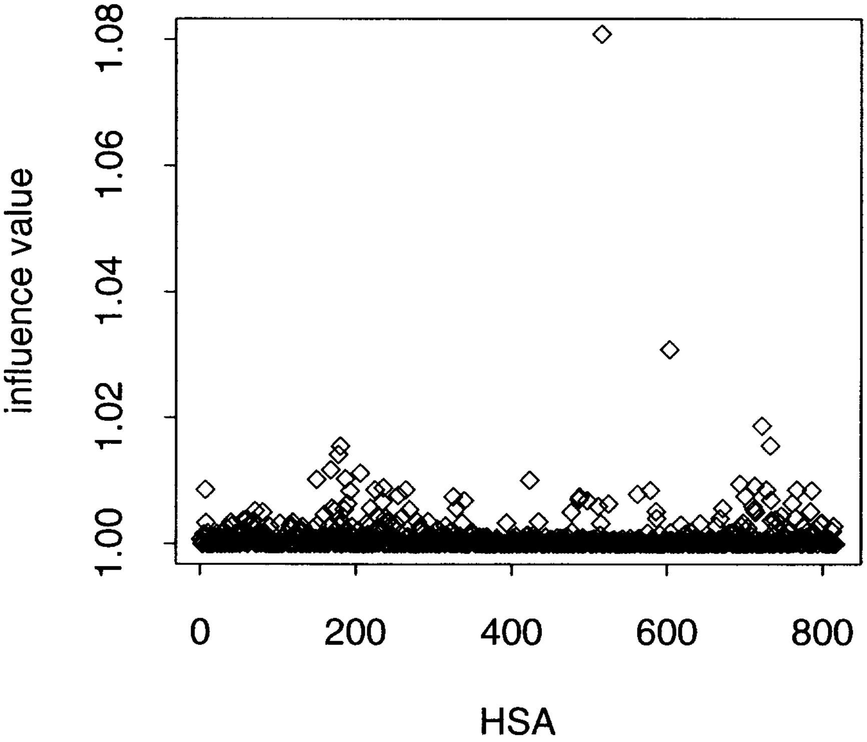
(h) NN=18

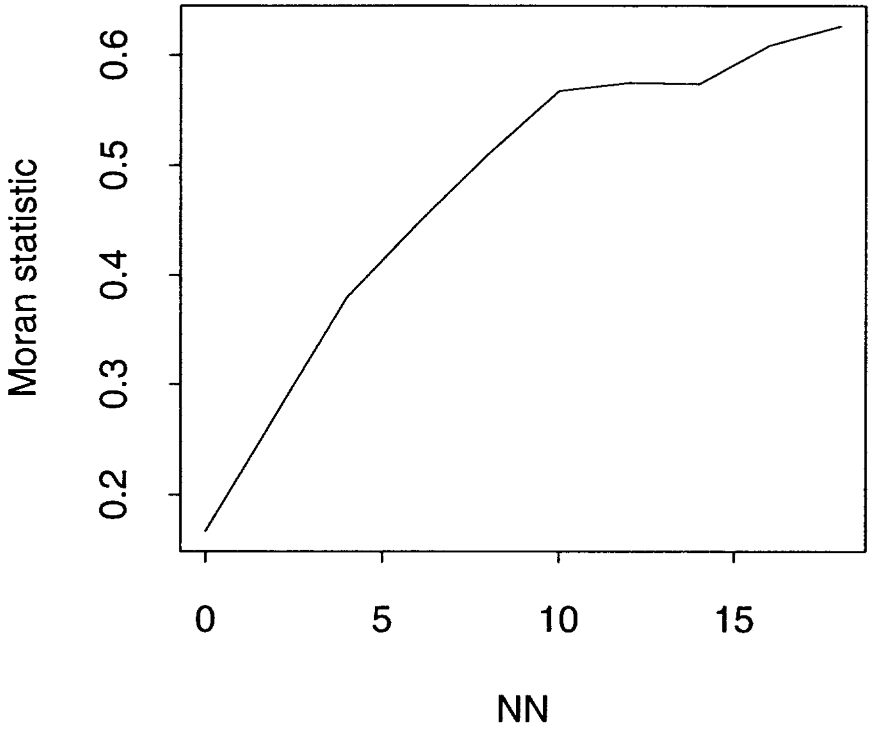

(a) predicted
(b) pred+noise
(c) NN=4
(d) NN=6
(e) NN=8
(f) NN=12
(g) NN=14
(h) NN=18
rate per 100,000
0.002 - 1.507
1.507 - 5.148
5.148 - 11.05
11.05 - 19.96
19.96 - 532.63

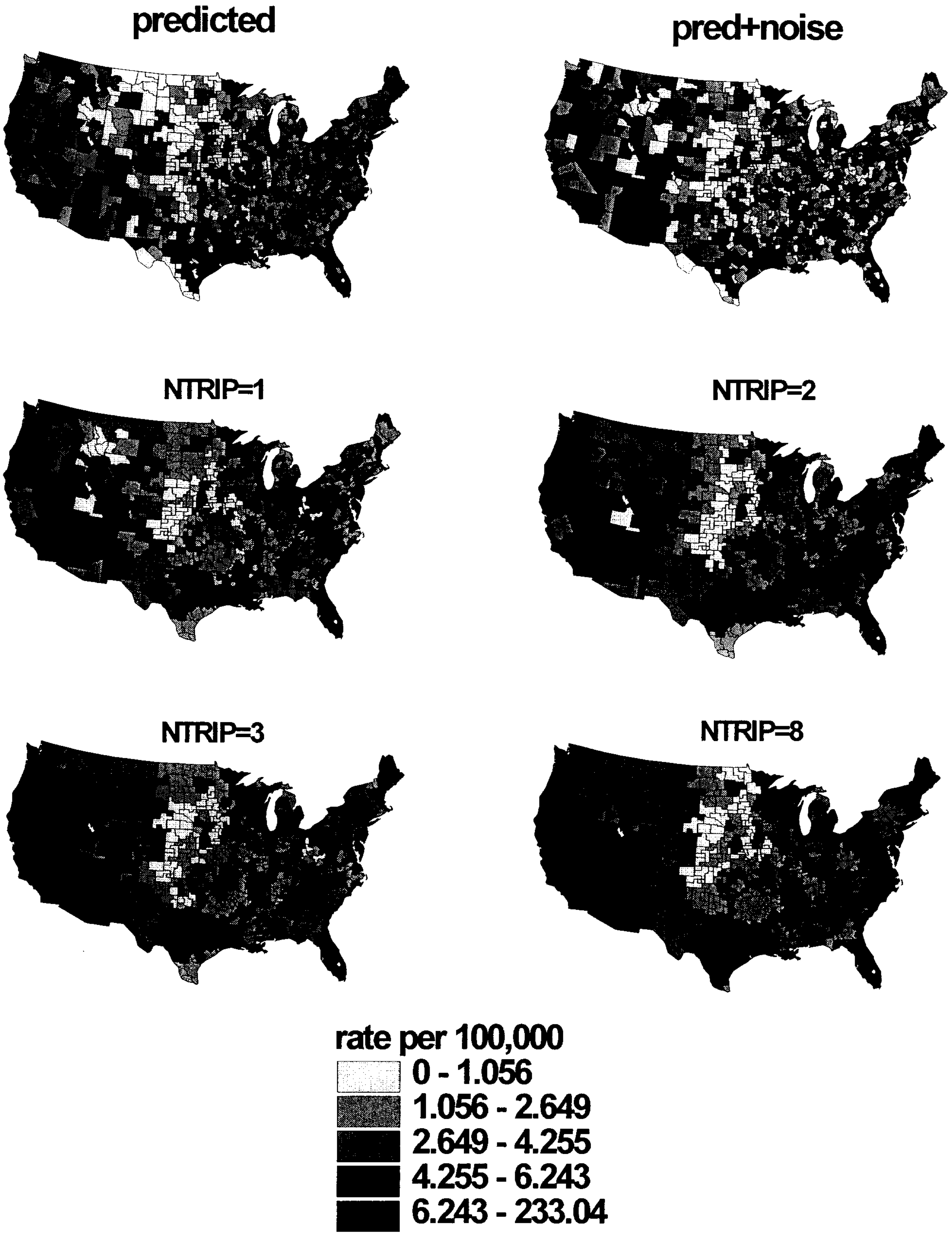

30% noise
10%
mean squared error
0
50
150
250
0
5
10
15
NN

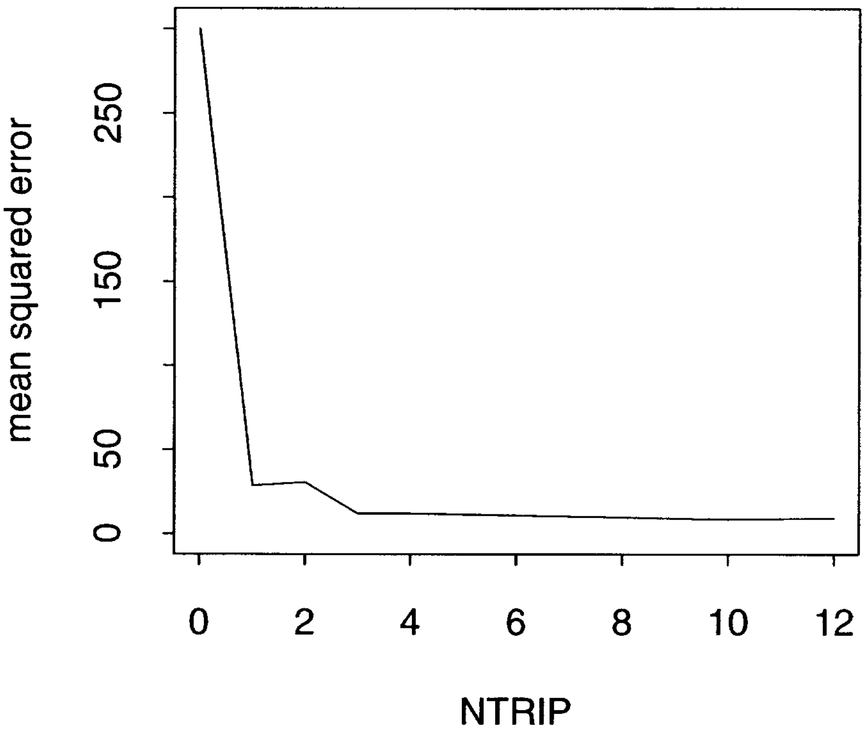

67% noise
47%
mean squared error
0
50
150
250
0
5
10
15
NN

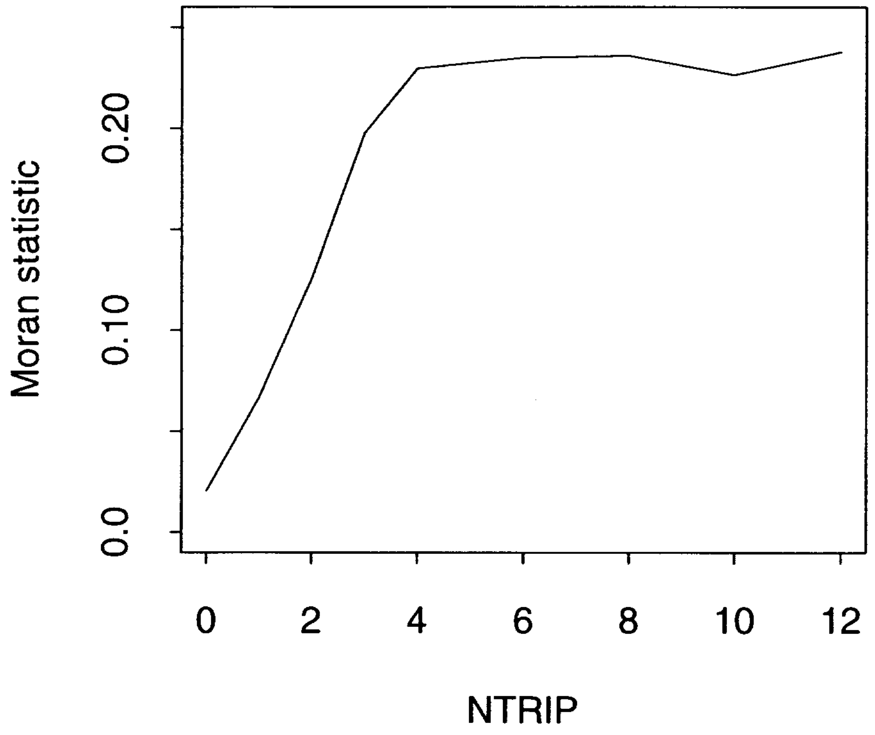

influence value
1.0
1.2
1.4
1.6
1.8
2.0
0
200
400
600
800
HSA


influence value
1.00
1.02
1.04
1.06
1.08
0
200
400
600
800
HSA


Moran statistic
0.2
0.3
0.4
0.5
0.6
0
5
10
15
NN


predicted
pred+noise
NTRIP=1
NTRIP=2
NTRIP=3
NTRIP=8
rate per 100,000
0 - 1.056
1.056 - 2.649
2.649 - 4.255
4.255 - 6.243
6.243 - 233.04


mean squared error
0
50
150
250
0
2
4
6
8
10
12
NTRIP


Moran statistic
0.0
0.10
0.20
0
2
4
6
8
10
12
NTRIP