

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 $\theta^*=135^\circ$ 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.

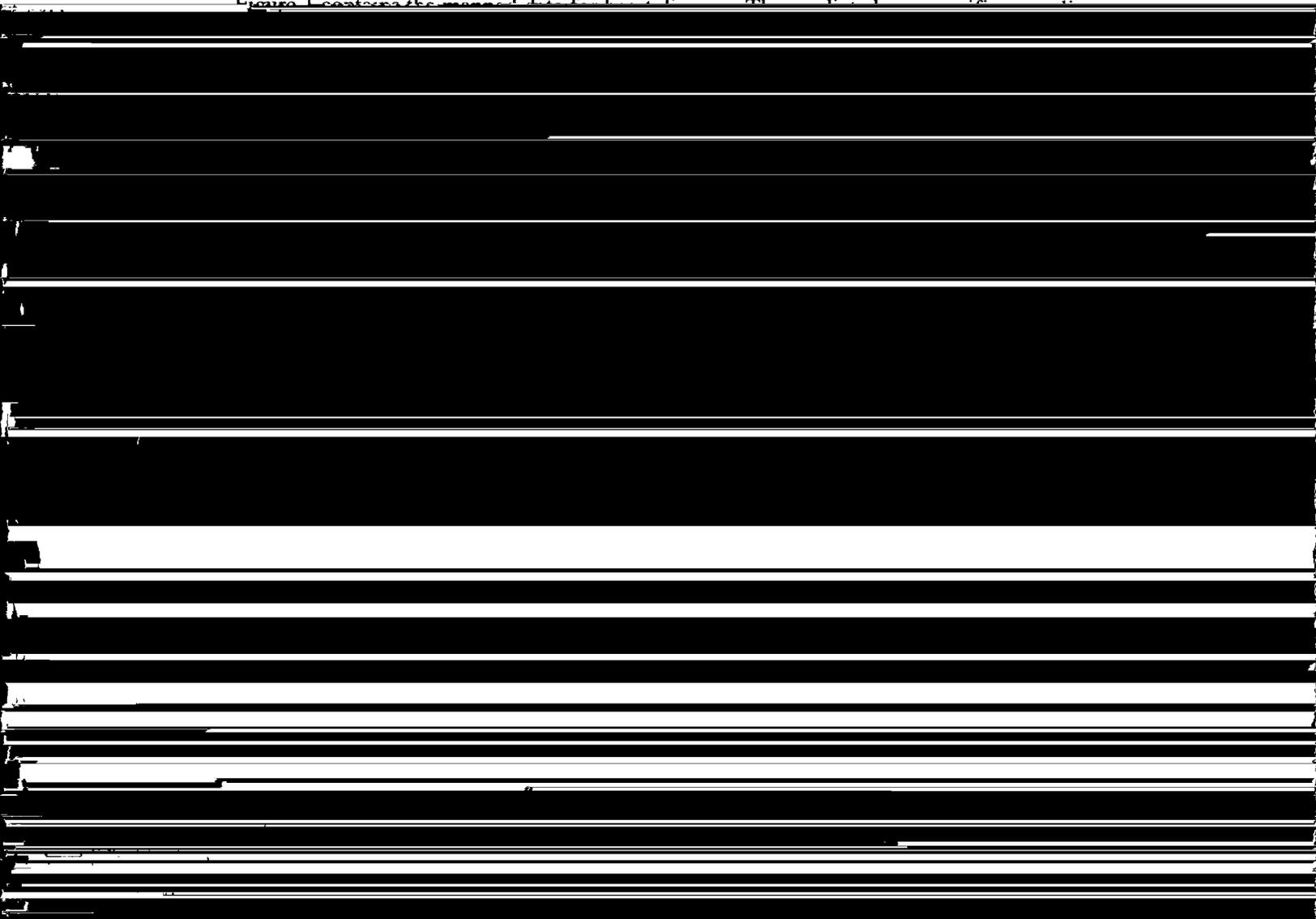
The process of selecting a criterion required that there be some appreciable change in the

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 mean and standard error of the predicted rates for heart disease and HIV for each of the three noise levels. The predicted rates are shown in the top row and the standard errors are shown in the bottom row. The predicted rates are shown in the top row and the standard errors are shown in the bottom row.



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 $NN \neq 6$ are used, the mortality rate for HSA number 83 is smoothed to values much closer to its predicted rate.

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

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 Statistician 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

Figure 1. Heart disease mortality rates for white males, ages 35-44. Predicted rates with added noise (30%) and smoothed results using various values of nearest neighbors (NN).

Figure 2. HIV mortality rates for white males, ages 35-44. Predicted rates with added noise (65%) and smoothed results using various values of nearest neighbors.

Figure 3. MSE by number of nearest neighbors for heart disease mortality rates for three levels of noise. The solid curves represent calculations including all HSAs and the dashed curve is for all HSAs except number 83 (67% noise).

Figure 4. MSE by number of nearest neighbors for HIV mortality rates for three levels of noise. The solid curves represent calculations including all HSAs and the dashed curve is for all HSAs except number 516 (30% noise).

Figure 5. Influence values for HIV mortality rates for each HSA.

Figure 6. Influence values for heart disease mortality rates for each HSA.

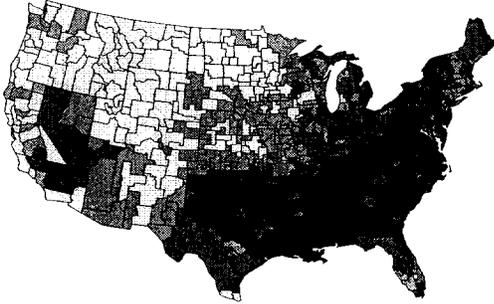
Figure 7. Moran statistic by number of nearest neighbors for heart disease mortality rates for white males, ages 35-44.

Figure 8. Suicide mortality rates for white females, ages 65-74. Predicted rates with added noise (117%) and smoothed results using various values of maximum number of triples.

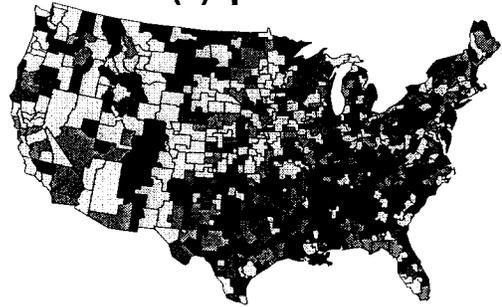
Figure 9. MSE by maximum number of triples for suicide mortality rates for white females, ages 65-74.

Figure 10. Moran statistic by maximum number of triples for suicide mortality rates for white females, ages 65-74.

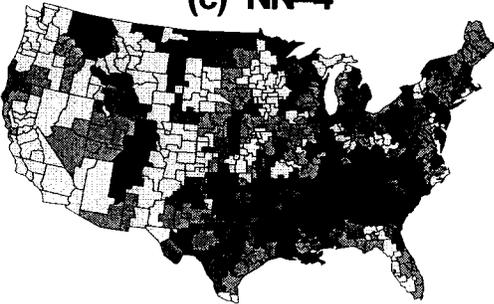
(a) predicted



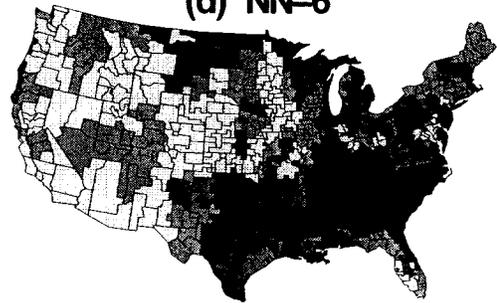
(b) pred+noise



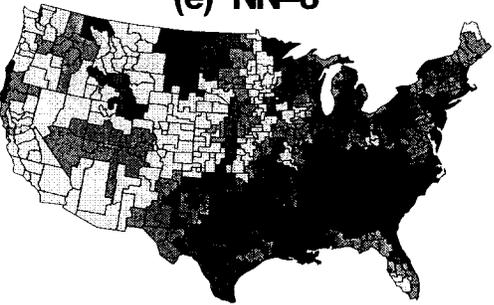
(c) NN=4



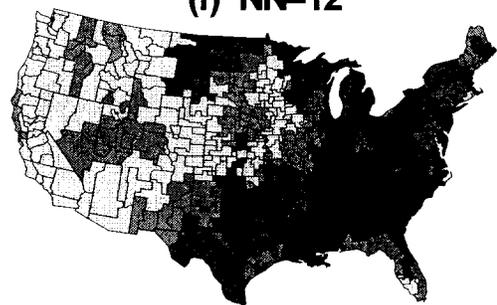
(d) NN=6



(e) NN=8



(f) NN=12



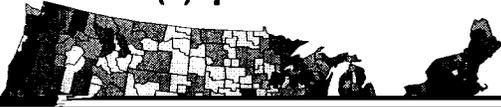
(g) NN=14



(h) NN=18

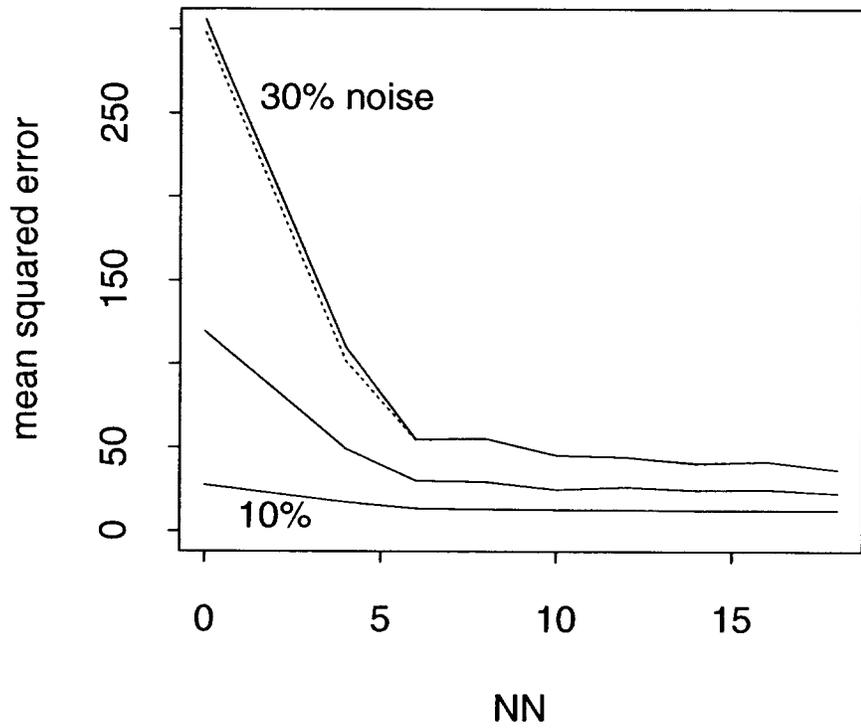


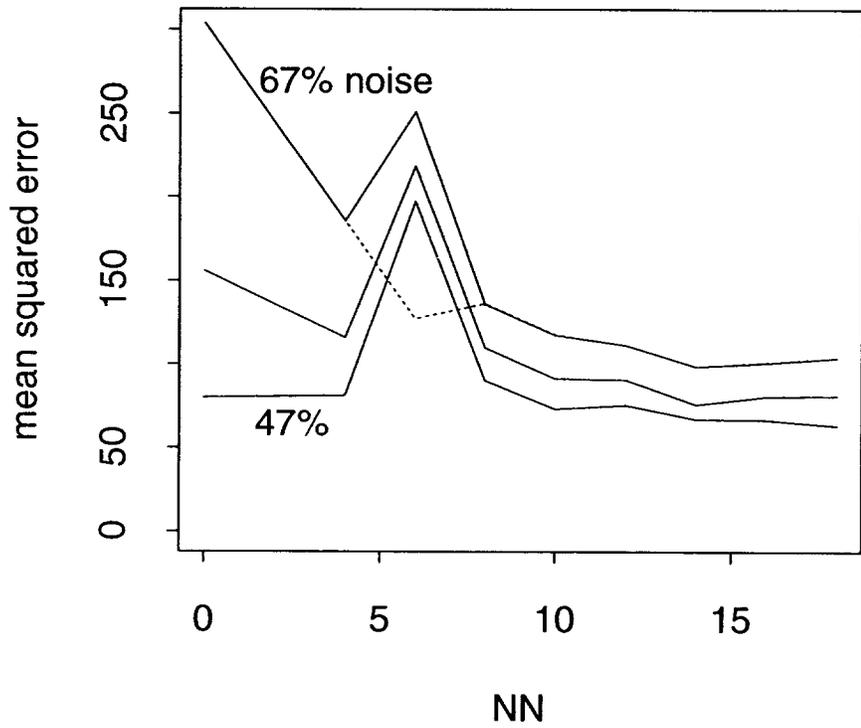
(a) predicted

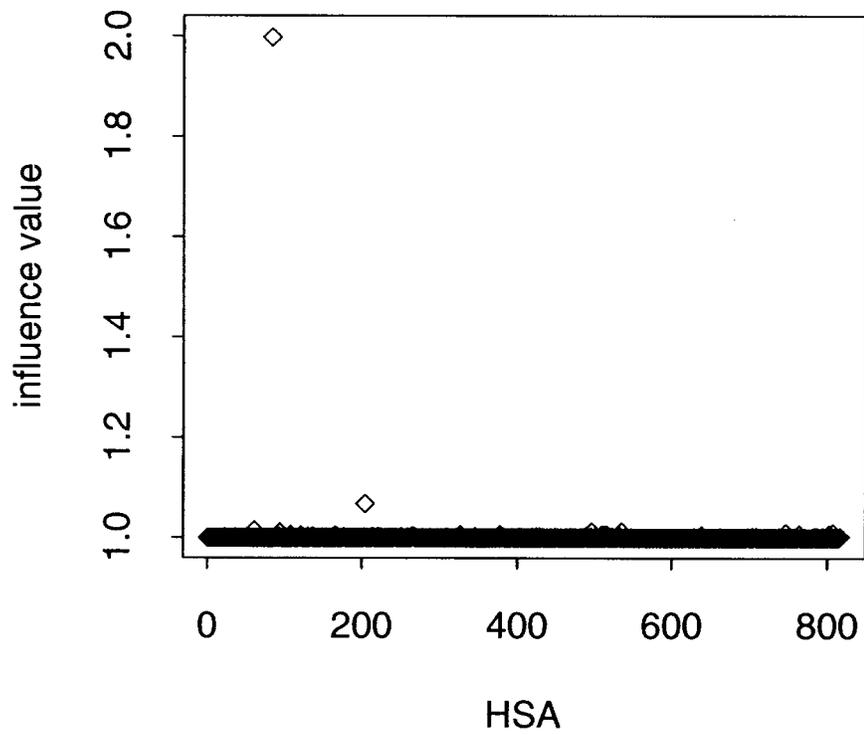


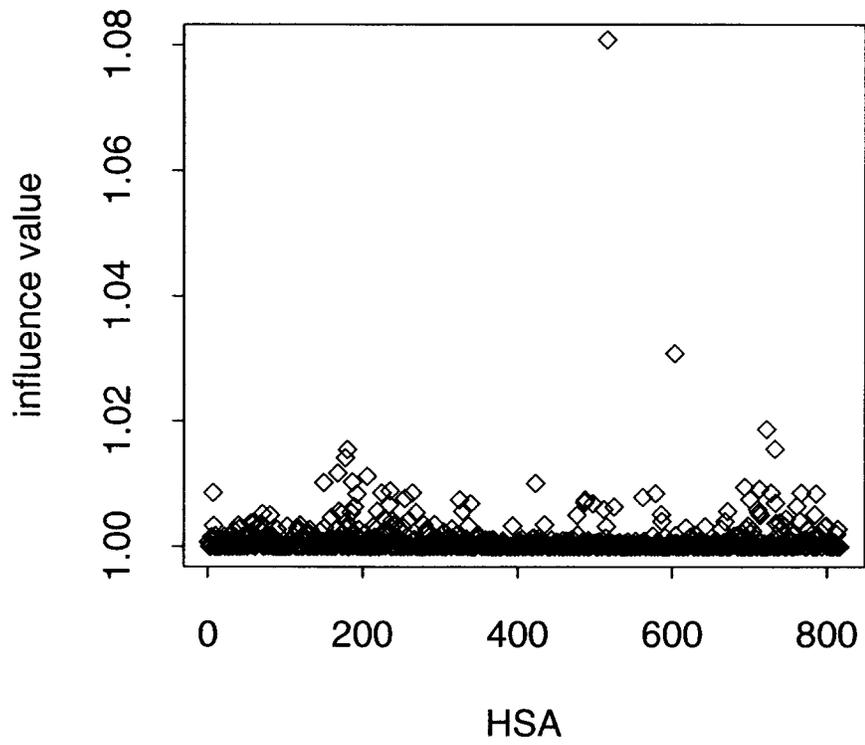
(b) pred+noise

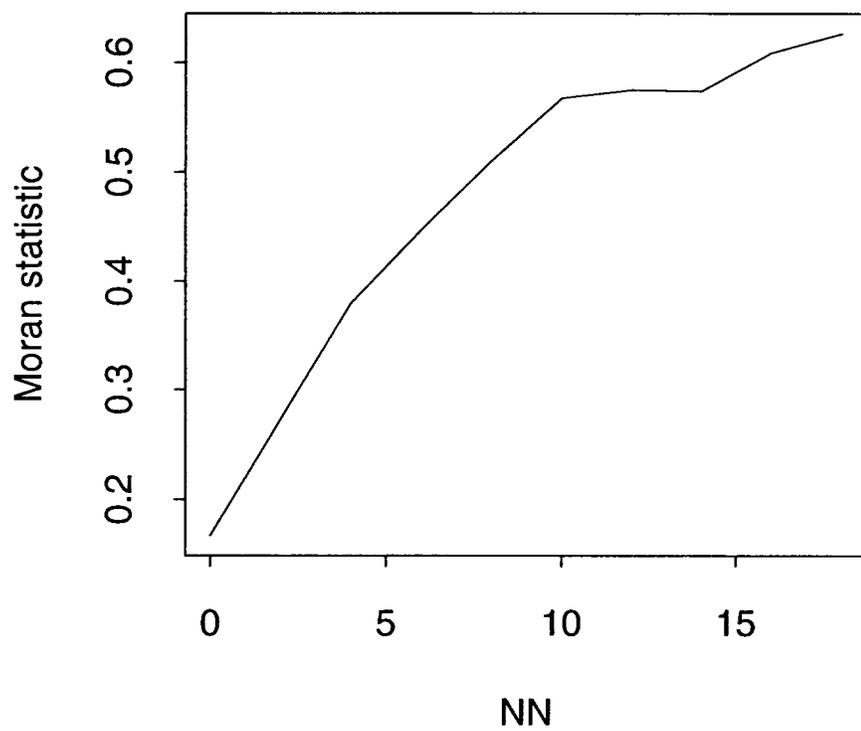












predicted



pred+noise



