ISSN 1946-7664. MCFNS 2021
Vol. 13, Issue 2, pp. 27–30
Mathematical and Computational
Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Feb. 2, 2020
Accepted: Oct. 30, 2021
Published: Oct. 30, 2021
Last Correction: Nov. 21, 2021

NEAREST NEIGHBOR BIAS IN THE SUBSTITUTION OF MISSING VALUES

Chris J Cieszewski[1*], Kim Iles[2]
[1] *University of Georgia, Athens, GA, USA.*
[2] *Kim Iles & Associates, Nanaimo, B.C., Canada.*
[*] *Corresponding Author*

Abstract. This is a short technical note illustrating the bias inherent in the general case of the Nearest Neighbor (NN) method used to substitute missing values. This presentation doesn't make any assumptions about the geometry of the sampled subjects. The general examples illustrate that the bias exists mainly at the limits of the data range and not necessarily within the center part of the range. However, the latter is also possible around any significant data gaps. The NN data domain stretches across an arbitrary subject characteristic rather than across the physical space. It is possible to reduce the discussed here biases by assuring that the domain range of the considered attribute is well-represented within its entire range, especially at its upper and lower limits and there are no major gaps in the training data.

Keywords: Mapping; Nearest Neighbor Imputation; Nearest Neighbor Bias; Large-area Forest Inventories; Multi-source Data Fusion.

1 Introduction

The basic principle of the Nearest Neighbor (NN) method is substituting a missing attribute of an unmeasured subject with the corresponding attribute of a similar measured subject. Some arbitrary proxy criteria correlated with the attribute determine the similarity between the subjects. The meaning of the nearest neighbor relates, in this case, to the degree of quantitative or qualitative similarity. It does not relate to the physical or geographic proximity, regardless of the correlation between the spatial distance and the physical similarities of various individuals. For example, the similarities in spectral signatures of LTM pixels can be the proxy criteria for assigning stand attributes in locations with ground measurements to the unmeasured stands corresponding to similar spectral signatures (e.g., Lowe and Cieszewski, 2014).

In another sampling example, Iles (2010) points out an inherent bias in the sampling-based on spatial nearest neighbors (i.e., defined by physical distance) for estimating the population mean (see also Czaplewski 2010). Iles uses an example of a line divided into three segments, A, B, and C, with unequal probabilities of selection, to describe the geometry of the problem. While Iles (2010) discussed the NN method sharing similar naming to what we discuss here, these two methods are different and have different bias properties.

The NN substitution method is well known and discussed in so many studies that it doesn't need much introduction. Scientists have also experimented with other techniques to contrast and enhance its outcomes (e.g., Haara and Kangas, 2012 and Magnusen et al., 2010). Some authors have discussed different aspects of potential biases associated with estimates derived using this method with various training data (Tomppo et al., 2009). McRoberts (2009) specifically considered the type of bias that occurs at the limits of the data ranges. He states accurately, "Nearest neighbors techniques are inherently biased because no prediction may be smaller or larger than the smallest or largest response variable observation in the reference set, respectively." This statement captures the essence of the problem at the outer edges of the data sets. However, it does not apply in every situation. It doesn't explicitly illustrate the bias calculations' mechanism, which can also apply to significant data gaps within the data range. The essence of the bias mechanism is strictly not the fact that there are no higher or lower values but the predicament in which the selection of the nearest neighbors is one-sided predisposed towards choosing the nearest neighbors only on the lower, or only on the higher, range of the values. Similar bias will occur when there are both lower and higher values, but either of them is too dissimilar to be selected as the nearest neighbor. Such a situation is likely to appear at the edges of any significant data gaps. Whenever a considerable data gap exists, the exact mechanism will apply around the void as if two or more disjointed datasets were used together, each with its maximum and minimum values.

This note symbolically shows the principles on which the bias is present in the NN substitution imputation, in the method used to substitute values by non-spatial nearest neighbors defined by any arbitrary proxy criteria.

2 The Bias Argument

Let the sampled population of the subjects be: A, B, and C. Regardless of their spatial distribution, A and C are the nearest neighbors of B, and vice verse, B is the nearest neighbor of both A and C, but A and C are not the nearest neighbors to each other. Therefore, as illustrated in Table 1, the attributes of central values, which is B, are assigned more frequently than those of peripheral values, such as A and C. Hence, the bias. The question is: what one would substitute for each missing value?

The pattern of bias with more subjects (e.g., Tab. 2) is that the central estimates are unbiased, but two peripheral estimates on each end are biased. At the lower end of the value range the estimates are overestimated, and at the higher end of the value range the estimates will be underestimated.

With a significant data gap, the situation is likely not to have any nearest neighbors on either side of the opening. If the values between D and X are missing (e.g., Tab. 3), D is unlikely to be the nearest neighbor to X because the distance between them is too great.

Table 1: NN substitution for missing values with three subjects. B gets selected as the nearest neighbor when either A, or C, is missing. Half of A and half of C comprise the nearest neighbor when B is missing. Thus, B is assigned more often than A and C. The result is changing the actual A+B+C to $\frac{A}{2}$+2B+$\frac{C}{2}$. The total might still be unbiased, but the outer estimates are not.

VALUE	SUBJECTS			TOTAL
Missing	A	B	C	A+B+C
Substituted	B	$\frac{A}{2}+\frac{C}{2}$	B	$\frac{A}{2}+2B+\frac{C}{2}$

The value of Y then will be asymmetrically assigned to substitute the missing X values. Analogically, X is unlikely to be the nearest neighbor of D, causing disproportionate assignments of C for the value of D. Hence the bias at the edges of the data gap.

3 Discussion

The NN method has gained much popularity, especially in large-area forest inventories. It is one of the most applicable methods for propagating information from inventory plots onto satellite imagery and combining data from different sources of spatially related sampling at varying levels of resolutions (e.g., Iles 2009). The method has universal applicability. Especially with the natural resource inventories and mapping, it may be likely the critical method of future operational and research analysis. The illustrative example presented here shows the bias mechanism in the NN process that is mainly at the limits of the estimated data range, not represented by either identical or surrounding (both greater and smaller) measurements. The exact bias mechanism will occur for the attribute values corresponding to any significant data gaps in the training datasets.

Table 2: NN substitution for missing values with over three subjects with no data gaps. With more subjects, two peripheral values on each end are biased. The subjects at each end are underrepresented, and their closest neighbors are overrepresented. The rest of the inner estimates are unbiased.

VALUE	SUBJECTS									TOTAL
Missing	A	B	C	D	. . .	W	X	Y	Z	A+B+C+...+X+Y+Z
Substituted	B	$\frac{A}{2}+\frac{C}{2}$	$\frac{B}{2}+\frac{D}{2}$	$\frac{C}{2}+...$	. . .	$...+\frac{X}{2}$	$\frac{W}{2}+\frac{Y}{2}$	$\frac{X}{2}+\frac{Z}{2}$	Y	$\frac{A}{2}+\frac{3B}{2}+C+...+X+\frac{3Y}{2}+\frac{Z}{2}$

Table 3: NN substitution for missing values with over three subjects and a significant data gap between the subjects D and W. In this example, all estimates are biased. A, D, W, and Z are underrepresented, and B, C, X, and Y are overrepresented.

VALUE	SUBJECTS									TOTAL
Missing	A	B	C	D	. . .	W	X	Y	Z	A+B+C+D+...+W+X+Y+Z
Substituted	B	$\frac{A}{2}+\frac{C}{2}$	$\frac{B}{2}+\frac{D}{2}$	C	. . .	X	$\frac{W}{2}+\frac{Y}{2}$	$\frac{X}{2}+\frac{Z}{2}$	Y	$\frac{A}{2}+\frac{3B}{2}+\frac{3C}{2}+\frac{D}{2}+...+\frac{W}{2}+\frac{3X}{2}+\frac{3Y}{2}+\frac{Z}{2}$

Removing the bias might be difficult. First, the multidimensionality of the similarity criteria, defined by multiple parameters, such as the various spectral signatures of the LTM images, makes the problem intractable. Second, the training data for the NN method is naturally much more limited than the estimated data sets. However, the bias will not occur if the training data represents a broader range of values than the calculated data. Conversely, the bias at the limits of the estimated data set will increase along with its dept in subjects if the training data represents a substantially narrower range of values than the calculated data. For example, if the training data comprised all the values from A to Z, while the imputations covered only subjects from B to Y, there would be no bias in the imputation. There are also possibilities of enhancing the robustness of the NN imputations with other statistical procedures, such as in Magnussen et al., (2010) suggesting a model-assisted solution.

Examining Table 1 may suggest that using a 2-neighbor averaging would mitigate the bias; however, this would cause a bias correction only in the presented simplified scenario of three subjects. It is unlikely that multiple neighbors would make the bias smaller. Various scientists use different numbers of the nearest neighbors. The consensus is that several neighbors are better than one. In our experience, in the application of the NN method to satellite images, the accuracy of the NN imputation deteriorates rapidly with increasing distances of the nearest neighbors—usually measured with the Euclidean distance defined by spectral values of different bands used for the analysis. Close neighbors are good predictors, but distant neighbors are disproportionably worse.

We deliberately simplified the presented here examples to illustrate the bias mechanism clearly. These examples assume an even-spaced representation of the known subjects, a linear correlation between the estimated attribute and the proxy criteria, and the availability of complete sets of training data. The distance between A and B is the same as between B and C, and so on, for all consecutive subjects. In implementations of the NN method, the distances of the nearest neighbors, measured by Euclidian distances, have absolute values and are indifferent to the smaller or larger neighbors. B is the nearest neighbor of A and C because, in the proxy criteria's Euclidian space, the distance between A and B is the same as between C and B. In the absence of C, only B is the nearest neighbor because the distance between A and B is smaller then the distance between A and D.

If two neighbors need to be selected, in the absence of A, the nearest neighbors of B are C and D rather than A and C. The problem is impossible to visualize in multidimensional spaces. For the same reason, it doesn't lend itself to intuitive reasoning as it is, for example, with spatial neighbors. However, the hypothetical examples provided here should help the understanding of the bias mechanism. We do not intend to invalidate the description of the bias provided by McRoberts (2009) but complement it. We give examples to make the phenomena more understandable and, hopefully, more intuitive. The understanding of the bias mechanism should help with a remedy for better training data preparation, which is critical in the bias creation.

Acknowledgement

Thanks are due to the two editors handling this manuscript and four anonymous reviewers, who provided helpful comments on the early version of this manuscript.

References

Czaplewski, R. 2010. Review of Nearest Neighbor Bias—A simple example. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS), 2(1):66–66. Retrieved from https://mcfns.net/index.php/Journal/article/view/MCFNS.2-66

Haara, A., & Kangas, A. 2012. Comparing K Nearest Neighbours Methods and Linear Regression—Is There Reason To Select One Over the Other?. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS), 4(1):50–65. Retrieved from https://mcfns.net/index.php/Journal/article/view/MCFNS.4%3A50

Iles, K. 2010. Nearest Neighbor Bias—A simple example. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS), 2(1):18–19. Retrieved from https://mcfns.net/index.php/Journal/article/view/MCFNS.2-18

Iles, K. 2009. "Total-Balancing" an inventory: A method for unbiased inventories using highly biased non-sample data at variable scales. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS), 1(1):10–13. Retrieved from http://mcfns.com/index.php/Journal/article/view/MCFNS.1-10

Lowe, R., & Cieszewski, C. 2014. Multi-source K-nearest neighbor, Mean Balanced forest inventory of Georgia. Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS), 6(2):65–79. Retrieved from https://mcfns.net/index.php/Journal/article/view/6_65/184

Magnussen, S., Tomppo, E., & McRoberts, R. E. 2010. A model-assisted k-nearest neighbour approach to remove extrapolation bias. Scandinavian Journal of Forest Research, 25(2):174–184.

McRoberts, R. E. 2009. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sens. Environ., 113(3):489–499.

Tomppo, E. O., Gagliano, C., De Natale, F., Katila, M., & McRoberts, R. E. 2009. Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. Remote Sens. Environ., 113(3):500–517. DOI:10.1016/j.rse.2008.05.021.