Vol 2, No 1, pp. 18–19, 2010
International Journal of
Mathematical and Computational
Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Nov. 24, 2009
Accepted: Feb. 14, 2010
Published: Feb. 28, 2010
Last Correction: Feb. 17, 2010

NEAREST NEIGHBOR BIAS

A SIMPLE EXAMPLE

Kim Iles

Kim Iles & Associates Ltd., 412 Valley Place, Nanaimo, BC, Canada. Ph.&FAX: 250.753.8095

Abstract. This is a short research note with an illustrative example describing the bias inherent in the nearest neighbor method.

Keywords: nearest neighbor, bias.

1 Background

You still hear, occasionally, the comment that nearest neighbor methods (there are many, of course) can be used to estimate totals, and that they are unbiased when doing so. It seems as if information to the contrary does not get around as fast as the desire for this to be true. While one can, of course, create a situation where it might be true, it is certainly not generally true.

As a very simple example, which of course applies in more dimensions, consider the following situation of sampling on a line where there are 3 possible items of different values that might be chosen (Figure 1). The arrows delimit the areas (thereby defining the probability of selection) where the item would be chosen using a random point along the line. There are 8 units (x) of length to the total line from border to border.

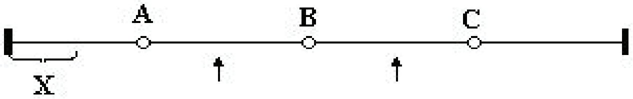

Figure 1: Illustration of sampling on a line with 3 possible items of different values that might be chosen.

The total for the line is obviously A+B+C. What is the probability of choosing A, B or C? In this example, the distance between the points and end points is equal (2x), so the probabilities and the total they calculate are easy to compute as:

Total = 3/8(A) + 2/8(B) + 3/8(C).

This is clearly not the correct total (or average per unit of length) in this simple example where the items are evenly distributed from each other. In other situations, where distances vary, the biases might be much worse. In forest sampling there is an obvious and well known bias in selecting the "nearest tree" for measurement which causes a bias toward isolated trees. Choosing the nearest tree from some direction (including a random direction) has a similar bias.

Is there any reasonable expectation that this kind of bias will disappear when more dimensions are used? It is easy to expand this example to two and three dimensions with this same reasoning, but might the problem be resolved at some point? I suspect not,; however I have not included a proof.. If the problem is mostly caused by borders, think of the modern situation where the forest map is cluttered with borders that become increasingly more sinuous and with increasingly larger ratios of edge to volume.

When faced with a bias, there are several classical ways to approach the problem.

1) Create a different design that eliminates the bias. In many situations, this can be done. There are forms of biased and unbiased ratio estimators, for instance (Cochran, 1977). Finding an unbiased form is often quite restricting in terms of practical application.

2) Determine the bias (or perhaps an unbiased estimate of it) and correct for it. A previous paper in this journal on "total balancing" would perhaps be an example of this (Iles, 2009). There are also some classic examples of this method in ratio estimation.

It seems to me that this is the most promising of the methods, since it allows the full range of practical issues to be dealt with, and has the greatest number of possible solutions to special cases.

3) Determine the probable extent of the bias, and if it is small enough then you can simply accept it. This is often done by simulations, and in rare cases with direct determinations; for example, calculations that would place an upper limit on the possible bias – again there are ex-

amples with ratio adjustments – always keeping in mind that there are assumptions involved which might not hold.

The problem with simulations, of course, is to convince people that the simulation used is equivalent, in terms of the bias effects to their own sampling situation – often in quite a different field.

In re-reading one of George Furnival's papers this year (perhaps an unpublished one, I have not been able to relocate it) I was struck by the economy and clarity of his view that forest inventory was almost entirely about 3 questions, each separate.

1) How much do you have?

2) Where is it?

3) How is it changing?

Surely these questions constitute, by far, the major part of forest inventory issues, and they are unlikely to be solved most efficiently in the same way. Perhaps a good way to view the problem is simply to solve each one efficiently, while using the others to limit that solution in ways that have theoretically valuable properties. If nearest-neighbor methods are the best way to distribute information, perhaps controlling the overall or group totals is the way to do this in a way which will eliminate bias.

Acknowledgements

I would also like to thank several anonymous reviewers for their input.

References

Iles, K. 2009. "Total-Balancing" an inventory: A method for unbiased inventories using highly biased non-sample data at variable scales. MCFNS [Online]. 1(1): 10-13.

Cochran, W. G. 1977. Sampling Techniques. 3d ed. New York: John Wiley and Sons. 428 p. See section 6.15.