Conditioning Inference on Line Orientation in Line Intersect Sampling [1]

1 Introduction

The literature on line intersect sampling (LIS) spans many decades and disciplines dating back at least to 1898 in an article by Rosiwal discussing the ascertainment of the quantitative contents of the mineral constituents in rocks. Of course, its similarity to the famous needle problem posed by Buffon in the 1770's is well known.

It was introduced to ecologists in a 1936 monograph by Bauer, but Canfield's 1941 "Application of the Line Interception Method in Sampling Range Vegetation" in a forestry journal is widely credited as the first quasi-statistical look at LIS within the field of ecology.

The probabilistic properties of LIS can be deduced geometrically, per the famous monograph by Kendall and Moran (1963) and appreciated by the renowned Swedish forest-statistician Bertil Matérn in his 1964 note on estimating the total length of roads by means of a line survey.

Our assessment of the literature on LIS (a bibliography with 164 entries is posted at http://www.yale.edu/forestry/gregoire/biblo.html) is that the basis of inference following its use to estimate descriptive parameters of a population is often unclear. This lack of clarity has resulted in abundant confusion to applied scientists who may not be aware of the distinction between design-based and model-based statistical inference. Indeed, a confusion is apparent in published works by forest biometricians, which was one motivation behind the 1998 article by Gregoire: "Design-based and model-based inference in survey sampling: appreciating the difference" published in the *Canadian Journal of Forest Research.*

Our remarks on conditional inference are prefaced with this historical overview in order to credit Kaiser (1983) properly as the conduit to our understanding of estimation following LIS from the design-based perspective. We have recently investigated the statistical properties of estimators – proposed or implicit – when using line transects with two or more segments (Gregoire and Valentine, 2003, henceforth GV03; Affleck, Gregoire, and Valentine, 2005, henceforth AGV05). [2]

[1]This research was supported by funds from the USDA Forest Service, Northeastern Research Station, RWU-4104, through a cooperative agreement between the School of Forestry and Environmental Studies at Yale University.

[2]We would be remiss if we neglected to mention current and very intriguing work being done in this area by L. Barabesi and colleagues at the University of Siena, Italy. At the Seventh International Meeting on Quantitative Methods for Applied Sciences, Siena, September 23-24, 2004, Barabesi and Marcheselli will present "Improved strategies for coverage estimation by using replicated line-intercept sampling". A pdf version of this paper can be obtained by emailing barabesi@unisi.it.

Our remarks today are restricted, however, to an inferential nuance made explicit by Kaiser (1983) which has gone under-appreciated, indeed barely noticed by many users of LIS.

2 LIS sampling design

Let the population of interest consist of discrete and stationary objects (trees, bushes, logs, rocks, dens, scat, etc.) which we call particles for sake of generality. Using the notation adopted in AGV04, the population is the collection $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ of discrete particles distributed over a region whose projection onto the horizontal plane is $\mathcal{A}$, the area of which is denoted by A. Each particle must be connected in the sense that any two points in P_i can be joined by a path contained entirely in P_i. In other words, a disjoint object constitutes one distinct particle. Particles can have any shape: they need not be circular nor needle-shaped with a central axis; particles need not be convex; and can be forked or have void interior regions. Furthermore, nothing about the spatial distribution of $\mathcal{P}$ on $\mathcal{A}$ is assumed: particles may be spread and oriented in any manner physically possible over $\mathcal{A}$ and may overlap in their projections onto $\mathcal{A}$.

Let y_i be a fixed, measurable characteristic of P_i that is independent of whether or how that particle is intersected by a transect. The target parameter to be estimated may be the aggregate quantity

$$\tau_y = \sum_{i \in \mathcal{P}} y_i \ ,$$

or perhaps the amount per unit area:

$$\lambda_y = \frac{\tau_y}{A}.$$

Evidently, when $y_i = 1, \forall i$ by definition, $\tau_y = N$, and hence population size and density are encompassed as special cases of τ_y and λ_y.

The sample design consists of M replicated transects on $\mathcal{A}$, the m^{th} of which is indicated by $T_m, m = 1, \ldots, M$. The total length of each transect is L. For the moment, assume each transect comprises a single straight-line segment. The location of T_m is referenced by the coordinate pair (x_m, z_m) on the horizontal plane of $\mathcal{A}$. By convention (x_m, z_m) is usually the midpoint or an endpoint of the transect. Transect locations are presumed to be selected independently and uniformly at random in $\mathcal{A}$. Systematic location introduces the usual limitations on variance estimation, but otherwise no essential incumbrances.

Relative to some reference direction, say $\theta = 0$, let θ_m denote the orientation of T_m. The orientation of T_m may be selected at random from a defined distribution, such as the uniform ($\theta_m \sim U[0, 2\pi]$). We allow for the possibility that transect orientations are selected independently for each of the M transects, or for the simpler case where an orientation is selected randomly and used for all transects. In the latter case, $\theta_m = \theta, \forall m$, say. Alternatively, the orientation of the transect may be fixed purposefully in advance of sampling, in which case $\theta_m = \theta, \forall m$, necessarily.

Irrespective of the choice of transect orientation, let $w_i(\theta_m)$ denote the width of P_i in a direction perpendicular to θ_m.

If P_i is intersected by T_m, P_i is included into the sample. Nothing precludes the intersection of P_i by two or more transects, although this event can be expected to have small probability of occurrence. We have dealt with partial and multiple intersections of P_i earlier (GV03, AGV04). Gregoire and Monkevich (1994) presented the reflection method to deal with reduced sample support near the edge when using LIS with single, straight-line transects. We currently are working on solutions to the boundary overlap problem when using radial and polygonal transects with two or more segments.

3 Estimation following LIS

When transect orientations are not randomly selected, estimation of τ_y is necessarily conditional on the choice of transect orientation. This arises when sampling logging slash, as in Warren and Olsen's (1964) motivating example, and one deliberately fixes the transect orientation perpendicularly to the predominant orientation of the felled needle-shaped tips. The conditional inclusion probability, denoted by $\pi_i^{\mathrm{c}}(\theta_m)$, is

$$\pi_i^{\mathrm{c}}(\theta_m) = Lw_i(\theta_m)/A, \tag{1}$$

as shown by Kaiser (1983). Based solely on the sample tallied at T_m, the Horvitz-Thompson (HT) estimator of τ_y conditionally on transect orientation is thus

$$\begin{aligned}\widehat{\tau}_{ym}^{\mathrm{c}} &= \sum_{i\in\mathcal{P}} \frac{t_{im} y_i}{\pi_i^{\mathrm{c}}(\theta_m)} \\ &= \frac{A}{L}\sum_{i\in\mathcal{P}} \frac{t_{im} y_i}{w_i(\theta_m)},\end{aligned} \tag{2}$$

where $t_{im} = 1$ if P_i is tallied by T_m, and $t_{im} = 0$, otherwise. At the conclusion of sampling, τ_y is estimated by

$$\widehat{\tau}_y^{\mathrm{c}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\tau}_{ym}^{\mathrm{c}}. \tag{3}$$

When transect orientation is randomly selected, the unconditional inclusion probability is

$$\begin{aligned}\pi_i^{\mathrm{u}} &= E_\theta\big[\pi_i^{\mathrm{c}}(\theta_m)\big] \\ &= Lc_i/A,\end{aligned} \tag{4}$$

where $c_i = E_\theta\big[w_i(\theta_m)\big]$. For any connected particle, c_i can be computed by measuring the circumference of the convex hull of P_i and dividing by the mathematical constant

π. The HT estimator of τ_y unconditionally on transect orientation is thus

$$\begin{aligned}\widehat{\tau}^{\mathrm{u}}_{ym} &= \sum_{i\in\mathcal{P}} \frac{t_{im}y_i}{\pi^{\mathrm{u}}_i} \\ &= \frac{A}{L}\sum_{i\in\mathcal{P}} \frac{t_{im}y_i}{c_i},\end{aligned} \tag{5}$$

and at the conclusion of sampling, τ_y is estimated by

$$\widehat{\tau}^{\mathrm{u}}_{y} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\tau}^{\mathrm{u}}_{ym}. \tag{6}$$

A point to be emphasized is that when transect orientations are selected at random, one can choose to estimate τ_y (λ_y) conditionally on the orientations that were chosen, or unconditionally over all possible orientations. Indeed, it seems reasonable to choose to condition, or not, based on a consideration of which estimator, $\widehat{\tau}^{\mathrm{c}}_{y}$ or $\widehat{\tau}^{\mathrm{u}}_{y}$, is more precise.

The variance of $\widehat{\tau}^{\mathrm{c}}_{y}$ is

$$V\left[\widehat{\tau}^{\mathrm{c}}_{y}\right] = \frac{1}{M^2}\sum_{m=1}^{M}\left[\sum_{i\in\mathcal{P}} y_i^2 \frac{1-\pi^{\mathrm{c}}_i(\theta_m)}{\pi^{\mathrm{c}}_i(\theta_m)} + \sum_{i\in\mathcal{P}}\sum_{\substack{i'\neq i\\ i'=1}} y_i y_{i'} \frac{\pi^{\mathrm{c}}_{ii'}(\theta_m) - \pi^{\mathrm{c}}_i(\theta_m)\pi^{\mathrm{c}}_{i'}(\theta_m)}{\pi^{\mathrm{c}}_i(\theta_m)\pi^{\mathrm{c}}_{i'}(\theta_m)}\right], \tag{7}$$

where $\pi^{\mathrm{c}}_{ii'}(\theta_m)$ is the joint conditional inclusion probability of P_i and $P_{i'}$.

Similarly, the variance of $\widehat{\tau}^{\mathrm{u}}_{y}$ is

$$V\left[\widehat{\tau}^{\mathrm{u}}_{y}\right] = \frac{1}{M}\left[\sum_{i\in\mathcal{P}} y_i^2 \frac{1-\pi^{\mathrm{u}}_i}{\pi^{\mathrm{u}}_i} + \sum_{i\in\mathcal{P}}\sum_{\substack{i'\neq i\\ i'=1}} y_i y_{i'} \frac{\pi^{\mathrm{u}}_{ii'} - \pi^{\mathrm{u}}_i\pi^{\mathrm{u}}_{i'}}{\pi^{\mathrm{u}}_i\pi^{\mathrm{u}}_{i'}}\right], \tag{8}$$

where $\pi^{\mathrm{u}}_{ii'}$ is the joint unconditional inclusion probability of P_i and $P_{i'}$.

Regrettably, these expressions provide little guidance as to whether $\widehat{\tau}^{\mathrm{c}}_{ym}$ is preferable to $\widehat{\tau}^{\mathrm{u}}_{ym}$, or not. Clearly, when dealing with particles that project circles onto $\mathcal{A}$, $\pi^{\mathrm{c}}_i(\theta_m) = \pi^{\mathrm{u}}_i$, and there is no distinction between the variances of the conditional and unconditional estimators. But for a collection of arbitrarily shaped particles haphazardly arranged on $\mathcal{A}$, it is difficult to deduce analytically whether one is apt to be more precise than the other. The combined effect on estimation variance of the joint inclusion probabilities, $\pi^{\mathrm{c}}_{ii'}(\theta_m)$ and $\pi^{\mathrm{u}}_{ii'}$, strikes me as especially difficult to discern. we are unaware that anyone has studied this matter either analytically or empirically.

The notion of conditioning inference on the transect orientations actually observed, rather than over all possible orientations that could have been observed, is not unlike poststratifying the sample and then conditioning inference on the observed post-strata sizes. Also, conditional inference is familiar to all of us in a model-based setting: having fitted a regression model, subsequent inference is conditioned on the set of predictor variable values used to estimate the parameters of the model. These familiar activities notwithstanding, conditioning after having randomly selected transect orientation is a course of action that many may find difficult to reconcile in the setting presented here. Unfamiliarity aside, there is nothing perverse about so doing.

Two questions seem relevant in this regard, however:

① Does the choice between the conditional versus unconditional estimator make much of a difference on the precision of estimation?

② Does choosing one or the other based on which has smaller estimated variance introduce bias?

In our view, the answer to the latter question is emphatically, NO. One is always free to choose a preferred estimator after sampling has been concluded and based on what has been observed in the sample, without risk of insinuating a design-based bias.

Our answer to the former question is more equivocal, because we suspect that it permits of no answer in complete generality, as it is too dependent on particle shape and pattern of distribution on the plane. Moreover, no study has been undertaken to provide empirical evidence to guide one's choice.

4 Incorporating an Auxiliary Variate into LIS

Let $q_i(\theta_m)$ be an auxiliary variate measurable on P_i whose value may depend on the angle of orientation, θ_m, of T_m. Therefore, when θ_m is a random variable, then $q_i(\theta_m)$ is random, also, whenever its value changes with θ_m.

Consider the conditional estimator

$$\widetilde{\tau}^{\mathrm{c}}_{ym} = \sum_{i \in \mathcal{P}} \frac{t_{im} q_i(\theta_m) y_i}{E\big[t_{im} q_i(\theta_m) \,|\, \theta_m\big]}. \tag{9}$$

The corresponding unconditional estimator is

$$\widetilde{\tau}^{\mathrm{u}}_{ym} = \sum_{i \in \mathcal{P}} \frac{t_{im} q_i(\theta_m) y_i}{E\big[t_{im} q_i(\theta_m)\big]}. \tag{10}$$

When $q_i(\theta_m) = 1, \forall i$, then (9) is identical to (2) and (10) is identical to (5). When this condition does not hold, then (9) and (10) are not HT estimators. For a judicious choice of the auxiliary variate, $q_i(\theta_m)$, the actual measurement of y_i may be obviated, as in the following example adapted from Kaiser (1983).

Example 1

Suppose y_i is the "coverage" onto $\mathcal{A}$ of P_i, so that one is interested in estimating the total area of $\mathcal{A}$ covered by $\mathcal{P}$. For irregularly shaped particles, y_i is difficult to measure directly. Define $q_i(\theta_m)$ to be the length of interception of P_i by the line in $\mathcal{A}$ which contains T_m as a subset ($q_i(\theta_m)$ may be longer than just the interception length of T_m in P_i). In this case, $E\big[t_{im}q_i(\theta_m)\,|\,\theta_m\big] = y_iL/A$, so that

$$\widetilde{\tau}^{\,c}_{ym} = \sum_{i\in\mathcal{P}} \frac{t_{im}q_i(\theta_m)y_i}{y_iL/A}$$

$$= \frac{A}{L}\sum_{i\in\mathcal{P}} t_{im}q_i(\theta_m) \tag{11}$$

In other words, one has replaced the tedious measurement of coverage area with the comparatively simpler measurement of its interception length.

Because $E\big[t_{im}q_i(\theta_m)\,|\,\theta_m\big] = y_iL/A$ does not involve the orientation of T_m explicitly, $E\big[t_{im}q_i(\theta_m)\big] = y_iL/A$, too. Therefore the estimator of coverage area unconditionally on orientation, $\widehat{\tau}^{\,u}_y$, is identical to $\widehat{\tau}^{\,c}_y$. We emphasize that this result, which does not hold in general, obtains here because of the particular choice of auxiliary variate to estimate coverage area. The auxiliary variate thus achieves two purposes: obviation of y_i and elimination of the dilemma between conditional versus unconditional estimation of τ_y.

References

[1] Affleck, D. L. R., Gregoire, T. G. and H. T. Valentine. (2005) Design unbiased estimation in line intersect sampling using segmented transects. *Environmental and Ecological Statistics* 12(2) 139 - 154.

[2] Bauer, H. L. (1936) Moisture relations in the chaparral of the Santa Monica Mountains, California. *Ecological Monographs*, 6(3): 409-454.

[3] Canfield, R. H. (1941) Application of the line interception method in sampling range vegetation. *Journal of Forestry*, 39: 388-394.

[4] Gregoire, T. G. and N. S. Monkevich (1994) The reflection method of line intercept sampling to eliminate boundary bias. *Environmental and Ecological Statistics*, 1: 219-226.

[5] Gregoire, T. G. and H. T. Valentine. 2003. Line intersect sampling: ell-shaped transects and multiple intersections. *Environmental and Ecological Statistics*, 10:263-279.

[6] Kaiser, L. (1983) Unbiased estimation in line-intercept sampling. *Biometrics*, 39: 965-976.

[7] Kendall, M. G. and P. A. P. Moran. (1963) *Geometrical Probability*. Hafner Publishing Company: New York

[8] MATÉRN, B. (1964) A method of estimating the total length of roads by means of a line survey. *Studia Forestalia Suecica*, 18: 68-70.

[9] ROSIWAL, A. (1898) Ueber geometrische Gesteinsaualysen. Ein einfacher Weg zur ziffremassigen Foxtstellung des Quantittsverhitnissos der Mineralbestandtheile gemengter Geneine. *Verhandlungen*, 5(6): 143-174.

[10] WARREN, W. G. and P. F. OLSEN. (1964) A line intersect technique for assessing logging waste. *Forest Science*, 10: 267-276.

Conditioning Inference on Line Orientation in Line Intersect Sampling

Timothy G. Gregoire, Yale University
David L. R. Affleck, Yale University
Harry T. Valentine, U.S. Forest Service

SamNordisk Skogsforskning, Inventory Working Group
Forest Inventory, Planning, and Modelling
Sjusjøun, Norway
6–8 September 2004