The University of Chicago
Center for Integrating Statistical and Environmental Science
www.stat.uchicago.edu/~cises

Chicago, Illinois USA

TECHNICAL REPORT NO. 46

Species Abundance Along a Curvy Margin: Correcting Sampling Biases for Coastlines, Rivers, and Other Convoluted Edges

Li Ma, Michael L. Stein, Mei Wang, A. Olaf Shelton, Catherine A. Pfister, Kenneth J. Wilder

November 26, 2008

This research is funded by
U.S. EPA-Science To Achieve Results (STAR) Program
Grant # R-82940201-0

Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201 to The University of Chicago, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Species abundance along a curvy margin: correcting sampling biases for coastlines, rivers, and other convoluted edges

Li Ma[*], Michael L. Stein[†], Mei Wang[‡]
A. Olaf Shelton[§], Catherine A. Pfister[¶], Kenneth J. Wilder[‖]

November 26, 2007

Abstract

Estimating species abundance via transects and quadrats has the advantages over other abundance estimators (such as mark-recapture) that they can be less expensive and do not require handling the animals. However, there can be diverse sources of bias in transect sampling. Here we address a bias that results when the margin from which sampling occurs has a complex geometry (*e.g.,* is convoluted) and the probability of sampling areas away from that margin is not uniform. We introduce a means of correcting that bias via a Horvitz-Thompson estimation procedure with a cubic-spline approximation of the margin. We apply these methods to the estimation of Pacific herring (*Clupea pallasi*) egg abundance along coastal southeast Alaska. We show the method corrects for an overestimation of abundance. This method has application to a number of habitats, including coastlines, rivers and forest margins.

[*]Dept. of Statistics, Stanford University, Stanford, CA 94305, USA; ma2@stanford.edu
[†]Dept. of Statistics, University of Chicago, Chicago, IL 60637, USA; stein@stat.uchicago.edu
[‡]Dept. of Statistics, University of Chicago, Chicago, IL 60637, USA; meiwang@stat.uchicago.edu
[§]Dept. of Ecology, University of Chicago, Chicago, IL 60637, USA; ole@uchicago.edu
[¶]Dept. of Ecology, University of Chicago, Chicago, IL 60637, USA; cpfister@uchicago.edu
[‖]Google Inc., Mountain View, CA 94043, USA; kjwilder@gmail.com

1 Introduction

Accurately estimating the abundance of species is a key component of ecological studies and, in many instances, quantitatively challenging. For species that are somewhat sedentary, sampling via transects and quadrats is common and estimation procedures can be found in foundational works (Seber 1973 [15], Burnham *et al.* 1980 [9], Borchers *et al.* 2002 [4], Buckland *et al.* 2001 [8]). Transect and quadrat sampling is often inexpensive and avoids handling the animals, in contrast with methods such as mark-recapture. For plants and for animals with low vagility or those that have a sedentary phase, transect sampling methods are ideal. However, as many authors have noted, there are diverse sources of bias in estimates of species abundances using transect sampling designs, including those associated with overlooking an organism when it was there [13], biases associated with aggregated distributions of species [11], and biases that arise from uneven coverage throughout the surveyed area [14].

Of the sampling biases that ecologists must contend with, perhaps the most difficult to overcome are those related to the geography inherent to the habitat to be sampled and the species distribution over that geography. One such bias is that associated with sampling along a margin that has complex geometry, such as a coastline, a river, or a forest margin. In this paper, we show that areas along such a non-linear, convoluted margin may have an unequal probability of being sampled under common transect sampling schemes, introducing biases in the estimates of species sampled in these landscapes. We introduce a way to correct this bias using Horvitz-Thompson estimation and a cubic-spline approximation of the margin (which, in the example we provide, is the coastline.) We include an artificial "buffer" zone to deal with the difficulty of measuring the exact area of the interested region. We apply these methods to a dataset on Pacific herring (*Clupea pallasi*) spawning grounds in Alaska, and show that a random sampling scheme without correction leads to an overestimate of their abundance as a consequence of the coastline geometry in that region.

2 Habitat geometry and non-uniform sampling

When estimating the abundance of species in an ecological system, designing a transect sampling scheme that achieves uniform coverage can be difficult, if not impossible. This difficulty arises in part because the spatial extent of the species of interest may not be well known *a priori*. A vital aspect of survey design then is to assign transect locations in such a way to account for this uncertainty. There are two components to assigning transect locations. First, the starting location for transects must be determined, then the direction the transect will proceed must be chosen. In habitats with clear boundaries, such as coastlines, rivers, or forest margins, the border between habitats provides a natural and repeatable location at which to initiate transects. Because species composition and abundance frequently change along elevation or distance gradients (Didham *et al.* 1998 [7], Burns 2004 [2], Power *et al.* 2004 [5]) that surround such edges, an easily determined, logical and unbiased direction in which to survey is perpendicular to the edge habitat. Transect starting positions can be randomly drawn using a linear, uniform distribution over the edge. Along each transect multiple sampling quadrats are surveyed following a predetermined set of rules. Such rules should ensure that the survey covers all parts of the region of interest.

In the Pacific herring spawning example, a quadrat is sampled every 5 meters along each transect and transects are terminated after either reaching a set distance away from the edge or if the samplers reach a predetermined cue (e.g. a given depth in coastal SCUBA survey). For this example the location of the first quadrat is effectively random because tidal fluctuations constantly change the initiation point of the transect. If the surveyed boundary follows a straight line, this sampling scheme is sufficient. However, many real boundaries are irregular and convoluted and randomized transect sampling along these boundaries results in biased, non-uniform sampling of the area.

To demonstrate the potential for sampling bias, we first define sampling probabilities.

The sampling probability of a point in the region of interest is the chance for that point to be reached from a randomly selected point on the edge. We say that a point is *reachable* from a transect origin on the edge if a quadrat lying on the transect extended from that origin location covers that point. The two step sampling procedure described earlier produces a probability distribution over the support of the interested region. However, this sampling distribution is typically not uniform within the survey area due to boundary geometry. The chance that a given point is sampled is determined by the length of the boundary from which this point can be reached. So it depends on both the number of boundary segments satisfying the reachability condition and the convexity (or concavity) of these segments. For example in Figure 1, Point 1 is much more likely to be sampled than Point 2.

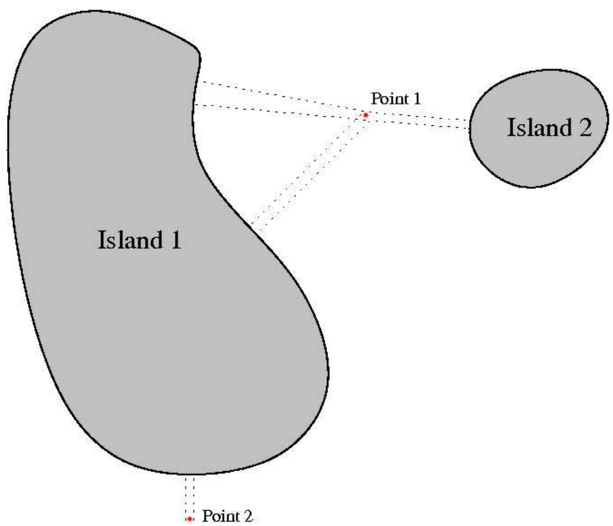

Figure 1: Two points with different sampling probabilities.

Because of the non-uniformity, simply taking the average of the densities in the sampled quadrats will produce a biased estimate of the true average density if the population's distribution is related in some way to the edge's geometry that determines the distribution of the sampling probability. For example, if the population of interest tends to occur in regions closer to the edge, and if points closer to the edge are overall more likely to be selected than points farther away, then this simple estimator will overestimate the true average density in the region. We demonstrate this bias in sampling data for Pacific herring.

In addition to finding the appropriate weights for correcting bias in the estimates of abundance, an additional challenge is to measure the exact size of the interested area, usually the area where the population is present. This problem also arises from the irregular geometric shape of the edge habitat. A simple way of estimating this area would be to take the product between the total linear length of the edge and the average length of the sampled transects. However, this method often does not produce a sensible estimate of the total area. For example in the simple idealized cases demonstrated in Figures 2(a) and 2(b), the product of the average transect length and the length of the edge underestimates the area of the censusing region in 2(a) and overestimates the area in 2(b). Only when the edge is a straight line does this product generally equal the true area of interest.

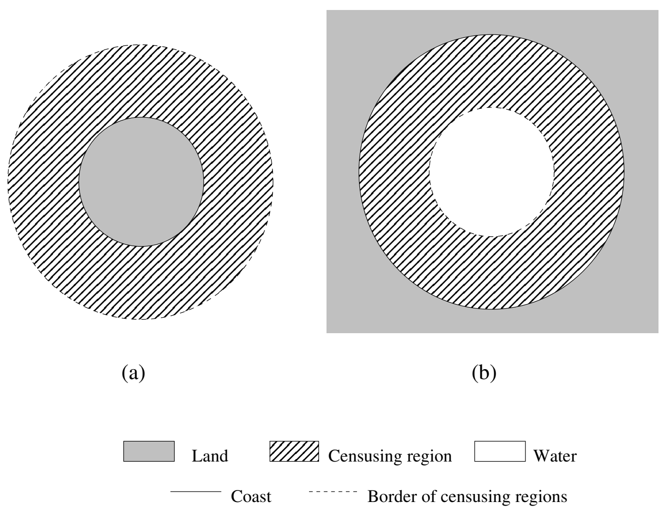

Figure 2: Two idealized censusing regions. The two censusing regions indicated in the left and right panels are of the same size, but the product of the total coast length and the average transect length underestimates the censusing area in (a) and overestimates it in (b).

3 Methodology

3.1 Unbiased estimation

The fundamental cause of the bias in the simple unweighted estimator of average population density is the non-uniformity in the sampling probability distribution arising from the edge sampling scheme. Horvitz and Thompson (1952) show how to correct the bias in such situations by weighting the sampled points inversely proportional to their sampling probabilities. The weighting technique is essentially the same as one gets from "importance sampling" (Liu 2001 [10]) but applied to probability masses rather than densities.

Before presenting the weighted estimator, we first introduce some basic notation. Write $Q(t,q)$ for the qth quadrat along the tth transect and Q_t for the total number of quadrats placed along the tth sampled transect. Let T be the total number of transects sampled, d the distance between two adjacent quadrats on each transect, π_{tq} the sampling probability of *the central point* of $Q(t,q)$, and finally X_{tq} the measured population density in $Q(t,q)$.

Under the key assumption that sampling probabilities vary little for points within the same quadrat, which is typically true when the quadrats are small compared to the scale of the transects, the Horvitz-Thompson (H-T) estimator for average population density is

$$\widehat{\mu} = \frac{\sum_{t=1}^{T} \sum_{q=1}^{Q_t} (1/\pi_{tq}) X_{tq}}{\sum_{t=1}^{T} \sum_{q=1}^{Q_t} 1/\pi_{tq}}, \tag{1}$$

which can be calculated by computing π_{tq} for all sampled quadrats. (See detailed derivation in Appendix A1.)

The variance of the H-T estimator is commonly estimated by computing second order inclusion probabilities [1]. Here we can avoid computing second order inclusion probabilities by treating the transects, rather than the quadrats, as our sampling units. The transect-quadrat sampling scheme allows us to do so because the sampled transects are selected *independently* from the space of all transects in the first step of the two-step procedure. (There is a one-to-one correspondence between transects and transect origins.) This random sampling over transects provides protection from the spatial correlation *across* transects. At the same time, treating the transects as the basic units saves us from dealing with the covariance structure among the quadrats *within* each transect. Therefore, under this sampling scheme we can bypass the entire covariance structure of the quadrats in estimating the variance of the H-T estimator.

To implement this transect-as-unit method, let Y_t be the weighted sum of population densities along the tth randomly selected transect, and Z_t the sum of the weights of all the quadrats sampled along that transect. That is,

$$Y_t = \sum_{q=1}^{Q_t} (1/\pi_{tq}) X_{tq} \quad \text{and} \quad Z_t = \sum_{q=1}^{Q_t} 1/\pi_{tq}.$$

For each t, Y_t and Z_t are two jointly distributed random variables over the space of all the transects that could be drawn, and (Y_1, Z_1), (Y_2, Z_2), $\ldots$, (Y_T, Z_T) are independent identically distributed bivariate random vectors. More specifically, let $\mu_Y = E(Y_t)$, $\mu_Z = E(Z_t)$, $\sigma_Y^2 = \text{Var}(Y_t)$, $\sigma_Z^2 = \text{Var}(Z_t)$ and $\sigma_{YZ} = \text{Cov}(Y_t, Z_t)$. Then

$$\begin{pmatrix} Y_t \\ Z_t \end{pmatrix} \sim \left(\begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{pmatrix} \right),$$

for all $t = 1, 2, \ldots, T$.

With these new notations, the H-T estimator of the average population density can be rewritten in terms of Y and Z,

$$\widehat{\mu} = \frac{\sum_{t=1}^{T} Y_t}{\sum_{t=1}^{T} Z_t}. \tag{2}$$

Its variance can be estimated by (see [3], p.155)

$$\widehat{\text{Var}}(\widehat{\mu}) = \frac{1}{T}\Big(\frac{\overline{Y.}}{\overline{Z.}}\Big)^2\Big(\frac{s_Y^2}{\overline{Y.}^2} + \frac{s_Z^2}{\overline{Z.}^2} - \frac{2s_{YZ}}{\overline{Y.}\,\overline{Z.}}\Big) \tag{3}$$

where

$$\overline{Y.} = \frac{1}{T}\sum_{t=1}^{T} Y_t, \quad \overline{Z.} = \frac{1}{T}\sum_{t=1}^{T} Z_t, \quad s_Y^2 = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \overline{Y.})^2, \quad s_Z^2 = \frac{1}{T-1}\sum_{t=1}^{T}(Z_t - \overline{Z.})^2,$$

and

$$s_{YZ} = \frac{1}{T-1}\sum_{t=1}^{T}(Y_t - \overline{Y.})(Z_t - \overline{Z.}).$$

Also note that the unweighted estimator $\widehat{\mu}^0$ coincides with the H-T estimator $\widehat{\mu}$ when all of the π_{tq}'s are equal. Hence we can use (3) to estimate the variance of $\widehat{\mu}^0$ as well.

3.2 Computing sampling probabilities

The key to compute the H-T estimator and its standard error is to find the appropriate weight, π_{tq}, or the inverse sampling probability, for the center of each sampled quadrat. Since the sampling probability for each point is proportional to the length of the edge segments from which that point can be reached, our task is to compute this length for each quadrat center.

The first component of computing sampling probabilities is to algorithmically define the edge where transects begin. Defining this edge is a trivial exercise if a high resolution digital map containing the edge as a distinct layer is available. In our example with herring below, we only had access to a digitized chart (a common form of spatial data) from which we extracted a one pixel wide coastal profile. We fit uniform cubic B-splines to this profile, which gives an analytical representation of the coast that will later be utilized to compute sampling probabilities. Of course by fitting cubic splines, we assume that the coastline satisfies the corresponding differentiability conditions. Note that a smoothness assumption is implicit in the *actual* sampling scheme, because SCUBA censusing attempted to be in a direction perpendicular from the coast, which is possible only if the coastline is assumed to be differentiable. We provide details for extracting and spline-smoothing the NOAA nautical chart coastline in Appendix A2. These techniques could be applied to other digitized media such as scaled aerial photographs or satellite images.

Once the smoothed chart was produced, we computed the sampling probabilities for the centers of sampled quadrats. For this purpose we divided the B-splines approximating the coast profile into tiny segments whose lengths are 0.002 pixel width. The choice of this length is arbitrary up to the extent that they must be small enough so that the coastal segments can be treated as points for numerical purposes. Within each of these tiny segments the edge is approximately a straight line, so the tangency of the coast in each of these segments is well defined and can be attained from the analytic form of the corresponding splines. This then determines the direction of the transect extending from that segment. For computational purposes, we can translate the reachability of a given quadrat center from a coastal segment into three criteria: (1) the distance from the point to the transect is less than half of the side length of the quadrats, (2) the line segment connecting the point and the coast segment must be contained in the survey region, and (3) there is no physical barrier (*e.g.,* a rock or deep water) between that point and that coast segment that would prevent the surveyers from sampling the entire transect. Summing up the length of all the tiny segments satisfying the above three criteria gives the sampling probability of that quadrat center (up to a constant).

3.3 Measuring the survey area

Now that an unbiased estimator for the average population density is available from H-T estimation, the remaining task is to measure the size of the survey area. Because habitat edges are not straight, simply estimating the average transect length and taking its product with the total linear edge length will not give a correct estimate of the survey area. However, we can solve this problem by noticing that it is unnecessary to measure the exact area where a population is actually present. If we could find a region that covers all the survey area, and if we pick this extended region (ER) carefully enough, then we would be able to easily measure the area of the ER, and use the method introduced in the previous subsection to estimate the average density in this extended region. The product of the average density in the ER and the size of the ER will be a sensible estimate of the total population size.

4 Application to Data

4.1 Study System

Along the west coast of North America, Pacific herring (*Clupea pallasi*) spawn annually in shallow coastal areas vegetated with marine algae (e.g. the kelps, Order Laminariales) and seagrasses (e.g. *Zostera (ital.) spp.*). Surveys of herring eggs deposited in spawning grounds provide an economical method to assess abundance of this ecologically important and commercially exploited species. Total egg abundance is a critical component for determining population size and setting fisheries catch limits. Spawn deposition occurs in April or May in southeastern Alaska. Egg densities can be extraordinary, exceeding 10^8 $eggs/m^2$ in places. Herring egg abundance estimates are derived from SCUBA surveys conducted at herring spawning grounds by the Alaska Department of Fish and Game (ADFG).

Annually since 1980, ADFG divers have estimated herring egg abundances at herring spawning sites by initiating a transect from a randomly selected point along the coast known to have received herring spawn and conducting a transect perpendicular to the coastline. Within a given transect the abundance of herring eggs attached to the substrate within a $0.1\,m^2$ quadrat is visually estimated every 5 meters. The number of quadrats sampled per transect depends on local bathymetry.

In this section we apply our method to the ADFG herring spawn survey data collected in Sitka Sound in southeastern Alaska in April of 2004. The map available for this region is a NOAA nautical chart at an approximate scale of 1:25000. The dataset includes the egg counts for 637 sampled quadrats along 42 transects randomly selected from 77 miles of coastline that received herring spawn in 2004. The numbers of quadrats included in each of the 42 transects range from 3 to 65 and average 15.

4.2 Extended region designation

Because herring only spawn in shallow waters close to the coast, one simple example of an extended (spawning) region (ESR) would be the sea region within a certain distance, D, to the coast. Under previous notations, this is equivalent to imposing the condition that $Q_t \leq D/d$. See an illustration of such an ESR in Figure 3.

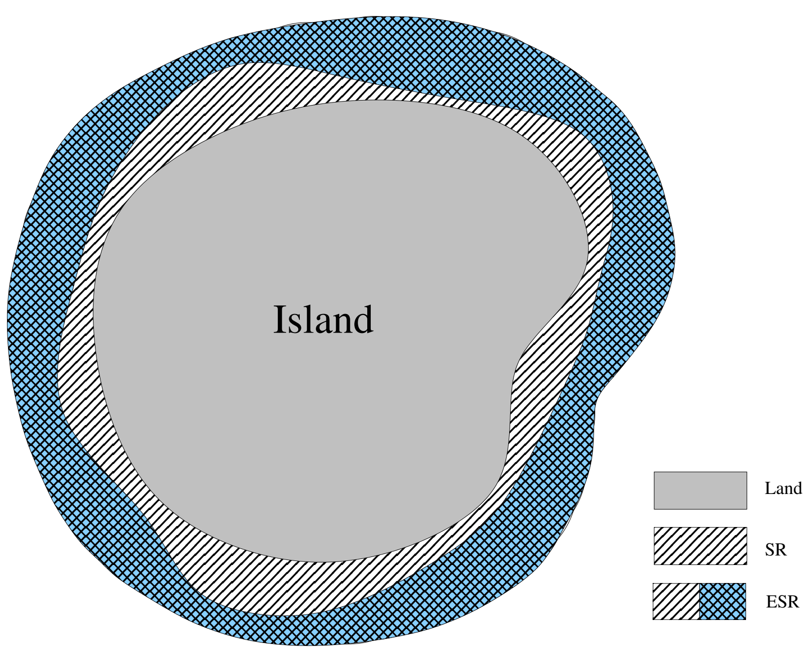

Figure 3: True spawning region (SR) and the extended spawning region (ESR).

For the ESR, most transects extending from the possible origin points on the coast include exactly D/d quadrats except those that are physically blocked by land. Also, because the divers stop at a water depth for which there are no eggs at greater depths, X_{tq} is hereafter defined to be 0 if it was not directly observed by the divers. That is,

$$X_{tq} = 0 \qquad \text{for } t = 1, 2, \ldots, T \text{ and } Q_t < q \leq D/d.$$

On the other hand, the area of the ESR, A_E, can be easily measured, essentially without error, on a digitized map by counting the pixels. Finally, we can obtain an estimator for the total number of eggs,

$$\widehat{N} = \widehat{\mu} A_E = \frac{\sum_{t=1}^{T} Y_t}{\sum_{t=1}^{T} Z_t} \cdot A_E, \tag{4}$$

and

$$\widehat{\mathrm{Var}}(\widehat{N}) = \frac{1}{T}\Big(\frac{A_E\,\overline{Y.}}{\overline{Z.}}\Big)^2\Big(\frac{s_Y^2}{\overline{Y.}^2} + \frac{s_Z^2}{\overline{Z.}^2} - \frac{2s_{YZ}}{\overline{Y.}\,\overline{Z.}}\Big), \tag{5}$$

where the Y's and Z's are now defined on the ESR.

5 Results and Discussion

Because herring typically do not spawn more than 400 meters away from the coast, we choose $D = 400$ meters as the upper bound of transect length in the definition of the ESR. A histogram of the weights for the sampled quadrats, *i.e.,* the inverse sampling probabilities for their centers, is given in Figure 4. The values are rescaled so that 1 represents the weight for a quadrat if the coast were a straight line and the sampling on the spawning region were uniform. The weights range from 0.074 to 11.5. Most of them are smaller than 1, which is consistent with the fact that points with higher sampling probabilities, and thus lower weights, are generally over-sampled. Figure 5 gives a color map of the sampling probabilities. It shows that most quadrat centers close to the coast have significantly higher chance of being sampled (indicated as blue) than if the coastline were a straight line.

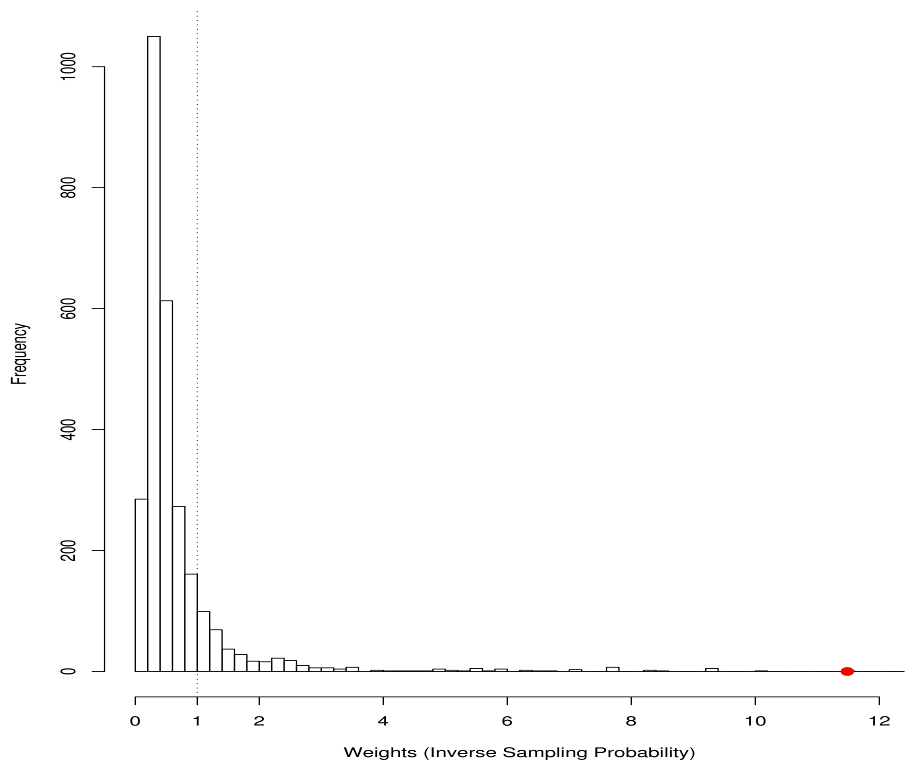

Figure 4: Histogram of the inverse sampling probabilities of quadrats in the Pacific herring surveys of coastal southeast Alaska. The red dot represents the point with the highest weight. Values to the left of the dashed line indicate those that were oversampled.

Table 1 presents the estimates of the egg density and the total number of eggs in the survey region using four different methods: (1) unweighted estimation on the extended spawning region, (2) H-T estimation on the extended spawning region, (3) unweighted estimation on the (unextended) spawning region, which was used by the ADFG, and (4) H-T estimation on the (unextended) spawning region. Note that strictly speaking there is no way to attain the weights for the original spawning region, because the exact boundary of this region is

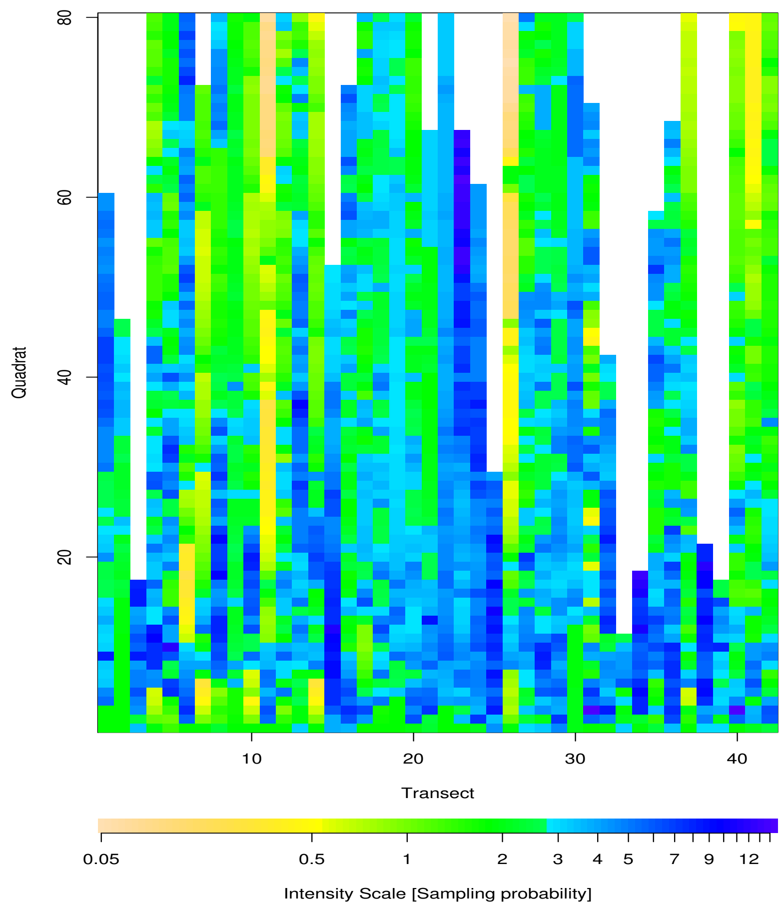

Figure 5: Map of the sampling probabilities for herring eggs in coastal southeast Alaska. Yellow indicates a low probability of being sampled while blue is high. The data have been rescaled so that "1" indicates the chance of being sampled if the coast were simply a straight line.

unknown and so the lengths of the possible transects are also uncertain. Here we use the weights from the ESR as an approximation for the weights from SR. The latter two methods give much higher estimates for the total egg count mainly because the estimate for the spawning area, which is attained by simply taking the product of the total coastal length and the average transect length, is probably larger than the true spawning area. For both the ESR and the SR, H-T estimation produces smaller estimates for the egg count than those given by the unweighted estimator. Also, the standard errors for the H-T estimators are almost the same as their counterparts for the unweighted estimators. So correcting the bias of the unweighted estimator using weights does not lead to increased variability in this case.

	On the ESR		On the SR	
	Unweighted	H-T	Unweighted	H-T***
Egg density ($1000/m^2$)	131 (22)	95 (23)	570 (73)	530 (76)
Coast length (km)	146	146	146	146
Spawning area (km^2)	-	-	11.1*	11.1*
Extended spawning area (km^2)	40.8	40.8	-	-
Total egg count (10^{10})	534 (90)	390 (92)	633 (81)**	588(87)**

Table 1: Estimated egg density and total egg count by method. Note: (∗) The true spawning area is unknown, and probably smaller than $11.1\,km^2$, which is the product of the total coastal length and the average transect length. (∗∗) The standard error for the estimated total egg count under the last method is computed assuming that contribution to the variance of the estimated total egg count from the variance of the estimated spawning area is dominated by the uncertainty in the estimated egg density. (∗∗∗) Because the actual spawning region is unknown, in theory there is no way to compute the appropriate weights for the H-T estimator on the SR. Here we use the weights computed from the ESR as an approximation.

The percentagewise difference between the H-T estimator for the egg count and the unweighted one is nontrivial for both the ESR and SR. However, the difference for the ESR, about 27%, is quite a bit larger than that for the SR, and is a finding we would expect because points further from the coast have smaller sampling probabilities and therefore larger weights in the H-T estimator (Figure 5). Therefore, the H-T estimator tends to put larger weights on the points in the ESR that fall out of the true spawning region as those points tend to be farther from the coast. Thus more weight is put on points with zero eggs, drawing down the H-T estimator relative to the unweighted estimator.

Because the upper bound on the length of each transect we impose to define the ESR in our example is quite arbitrary, it is interesting to see how sensitive the estimates are for different choices of the ESR. For this reason, we compute the estimates for the total egg count for several different feasible ESRs. We construct a sequence of different ESRs by changing the maximum transect length from 325 meters (i.e., 65 quadrats), which is the observed length of the longest transect, up to 525 meters (i.e., 105 quadrats). As Table 2 shows, both the H-T estimate for the total egg count and its standard error are quite stable over this rather generous range of ESRs.

Q_M	65	70	75	80	85	90	95	100	105
Estimated total	424	417	401	390	387	393	392	392	395
egg count	(99)	(95)	(92)	(92)	(91)	(92)	(93)	(94)	(95)

Table 2: Estimated total egg count ($/10^{10}$) using H-T estimation and its standard error for different ESRs. Q_M is the maximum number of quadrats in a transect, which defines our ESRs.

6 Concluding Remarks

In habitats with complicated geometry, transect sampling schemes that randomly select transect origins along the edge of a habitat can result in an uneven distribution of sampling probabilities over the surveyed area. This uneven distribution leads to bias in the unweighted estimator of the average population density. We introduce a method that incorporates the geometry of habitat edges to compute the appropriate weights for the sampled quadrats, thus correcting the sampling bias. When applied to the ADFG herring data, this weighting procedure causes no extra variability in the estimation while correcting significant bias. Cluster sampling coupled with this Horvitz-Thompson estimation method provides an effective, economic and generic way of sampling an area. Such a sampling design is particularly valuable when direct (spatial) uniform sampling is too costly or the survey area of interest is not easily defined, as a result of a lack of high resolution maps, such as the lack of detailed bathymetric data for the herring dataset.

An important issue in implementing the techniques introduced in the paper is the choice of the buffer zone, or the ER. The general guiding principle is to pick a region that covers the true interested region and exceeds the latter *as little as possible*. (Ideally, the ER will equal the true region of interest.) Of course another principle is that the area of the ER should be easy to measure or compute. A good choice of the ER should balance these two contrasting criteria. In this light, the ER we used for the herring data was based on a fixed maximum transect length, and was thus rough. It meets the second criteria very well while almost completely ignoring the first. A better ER can be obtained by techniques such as photographing the interested region to determine at least approximately the appropriate maximum transect lengths for different parts of the boundary. This should be done before the sampling procedure takes place. With such information at hand, the divers (or other surveyers) should sample each transect up to the designated maximum length of that transect if possible, rather than determining subjectively where they should stop. While we apply the H-T estimator to a coastal situation, it should be emphasized that the procedure outlined above is generic and applicable to a wide variety of ecological settings—where detailed knowledge of a survey area is limited and habitat is irregularly shaped.

Acknowledgments

We are thankful to ADFG for collecting and S. Dressel of the ADFG for providing the Pacific herring spawning data. M. Pritchett provided GIS maps. Funding was provided from an EPA-STAR Cooperative Agreement #R-82940201-0 to the University of Chicago. Although the research described in this article has been funded in part by the U.S.E.P.A., it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. J. Forrester made helpful comments on the manuscript.

Appendix A1. Deriving the H-T estimator

We use a pair of coordinates (s, r) to represent a point in the sea, where s indexes the position of the corresponding transect origin on the coast and r the perpendicular distance of the point to that transect origin. Because a point in the sea could have more than one corresponding origin points on the coast (see Figure 1), different pairs of coordinates could refer to the same spot. Next, we let $p(s, r)$ be the sampling probability for Point (s, r), and $X(s, r)$ the population density at Point (s, r). Now if we let A_0 be the area of each quadrat, then under the key assumption that sampling probabilities vary little within each quadrat, the Horvitz-Thompson estimator can be computed as

$$\begin{aligned}\widehat{\mu} &= \frac{\sum_t \sum_q \int_{Q(t,q)} X(r,s) \cdot 1/p(r,s)\, dA}{\sum_t \sum_q \int_{Q(t,q)} 1/p(r,s)\, dA} \\ &\approx \frac{\sum_t \sum_q 1/\pi_{tq} \int_{Q(t,q)} X(r,s)\, dA}{\sum_t \sum_q 1/\pi_{tq} \int_{Q(t,q)} dA} \\ &= \frac{\sum_t \sum_q (1/\pi_{tq}) X_{tq} A_0}{\sum_t \sum_q (1/\pi_{tq}) A_0} = \frac{\sum_{t=1}^{T} \sum_{q=1}^{Q_t} (1/\pi_{tq}) X_{tq}}{\sum_{t=1}^{T} \sum_{q=1}^{Q_t} 1/\pi_{tq}}.\end{aligned}$$

Appendix A2. Extracting and smoothing the coastline

Coastline extraction and thinning

Extracting the coastline would be a trivial matter if the map had a separate layer identifying the coast or it used a unique color to indicate coastal points. In our case, however, the map was a single digitized image of Southeast Alaska. Both the coastline and all the legends and comments are represented by black pixels in the image; land, shallow water, and deep water, on the other hand, are encoded as camel, blue, and white pixels respectively. Also, in the map the coastline had a width of two to three pixels.

To extract the coastline we distinguished the coastal pixels from all the other black pixels, and that required some definition of a "coastal pixel". The appropriate definition depends

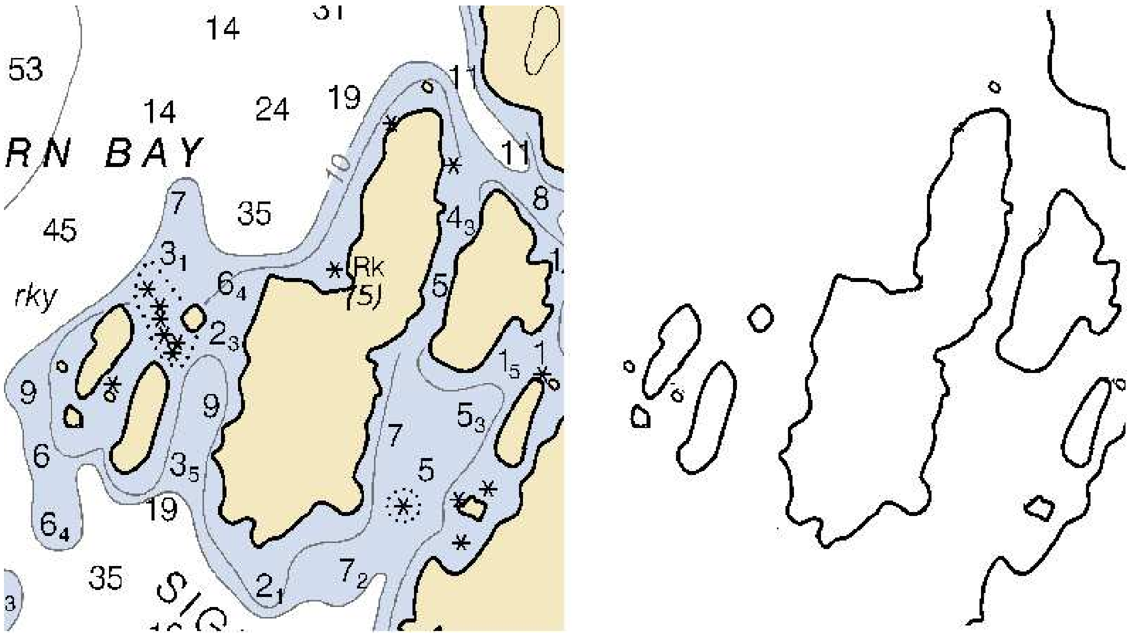

Figure 6: Coastline extraction of a portion of the map.

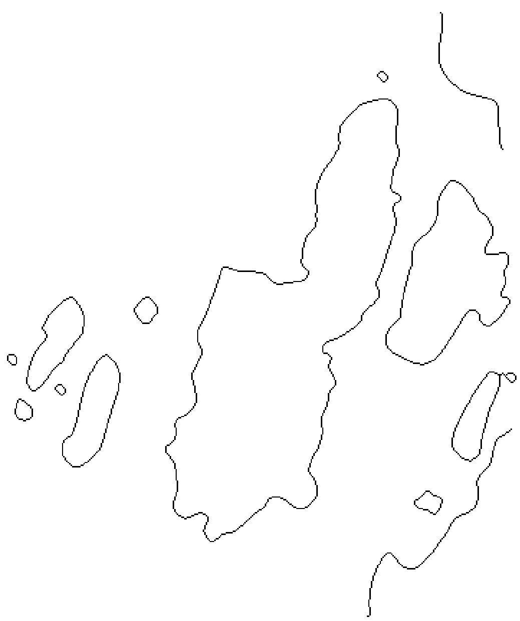

Figure 7: A thinned version of the coastline in Figure 6.

on the specific properties of each map. For the ADFG herring project, the criteria we chose was that if the 5 pixel by 5 pixel square centered at a black pixel included both land and water, then that black pixel in the middle was considered a coastal pixel. Picking out coastal pixels based on this rule produces a two-to-three-pixel wide coastline. Figure 6 illustrates the effect of this algorithm on a typical portion of the map. Overall, this algorithm does a satisfactory job in identifying the coastal points.

After extracting the coastal pixels, the next step was to thin the coastline to one-pixel wide. There are standard 2D image processing algorithms for this particular purpose that preserve the Euler number of the coastline image [12]. Such algorithms are implemented in standard image processing packages such as the image processing toolbox in MATLAB, which we used. The result of the thinning procedure when applied to the coast segment in Figure 6 is presented in Figure 7.

Coastline smoothing

After obtaining a one-pixel wide coastline, we smoothed it by fitting uniform cubic B-splines with the coastal pixels as control points [6]. A uniform cubic B-spline is twice continuously differentiable and its x and y coordinates can respectively be parametrized as piecewise cubic polynomials. Each piece between two adjacent pixels is determined by the four adjacent pixels around that segment, and the x and y coordinates of each piece are parametrized as cubic functions of a "speed" parameter $t \in [0, 1]$. For a more detailed discussion of these and other properties of B-splines, see [6]. The fitted B-splines give an explicit mathematical formula for the coast, which facilitates the computation of both the length of a coastal segment and the tangency at any coastal point.

References

[1] P. Ardilly and Y. Tillé. *Sampling Methods.* Springer, New York, 2005.

[2] T. P. Burns. Hard-coral distribution and cold-water disturbances in south florida: variation with depth and location. *Coral Reefs*, 4:117–124, 2004.

[3] W. G. Cochran. *Sampling Techniques.* Wiley, Hoboken, 3rd edition, 1977.

[4] D. L. Borchers *et al. Estimating animal abundance.* Springer, London, 2002.

[5] M. E. Power *et al.* River-to-watershed subsidies in an old-growth conifer forest. *Food webs at the Landscape Level*, pages 217–240, 2004.

[6] R. H. Bartels *et al. An Introduction to Splines for use in Computer Graphics and Geometric Modeling.* Morgan Kaufmann, 1st edition, 1987.

[7] R. K. Didham *et al.* Beetle species responses to tropical forest fragmentation. *Ecological Monographs*, 68:295–323, 1998.

[8] S. T. Buckland *et al. Introduction to distance sampling.* Oxford University Press, Oxford, 2001.

[9] J. L. Laake K. P. Burnham, D. R. Anderson. Estimation of density from line transect sampling of biological populations. *Wildlife monographs*, 72:202, 1980.

[10] J. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, New York, 2001.

[11] J. H. Pollard and S.T. Buckland. Adaptive distance sampling surveys. *Advanced distance sampling*, pages 229–259, 2004.

[12] W. K. Pratt. *Digital Image Processing.* Wiley, Hoboken, 2nd edition, 1991.

[13] T. J. Quinn and V. F. Gallucci. Parametric models for line-transect estimators of abundance. *Ecology*, 61:293–302, 1980.

[14] S. T. Buckland S. Strindberg and L. Thomas. Design of distance sampling survey and geographic information systems. *Advanced distance sampling*, pages 190–228, 2004.

[15] G. A. F. Seber. *The estimation of animal abundance and related parameters.* Griffin, London, 1973.