Mixture models for distance sampling detection functions

David L. Miller[1,*], Len Thomas[1]
1 School of Mathematics and Statistics, and Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews KY16 9LZ, Scotland
∗ E-mail: dave@ninepointeightone.net

Abstract

We present a new class of models for the detection function in distance sampling surveys of wildlife populations, based on finite mixtures of simple parametric key functions such as the half-normal. The models share many of the features of the widely-used "key function plus series adjustment" (K+A) formulation: they are flexible, produce plausible shapes with a small number of parameters, allow incorporation of covariates in addition to distance and can be fitted using maximum likelihood. One important advantage over the K+A approach is that the mixtures are automatically monotonic non-increasing and non-negative, so constrained optimization is not required to ensure distance sampling assumptions are honoured. We compare the mixture formulation to the K+A approach using simulations to evaluate its applicability in a wide set of challenging situations. We also re-analyze four previously problematic real-world case studies. We find mixtures outperform K+A methods in many cases, particularly spiked line transect data (i.e., where detectability drops rapidly at small distances) and larger sample sizes. We recommend that current standard model selection methods for distance sampling detection functions are extended to include mixture models in the candidate set.

Introduction

Distance sampling [1, 2] is a suite of methods for estimating the size or density of biological populations. There are two main variants: line and point transects. In both, an observer visits a randomly-located set of transect lines or points and records the distance, y, from the transect to each object of interest (i.e., animals or plants of the target species) that is detected within some truncation distance w. Not all objects within w are assumed to be detected; instead the observed distances are used to estimate the parameter vector, θ, of a detection function model, $g(y;\theta)$, which describes how the probability of detection declines with increasing distance. An assumption of the conventional method is that $g(0;\theta) = 1$. Given an estimate of θ, it is straightforward to estimate population size or density (see below).

A key part of distance sampling, therefore, is specification of the detection function model. Buckland et al. (Chapter 2) [1] provide a set of criteria for judging the utility of candidate model classes. Detection function models should be:

1. flexible, so that they can take a wide variety of shapes;
2. efficient, in the sense that many plausible shapes can be represented using few parameters;
3. flat at zero distance (i.e., $g'(0;\theta) = 0$), indicating that objects in the immediate vicinity of the observer are equally detectable; and,
4. monotonic non-increasing with increasing distance (i.e., $g'(y;\theta) \leq 0$ for $0 < y \leq w$), as it is typically unrealistic for objects to become more detectable with increasing distance.

The semiparametric "key function plus series adjustment" (K+A) modelling approach developed by Buckland [3] has become by far the most popular in practice, partly due to its inclusion in the industry-standard distance sampling analysis software Distance [4] and the R package `mrds`. However, as we demonstrate below, the approach has some drawbacks; in particular, although it meets criteria 1-3, it

does not necessarily meet the 4th. Our purpose in this article is to propose an alternative class of models, based on mixtures, that meets all 4 criteria and to evaluate its utility.

The approach of Buckland [3] was extended by Marques and Buckland [5] to allow covariates in addition to distance to be included in the detection function, and, for maximum generality, it is this K+A formulation that we describe here. The detection function is thus denoted $g(y, \mathbf{z}; \theta)$ where $\mathbf{z}$ is an observation-specific vector of covariates; the formulation of Buckland [3] is simply a special case of this model where there are no additional covariates.

In Marques and Buckland [5], the detection function is modelled as a parametric key function k and series expansion s of even functions (known as *adjustment terms*) with some parameters θ. g is then written as:

$$g(y, \mathbf{z}; \theta) = \frac{k(y, \mathbf{z}; \theta)\{1 + s(y, \mathbf{z}; \theta)\}}{k(0, \mathbf{z}; \theta)\{1 + s(0, \mathbf{z}; \theta)\}},$$

where k may be a half-normal, hazard-rate or uniform function and s may be zero (i.e., there are no adjustment terms), cosine, simple even polynomial or Hermite polynomial series (though note a uniform detection function may not include covariates). The denominator ensures that detection function evaluates to 1 at zero distance (i.e., $g(0, \mathbf{z}; \theta) = 1$). Model parameters are estimated using maximum likelihood. The recommended strategy for most situations is to choose a small set of key function and adjustment combinations, and for each combination to choose the number of adjustment terms using forward selection, i.e., start with no adjustment terms and fit an increasing number of terms, stopping when the Akaike Information Criterion (AIC) fails to decrease [4]. The combination with the lowest AIC is then selected as the best model. This strategy works well in practice in many cases: the key functions cover a range of realistic shapes for the detection function, so that often zero or one adjustments are sufficient to provide a good fit to the data, resulting in flexible and yet efficient estimation.

The resulting detection functions are capable of being flat at zero distance and the key functions are non-increasing. However, adding adjustment terms can result in non-monotonic functions. Further, when both covariates and adjustments are included in the model the range of the resulting detection function may not be [0,1]. When there are no additional covariates, one solution is to use constrained maximization, e.g. taking M equally spaced distances $y_1 = 0, \ldots, y_M = w$ and ensuring that $g(y_i; \hat{\theta}) \geq g(y_{i+1}; \hat{\theta})$ and that $g(y_{i+1}; \hat{\theta}) \geq 0$ for $i = 1, \ldots, M - 1$. In Distance this constraint is implemented using the NLPQL routine [6] and in the R package `mrds`, the SOLNP algorithm [7] is used.

A constrained optimisation solution presents a number of problems. First, constrained maximization is a more complex optimization problem than unconstrained maximization; this means that in practice optimization algorithms may fail to find the constrained maximum. Second, constrained maximum likelihood estimates do not have the same appealing properties as their unconstrained relatives – for example the usual estimator of the standard error of the parameters (square root of the inverse of the information matrix) can be biased. Third, constraints can only be applied at a finite number of points ($M = 10$ is used in Distance and $M = 20$ in `mrds` by default), which can lead to the constraint points missing non-monotonic parts of the function. Though increasing the number of points is an option, this incurs additional computational cost. An example of constrained maximisation failing is shown in the left panel of Figure 1. Finally, it is not clear how to implement the constraints in the case where there are additional covariates, particularly continuous covariates. One computationally expensive option would be to apply the constraints at every observed covariate combination (at present both Distance and `mrds` use unconstrained optimization when additional covariates are in the model). The central and right panels of Figure 1 from Pike et al. [8] show an example of covariate models fitted using unconstrained optimisation: a strongly non-monotonic function has been fitted for some covariate values. Detection probability estimates outside the range $[0, 1]$ are sometimes encountered during maximization when models include covariates. Given the above issues, it seems appealing to use a formulation that guarantees monotonicity from the outset.

Mixture models have been applied in the capture-recapture literature [9–12]. The main utility of

mixture models in capture-recapture is in better accounting for between-individual heterogeneity, which can cause severe bias if unmodelled [13]. Unmodelled heterogeneity is not generally considered an issue in distance sampling, provided that detection at zero distance is certain, heterogeneity is not extreme and a flexible detection function model is used [2 Section 11.12]. Mixture models offer the potential for flexible modelling since the individual parts of the mixture model (the *mixture components*) can be combined to obtain flexible detection functions, and provided each component is monotonic non-increasing, the resulting combination will also be monotonic non-increasing. In addition, mixture models are potentially well suited to deal with highly heterogeneous detection probabilities, where some part of the population is only observable at close distances while others are readily detected almost regardless of distance (for example bird species where males are more vocal than females). Such a situation results in a "spiked" detection function with a long flat tail – Figure 1 shows relatively mild examples. In a mixture model, different parts of the sample could be represented by different components, providing a good fit to spiked data and an appealing adequate conceptual explanation for the underlying data.

Here we introduce a new class of distance sampling detection function models, based on mixtures of simple parametric key functions. In the next section, we describe the models. We then illustrate their use and explore their performance by applying them to simulated data, and to real data from a number of studies. We also compare results with those obtained from the current standard K+A approach, and by using a combined approach where both the mixtures and K+A models are applied and a final model selected using AIC. An R [14] package, `mmds` (for Mixture Model Distance Sampling), implementing the methods is available from the Comprehensive R Archive Network (CRAN).

Methods

Finite mixture model detection functions: Formulation

Denoting the detection function as g, we consider a sum of J mixture components g_j, scaled by some mixture proportions ϕ_j:

$$g(y, \mathbf{z}; \theta, \phi) = \sum_{j=1}^{J} \phi_j g_j(y, \mathbf{z}; \theta_j),$$

where $\sum_{j=1}^{J} \phi_j = 1$. The distance is denoted y, the θ_js are vectors of parameters for function g_j, θ is a vector of all of the θ_js, ϕ is a J-vector of all of the ϕ_js, and $\mathbf{z}$ is a K-vector of the associated covariates.

Although other monotonic functions such as hazard-rate could be chosen, and the g_js need not all have the same form, here we let the g_js be half-normal functions:

$$g(y, \mathbf{z}; \theta, \phi) = \sum_{j=1}^{J} \phi_j \exp\Big(- \frac{y^2}{2\sigma_j(\mathbf{z})^2}\Big).$$

We assume that each mixture component has a different scale but that the covariates affect the scale parameters in the same way (though other, more complex, models may be possible).

Covariates are included as in Marques and Buckland [5], by decomposing the scale parameter σ (see also Marques et al. [15]). Using i to subscript each observation, our formulation for the scale parameter σ_{ij}, is

$$\sigma_{ij} = \exp(\beta_{0j} + \sum_{k=1}^{K} \beta_k z_{ik}),$$

where z_{ik} is the k^{th} covariate for the i^{th} observation. In this case θ will contain the β_{0j}s and β_ks.

We can write the pdf of the observed distances conditional on the observed covariates as [2]:

$$f(y|\mathbf{z};\theta,\phi) = \frac{\pi(y)g(y,\mathbf{z};\theta,\phi)}{\int_0^w \pi(t)g(t,\mathbf{z};\theta,\phi)\mathrm{d}t}.$$

where $\pi(y)$ is the pdf of object distances (observed and unobserved). The likelihood can then be formed by taking product of these pdfs over the n observations. The specific form of the likelihood differs between line and point transects, because sampler geometry means that the form of $\pi(y)$ is different for lines and points. For line transects, with random line placement, we expect an equal number of objects at all distances from the line, and hence $\pi(y) = 1/w$ (where w is again the truncation distance). The likelihood is then given by:

$$\begin{aligned}\mathcal{L}(\theta,\phi;\mathbf{y}|\mathbf{z}_1,\ldots,\mathbf{z}_n) &= \prod_{i=1}^{n} f(y_i|\mathbf{z}_i;\theta,\phi)\\ &= \prod_{i=1}^{n} \frac{g(y_i,\mathbf{z}_i;\theta,\phi)}{\mu_i(\mathbf{z}_i)}\\ &= \prod_{i=1}^{n} \frac{\sum_{j=1}^{J}\phi_j g_j(y_i,\mathbf{z}_i;\theta_j)}{\mu_i(\mathbf{z}_i)}\end{aligned}$$

where $\mu_i(\mathbf{z}_i)$ (called the *effective strip width*) is given by:

$$\mu_i(\mathbf{z}_i) = \sum_{j=1}^{J}\phi_j \int_0^w g_j(y,\mathbf{z}_i;\theta_j)\mathrm{d}y. \tag{1}$$

For point transects, with random point placement, the expected number of objects increases with increasing distance from the point, and hence $\pi(r) = 2r/w^2$, giving

$$\begin{aligned}\mathcal{L}(\theta,\phi;\mathbf{y}|\mathbf{z}_1,\ldots,\mathbf{z}_n) &= \prod_{i=1}^{n} f(y_i|\mathbf{z}_i;\theta,\phi)\\ &= \prod_{i=1}^{n} \frac{2\pi y_i g(y_i,\mathbf{z}_i;\theta,\phi)}{\nu_i}\\ &= \prod_{i=1}^{n} \frac{2\pi y_i \sum_{j=1}^{J}\phi_j g_j(y_i,\mathbf{z}_i;\theta_j)}{\nu_i}\end{aligned}$$

where ν_i (the *effective area of detection*) is defined as:

$$\nu_i = 2\pi\sum_{j=1}^{J}\phi_j \int_0^w y g_j(y,\mathbf{z}_i;\theta_j)\mathrm{d}y. \tag{2}$$

For both line and point transects, parameters are estimated using maximum likelihood. Practicalities associated with this maximization, along with analytic derivatives of the likelihood are described in Appendix S1 and Text S1. The best number of mixture components to use for inference can be determined using standard model selection techniques, such as Akaike's Information Criterion (AIC), and goodness-of-fit of fitted models can be assessed just as for K+A models using, for example quantile-quantile plots and Kolmogorov-Smironov tests (see Buckland et al. [2], Section 11.11).

In this article, we assume the distance data are in the form of "exact" object-transect distances; alternatively, distances can be grouped into intervals, with pre-defined cutpoints (e.g., 0-10m, 10-20m,

etc.), so that the data are the distance interval of each observation. In this case, a multinomial likelihood is obtained (see, e.g. Buckland et al. [1], Section 3.3.2). Also, in some cases (e.g., some aerial surveys), objects below a defined distance are not counted – so-called "left truncation" [1 Section 4.3.2]. The likelihood is readily amended to account for this, by changing the lower limit of integration in equation (1) or (2).

Estimating population size

Population size can be estimated using the Horvitz-Thompson-like estimator [5]:

$$\hat{N} = \frac{A}{a} \sum_{i=1}^{n} \frac{1}{\hat{p}_i} \tag{3}$$

where A is the area of the study region for which population size is being estimated, a is the size of the sampled area, and p_i is the probability of the i^{th} observation being detected given it is within the sampled area. For line transects, $a = 2wL$ where L is the total line length, and

$$\hat{p}_i = \frac{1}{w} \sum_{j=1}^{J} \hat{\phi}_j \int_0^w g_j(y, \mathbf{z}_i; \hat{\theta}_j) \mathrm{d}y.$$

For point transects, $a = \pi w^2 k$ where k is the number of points, and

$$\hat{p}_i = \frac{2\pi}{w^2} \sum_{j=1}^{J} \hat{\phi}_j \int_0^w y g_j(y, \mathbf{z}_i; \hat{\theta}_j) \mathrm{d}y.$$

A standard summary statistic is the average detection probability for an animal within the sampled area, $\hat{P}_a$, which is given by:

$$\hat{P}_a = n/\hat{N}.$$

Estimators for the variances of $\hat{N}$ and $\hat{P}_a$ are given in Text S2.

Examples

Simulated data. Extensive simulations were carried out to investigate performance (in terms of the accuracy of estimation of P_a) when the true detection function model is not known to the estimation procedure. Buckland et al. [1] show that accurate results are readily obtained in situations where there is a wide "shoulder" of high detection probability at small and medium distances: in such situations, the dependence on having a good detection function model is only slight. Hence, we focus here on a variety of more challenging scenarios.

Each simulation involved generating 200 replicate datasets from a specified detection function model (assuming the entire study area was included within the surveyed transects, i.e., $A = a$ in equation (3), and a truncation distance of $w = 1$), fitting each dataset with a range of mixture and key series plus series adjustment (K+A) models, and in each case recording estimated parameter values and abundance from the model with the lowest AIC in each of: mixture models, K+A models, and both combined. Mixture models with 1-, 2-, and 3-point half-normal components were fitted to the data along with two K+A models: half-normal plus cosine adjustments and hazard rate plus simple polynomial adjustments, both with monotonicity constraints implemented as described above and with a maximum of 3 adjustments. Mixture models and K+A models were fitted using the R packages `mmds` (version 1.1) and `Distance` (a simplified interface to `mrds`; version 0.6.1) respectively, both written by the authors.

Fourteen different simulation scenarios were investigated, in five groups, as described below and illustrated in Figure 2, one line per group. True parameter values and summary statistics are given in

Appendix S2. For each scenario, a simulation was performed at each of five sample sizes (number of observations, n): 30 (low), 60 (recommended minimum for line transects [1]), 120 (adequate), 480 (large) and 960 (very large). We anticipated performance would depend upon sample size, because: the methods are likelihood-based and hence only asymptotically unbiased even if the correct model is fitted; the use of AIC to select model complexity meant that more flexible (and hence accurate) models could be expected to be selected given larger sample sizes. Mixture models are "parameter hungry" compared with K+A models, in the sense that each additional mixture component requires 2 extra parameters, while each additional adjustment term requires only one and hence, given the use of AIC for model selection, the relative performance of the two approaches may change at different sample sizes.

Group A. Line transect with 2-point half-normal mixture detection functions. Four scenarios were tested, representing a range of potentially challenging detection functions. Scenarios A1 and A2 both have mixture components with quite different scale parameters, but in A1 the majority of data come from the less detectable component while in A2 it comes from the more detectable component. A3 tests the behaviour of the models when the scale parameter of one of the mixture components is very large relative to the truncation distance. A4 has a large spike (i.e., a sharp decline in detectability at small distances), which is similar to some of the data we analyse in the case studies, below.

Group B. Point transect with detection functions as in the previous scenario. The geometry of point transect sampling means there are few animals close to the point relative to those at larger distances. Hence, for a given sample size of observations, there are far fewer at small distances than for line transects, making it harder to accurately model the detection function in the critical region close to the point. We therefore anticipate that performance will be worse for point transects. For this group, Figure 2 shows pdfs of the observed distances.

Group C. Line transect with 3-point half-normal detection functions. Two scenarios were tested. C1 has a detection function much like A2, enabling us to investigate the efficacy of model selection (i.e., we expect a 2-point mixture to be selected and to produce good results). C2 is a more complex shape that could only be created using a 3-point mixture; it has the added complication (as with A3) that one of the components has a large scale parameter relative to truncation distance.

Group D. Line transect with 2-point half-normal detection functions, and additional covariates. We used covariate models to test two aspects of model robustness. In the first, we assumed the covariate values were observed, and included covariate models in the candidate set, along with distance-only models. Our prediction was that (at large sample sizes at least) covariate models would be selected and estimation of P_a unbiased. In the second, we assumed the covariate values were not observed, and hence covariate models were not in the candidate set. Our expectation was that (at larger sample sizes) more complex mixture distributions would be selected to compensate for the additional unobserved complexity, and that estimation of P_a would not be greatly affected. Two scenarios were tested. D1 had a binary factor covariate, with half the observations having one covariate value and half the other. D2 had a continuous covariate, whose fixed values were generated from a standard normal distribution function. Detection functions are shown in the fourth row of Figure 2, along with the marginal detection functions for the levels/quartiles of the covariates. Note that, for the unobserved covariate models, D1 is equivalent to a 4-point mixture, while D2 is equivalent to a 2-point continuous mixture; neither of these models were in the candidate model set. In the case of the K+A models, and in line with common practice, if covariates were included in the models then adjustment terms were not.

Group E. Line transect with other detection functions. The above models all use the same functional form for g_j in generation and fitting. Here we tested the model robustness using two alternative data generating functions, not in the candidate model set (see Appendix S2 for formulation). E1 used an exponential power series function (a generalization of the half-normal function with an additional shape parameter); E2 used a mixture of two hazard-rate functions, giving a shape that may be difficult to fit with half-normal models.

Case studies. The first two case studies return to the datasets depicted in Figure 1, and demonstrate

how the mixture formulation solves the issue of non-monotonic detection functions. The first case study also includes two other species, illustrating how the new approach can fit real data as well as, or better than, the K+A approach. The second case study gives an example of when covariate models can cause non-monotonicity. The third case study demonstrates modelling of spiked line transect data, while the fourth gives a point transect example, with covariates.

Case study: British Columbia marine mammals. Williams and Thomas [16] used a data from a line transect survey to study several species of marine mammal off the coast of British Columbia, Canada. Here, we investigate three species: harbour seal (*Phoca vitulina*) in water (the data also contained observations of hauled-out animals, which were not analysed here), harbour porpoise (*Phocoena phocoena*) and humpback whale (*Megaptera novaeangliae*). Truncation distances were set at 500m, 500m and 2000m for each species respectively, giving sample sizes of 232, 59, and 70 observations.

Case study: Long-finned pilot whales. Pike et al. [8] analyzed observations of 84 pods of long-finned pilot whales (*Globicephala melas*), sighted as part of a line transect survey, the North Atlantic Sightings Survey NASS-2001. The Beaufort sea state was recorded as a covariate during the survey and enters the authors' model as either a continuous variable, or a factor with 2 levels (0-1, 2+), 3 levels (0-1, 2, 3+), or 5 levels (0, 1, 2, 3, 4, with one value of 3.5 coded as 4).

Case study: Wood ants. Borkin et al. [17] analyse data on two species of wood ant (*Formica aquilonia* and *Formica lugubris*) collected during a line transect survey of the Abernethy Forest, Scotland, in 2003. The number of nests sighted was 150, with the farthest being 72.04m from the transect, although 45% of the nest sightings lay within 4m of the line. As part of their analysis, several different truncation distances were used. Larger truncation distances led to a large variance in the encounter rate estimates and hence in overall abundance estimates [17]. This is due to the spike caused by the large number of detections close to the line (see Figure S4). As well as distances, three covariates were recorded: habitat type (a four level factor), the size of each nest (a continuous variable, calculated as half-width multiplied by height) and species (a two level factor).

Case study: Amakihi. Marques et al. [15] analyse point transect data on a Hawaiian songbird, the Amakihi (*Hemignathus virens*). The data consist of 1243 observations (after truncation at 82.5m), collected at 41 points between 1992 and 1995, together with three covariates in addition to distance: the observer (a three level factor), minutes after sunrise (continuous) and hours after sunrise (a six level factor).

Results

Simulation results

Figure 3 summarizes the estimates of P_a obtained if the candidate model set contains only mixture models (including 1-point mixtures); the numbers below each boxplot are the proportion of times the correct model was selected. Figure S1 shows the distribution of estimates when only K+A models are used, giving a baseline to compare the mixture model results against (Figure S2). Estimates using the recommended modelling strategy of both mixture and K+A models is shown in Figure S2, and the proportion of time each model is chosen using the combined modelling strategy is given in Figure S3.

For Group A, the mixture approach produced unbiased results for scenarios A1 and A3 at all but the lowest sample size; even for the $n = 30$ scenarios the bias was small, despite the correct 2-point mixture model being selected only 48-60% of the time (Figure 3 – the half normal model was selected the remainder of the time). The K+A approach also performed well (Figure S1). Unsurprisingly, therefore, the combined approach performed well (Figure S2); what was a little surprising was that the correct model was only selected 60-76% of the time at the highest sample sizes for scenario A1, with the hazard-rate K+A model selected the remainder (Figure S3). Scenarios A2 and A4 showed positive bias at smaller sample sizes under the mixture approach; bias reduced substantially by 480 observations, where a large

proportion of the selected models were 2-point mixtures. Unlike scenarios A1 and A3, the detection functions in scenarios A2 and A4 were evidently not well approximated by a half-normal, and hence at lower sample sizes where the two point mixture tended not to be selected, the results were biased. The K+A approach did not fare well with these scenarios, showing strong positive bias even at the largest sample sizes. In combination, the mixture models were chosen over K+A models at larger sample sizes, and so the combined modelling approach produced much better results than K+A alone.

As expected, results were worse for the point transect scenarios of Group B. Estimates from the mixture approach were biased at low sample sizes for B1, when the two-point model was rarely selected, but were unbiased given 120 observations and greater. Estimates for B3 were unbiased. For B2 and B4, results were positively biased at small sample sizes, just as with A2 and A4, but unlike the line transect scenarios the bias did not disappear even at the largest sample size. This is unsurprising given the very small number of detections coming from the less detectable mixture component (see Figure 2 – the marginal pdf is almost identical to that of the easier to detect mixture component). Bias was generally worse with the K+A approach (Figure S1), and the combined approach (Figure S2) produced marginally better results than K+A alone; for scenarios B1 and B3 the combined results were much better than K+A alone.

Group C were the 3-point mixture scenarios. For C1, results were similar to A2 – unsurprising, given the similarity in detection functions. A 3-point mixture model was almost never chosen by AIC (Figure 3). For C2, estimates were surprisingly good, even when the 3-point mixture was not the selected model, at lower sample sizes. Evidently, the function is well approximated by a 2-point mixture, although at larger sample sizes ($n = 480$ and above), the 3-point model is preferred by AIC. In both cases, the K+A results were worse (Figure S1), although they were not far from unbiased for C2. In the combined results, the mixture models were chosen most (71-84%) of the time for model C1, while for C2 the mixture models were chosen less often (40-69% of the time); despite this, the results were just as good as those using mixtures alone (Figure S2).

We first address the results of the Group D simulations when covariates were available for inclusion in candidate models. Results for D1 were positively biased at lower sample sizes, but less so as the sample size increased, and almost unbiased by 120 observations, where the correct model was selected most of the time (Figure 3). Results for D2 were close to unbiased at all sample sizes. For the K+A models, estimates were positively biased at almost all sample sizes, with no patten of decreasing bias with increasing sample size (Figure S1). Positive bias is not surprising for the K+A models with covariates because no adjustment terms were used in that case. The combined results were as good as, or better than, either method alone, even though the mixture models were commonly not the ones selected by AIC (Figure S2 - even at the largest sample size, the correct model was selected 64% and 61% of the time for scenarios D1 and D2 respectively, echoing the results from Group A).

When covariate information is not available for fitting the model, the mixture model detection functions still performed well, showing that when covariates are not available mixture components can compensate, though not through using additional components (see Figure S3, 3-point mixtures are never AIC-best models). For the K+A models without covariates, performance was also similar to that from the covariate models, indicating that the flexibility provided by the series adjustment can compensate for lack of covariate information in that framework. However, results were still slighlty biased even at large sample sizes (Figure S1). As might be expected, bias was less when both approaches were combined (Figure S2).

The Group E results were encouraging. Although the mixture formulation was biased even at large sample sizes, the bias was always small (Figure 3), and generally no worse than that under the K+A formulation, which also showed a small bias (Figure S1). An exception was scenario E2 (the mixture of hazard-rate functions) where the K+A formulation was unbiased at the largest sample size. We had anticipated good performance of the mixture models for scenario E1, since the detection function shape is not far from half-normal; however it was not obvious that performance would be good for E2, where

the marginal shape cannot be approximated well by a mixture of half-normal functions. The combined strategy was no worse than either formulation alone in terms of bias, although the variability in estimates between simulation runs was generally higher than for the K+A approach alone (Figure S2).

Case studies results

British Columbia marine mammals. Results are summarised in Table 1 and detection functions for the AIC-best models are shown in Figure 4. In each case mixture models were two component models. For harbour seal, the mixture model had a lower AIC than for the K+A model reported in Williams and Thomas [16]. The $\hat{P}_a$ is approximately 20% lower, implying that the previous estimate of $\hat{N}$ may have been an overestimate. For harbour porpoise, the mixture model AIC is almost 1.5 points higher than the K+A model, which was a hazard-rate with no adjustments. Hence, the model deviances are very similar, but the penalty due to the 2-point mixture having an additional parameter prevents it from being selected. The $\hat{P}_a$ from the two models are very close. Lastly, for humpback whales, the mixture model AIC is almost 3 points higher than the K+A model – however, one advantage of the mixture model is that the fitted function is monotone (Figure 4) while the K+A function is not (Figure 1). Again, the estimated $\hat{P}_a$s are very similar.

Long-finned pilot whales. A mixture model detection function was fitted with each covariate, as well as a model with no covariates. The best model by AIC score (Table 1) was a 2-point mixture with Beaufort sea state included as a continuous covariate. Figure 5 shows the average detection function (in the sense that a detection function was evaluated over the range $(0, w)$ for each covariate combination and was then averaged point-wise) and the marginal detection function with the quartiles of Beaufort sea state. None of the non-monotonic behaviour seen in Figure 1 can occur when a mixture is used.

Wood ants. All combinations of main effects were fitted (Table 1), and the best model by AIC was a 2-point mixture with nest size and habitat as covariates (Figure S4). This model had an AIC that was considerably (6 points) lower than the AIC-best K+A model, a hazard-rate with the same covariates. $\hat{P}_a$ is about 10% lower when estimated using the mixture model.

Amakihi. The AIC-best mixture model was a two point mixture with observer and minutes after sunrise as covariates (Figure S5), closely followed by the model with only observer as a covariate (Table 1). In this case a hazard-rate with observer and minutes after sunrise as covariates performed better than mixtures in AIC terms, although by less than 1 AIC point. The difference in $\hat{P}_a$ between these two models is about 15%. It is encouraging that there is such a small difference in AIC, and that covariate mixture models were selected over mixture models without covariates, despite the large number of parameters that such models entail.

Discussion

We have investigated and demonstrated the utility of detection functions constructed from mixtures of half-normal functions in both line and point transect distance sampling. We also show that covariates can be readily included in such models. Further, these mixture detection functions can be simply "dropped into" other extensions of conventional distance sampling such as: methods for dealing with incomplete detection at zero distance [18, 19] (for these models, there is an additional mark-recapture component to the likelihood, where mixture models could also be used) or spatial models for distance sampling data [20, 21].

We have shown that the mixture models perform well on both simulated and real data where traditional methods produce suboptimal results. In many cases the proposed model outperformed K+A models in AIC terms, which is surprising given that the K+A formulation is designed to produce parsimonious and realistic fits, and that the mixture models in question often had more parameters. In particular mixture model detection functions appear useful when dealing with line transect data that has

a spike in detection probability at small distances, though we note that it is better to avoid collecting such data in the first place, where possible [1]. Also, other non-detection-related factors can cause a spike, such as rounding of measurements or responsive animal movement, and if present in the data these should be dealt with using other analysis strategies or field methods [1 see]. For line transect surveys, unbiased estimation of P_a was possible even for very spiked detection functions, so long as the sample size of observations was large (Scenarios A2 and A4). By contrast, estimates remained badly biased at all sample sizes for the equivalent point transect scenarios (B2 and B4). For such surveys, where such a small proportion of the data comes from the closer distances, then perhaps the only effective solution is to constrain the fit, for example using a Bayesian approach with strong priors on the detection function parameters.

We note that in our case studies, a larger coefficient of variation was reported with mixtures than with half-normal K+A models but mixtures seemed to have lower CVs than hazard-rate K+A models (in the line transect case, ignoring fin whales since the K+A model was non-monotonic; see Table 1). This can be explained by considering the behaviour of the detection function near zero distance. For a a single half-normal function the relatively swift drop-off in probability of detection leads to small uncertainty in estimates near zero. However, for the hazard-rate the shoulder can be very large (depending on the shape parameter), so the uncertainty surrounding the estimate of detection probability at zero distance is large. Mixtures of half-normals lie somewhere in between these two options (summing smaller variances).

Simulations show that small sample sizes do not support the use of mixture models with a high number of components, even when the data were generated from such a model. We avoid poorly fitting models of this sort by using both K+A and mixture detection functions and selecting the best between them (comparing Fig. 3 with Figure S1). This integrated approach is builds upon current model selection procedures for a detection function analysis – currently selection is made between different K+A formulations and number of adjustment terms using AIC; mixture models simply add another alternative detection function where rather than adjustment terms, mixture components are selected. So existing key-only models are special cases of the mixture detection functions.

In simulation we observed that 3-point mixture did not act as good surrogates for missing covariate information; 2-point mixtures were generally chosen by AIC as good models (however these models were useful). In our case studies, 2-point mixtures consistently provided the best fit. Only examination of further data will show whether 3-point and higher mixtures can be supported, however we note that when the K+A series formulation is used, detection functions with 5 or more parameters are rarely selected by AIC (a 3-point mixture with no covariates requires 5 parameters).

We have compared the new mixture approach for modelling detection functions with the most widely used alternative, K+A. However, other approaches exist, for example nonparametric and semiparametric kernel estimators (see Eidous [22] and references therein). So far as we are aware, all current alternatives fail some of the criteria given in the introduction – for example, the kernel functions can be non-monotonic. Giammarino & Quatto [23] have proposed a "mixture model" detection function – their model takes a rather different form to the mixtures we describe here (simply $\exp\left(-x^2/(2\sigma^2)\right) - x/\tau\right)$), though according to their results there seems to be little difference between their model and K+A approaches. Nevertheless, it may be useful to extend the comparisons presented here to include a wider selection of methods.

The mixture component used here was a half-normal, but other component functions may prove useful. In particular, a mixture of hazard-rate functions with different shape and/or scale parameters for each component may be better at fitting detection functions with a wide shoulder, a steep drop-off and then a second plateau in detectability (see E2 in Figure 2, which was generated from a mixture of two hazard-rate functions). Further, a mixture of a half-normal (or hazard-rate) and a uniform kernel may proove useful – this would have only two (or three) parameters, and hence may be more competitive (in AIC terms) with K+A models.

Another potentially useful extension is continuous mixtures of the form

$$g(x) = \int_{\mathbf{R}} \varphi(\kappa) g_\kappa(x, \mathbf{Z}; \theta, \kappa) \mathrm{d}\kappa$$

where $\varphi(\kappa)$ is a weighting function that controls the mixing of g_κ. Provided that an appropriate function can be chosen for φ, more flexible models could be used whilst keeping the number of parameters low. In addition, a combination of both finite and continuous mixtures could be used, echoing the work in capture-recapture [12].

Mixture model detection functions based on half-normal components are available as an R package, `mmds`, which is available on CRAN. These models will be added to the next version of the Distance software and the R package `Distance`.

Supporting Information

Appendix S1 Optimization details (PDF)

Appendix S2 Simulation parameters (PDF)

Text S1 Derivatives of the likelihood (PDF)

Text S2 Variance estimation for mixture model detection functions (PDF)

Figure S1 Simulation results: boxplots of the estimated average detection probabilities, P_a, for the best K+A model (by AIC score). Grey lines indicate the true value of the average detection probability. (PDF)

Figure S2 Simulation results: boxplots of the estimated average detection probabilities, P_a, for the best model (by AIC score) for both mixture and K+A models. In each case the best overall model was selected, reflecting the modelling approach undertaken in practice. Grey lines indicate the true value of the average detection probability. Numbers underneath each boxplot give the proportion of AIC best models that were of the same form as the model that the data was simulated from (e.g., in scenario D1, the proportion of AIC best models that were 2-point mixtures that included the covariate in the model). Numbers above each model give the proportion of times that the AIC best model was a 2- or 3-point mixture model. (PDF)

Figure S3 Simulation results: stacked bar charts showing the number of models selected by AIC that fall into the given model classes. Layout is as in Figure S2. "hn" is a half-normal detection function and "hr" is a hazard-rate detection function (no adjustments/1-point mixture). K+A indicates a key function plus adjustment term model where "cos" is cosine and "poly" are simple polynomial adjustments. MMDS is a mixture model with 2 or 3 components ("2-pt" or "3-pt", respectively). "(cov)" indicates that covariates were included in the model (no adjustments were allowed when covariates were used). (PDF)

Figure S4 Plot of the detection functions for the AIC best model for the ants data set (2-point mixture with nest size and habitat as covariates). The first panel shows the average detection function (dashed lines are the two mixture components of the detection function, averaged over covariate values). The second and third panels show the quartiles of nest size and the levels of habitat type respectively.

Figure S5 Plots of the (AIC) best mixture model for the Amakihi data: a 2-point mixture with observer and minutes after sunrise as covariates. Top row: detection function averaged over covariates (dashed lines are each mixture component averaged over covariates), marginal detection function showing the levels of observer (averaged over the values of minutes after sunrise) and marginal detection function for minutes after sunrise ranging between 0 and 300 minutes (averaged over the levels of observer), as in Marques et al (2007). Bottom row: pdf of distances averaged over the covariate values.

Acknowledgements

DLM acknowledges the UK EPSRC for financial support during his PhD, and Simon Wood for useful discussions. Both authors thank David Borchers, who suggested the parametrisation for the mixture proportions (Appendix S1), Tiago Marques for helpful comments on an earlier draft, and the following people, who provided case study data: Rob Williams (B.C. marine mammals), Daniel Pike (pilot whales), Kerry Borkin (wood ants) and Steven Fancy (Amakihi).

References

1. Buckland ST, anderson DR, Burnham KP, Laake JL, Borchers DL, et al. (2001) Introduction to Distance Sampling. Oxford University Press.

2. Buckland ST, anderson DR, Burnham KP, Laake JL, Borchers DL, et al. (2004) Advanced Distance Sampling. Oxford University Press.

3. Buckland ST (1992) Fitting density functions with polynomials. Journal of the Royal Statistical Society Series C: Applied Statistics 41: 63–76.

4. Thomas L, Buckland ST, Rexstad EA, Laake JL, Strindberg S, et al. (2010) Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology 47: 5–14.

5. Marques F, Buckland ST (2003) Incorporating covariates into standard line transect analyses. Biometrics 59: 924–935.

6. Schittkowski K (1986) NLPQL: A Fortran subroutine for solving constrained nonlinear programming problems. Annals of Operations Research 5: 485–500.

7. Ye Y (1987) Interior Algorithms for Linear, Quadratic, and Linearly Constrained Convex Programming. Ph.D. thesis, Stanford University.

8. Pike DG, Gunnlaugsson T, Vikingsson AG, Desportes G, Mikkelson B (2003) An estimate of the abundance of long-finned pilot whales globicephala melas from the NASS-2001 shipboard survey. Technical Report SC/11/AE/10.

9. Pledger S (2000) Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56: 434–442.

10. Dorazio RM, Andrew Royle J (2003) Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 59: 351–364.

11. Pledger S (2005) The performance of mixture models in heterogeneous closed population capture-recapture. Biometrics 61: 868–873.

12. Morgan BJT, Ridout MS (2008) A new mixture model for capture heterogeneity. Journal of the Royal Statistical Society Series C: Applied Statistics 57: 433–446.

13. Link W (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. Biometrics 59: 1123–1130.

14. R Core Team (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria. URL `http://www.R-project.org/`.

15. Marques TA, Thomas L, Fancy S, Buckland ST (2007) Improving estimates of bird density using multiple-covariate distance sampling. The Auk 124: 1229–1243.

16. Williams R, Thomas L (2007) Distribution and abundance of marine mammals in the coastal waters of British Columbia, Canada. Journal of Cetacean Research and Management 9: 15.

17. Borkin KM, Summers RW, Thomas L (2012) Surveying abundance and stand type associations of *Formica aquilonia* and *F. lugubris* (*Hymenoptera: Formicidae*) nest mounds over an extensive area: Trialing a novel method. European Journal of Entomology 109: 47–53.

18. Laake JL, Borchers DL (2004) Methods for incomplete detection at zero distance. In: Buckland ST, anderson DR, Burnham KP, Laake JL, Borchers DL, et al., editors, Advanced Distance Sampling, Oxford University Press. pp. 48–70.

19. Laake JL, Collier BA, Morrison ML, Wilkins RN (2011) Point-based mark-recapture distance sampling. Journal of Agricultural, Biological, and Environmental Statistics 16: 389–408.

20. Hedley SL, Buckland ST (2004) Spatial models for line transect sampling. Journal of Agricultural, Biological, and Environmental Statistics 9: 181–199.

21. Miller DL, Burt ML, Rexstad EA (2013) Spatial models for distance sampling data: recent developments and future directions. Methods in Ecology and Evolution 4: 1001–1010.

22. Eidous O, Shakhatreh MK (2011) Asymptotic Unbiased Kernel Estimator for Line Transect Sampling. Communications in Statistics - Theory and Methods 40: 4353-4363.

23. Giammarino M, Quatto P (2014) On estimating Hooded crow density from line transect data through exponential mixture models. Environmental and Ecological Statistics .

Tables

Figure Legends

Figure 1. Two examples of detection functions that are not monotone, fitted using conventional key function plus adjustment methods in the software Distance. The left panel shows data from humpback whale: a half-normal detection function with cosine adjustments was selected by AIC [16] but even with constraints in place the detection function is non-monotonic, with a small secondary peak at approx. 1500m. The second and third panels show data and models fitted to long-finned pilot whale where a half-normal detection function was selected with cosine adjustments and Beaufort sea state as a covariate [8]. Due to the inclusion of covariates, no monotonicity constraints could be employed. The middle panel shows the detection function averaged over the covariate values and the right panel the marginal detection function for 25th, 50th and 75th quantiles of the Beaufort sea state covariate; non-monotonicity occurs at approx. 2500m.

Table 1. Comparison of case study analysis results.

Species	**Model**	**ΔAIC**	$\hat{P}_a$	$\%CV\hat{P}_a$	**K-S** p
Harbour seal (in water)	Hn + cos(2)	1.19	0.425	7.55	0.52
	Hn 2-pt	0	0.335	15.38	0.94
Harbour porpoise	Hr	0	0.212	32.0	0.99
	Hn 2-pt	1.43	0.254	18.18	0.99
Humpback whale	Hn + cos(2)	0	0.386	12.64	0.67
	Hn 2-pt	2.88	0.381	18.48	0.64
Long-finned pilot whales	Hn + cos(2) `BSS` (cont.)	1.94	0.452	8.69	0.48
	Hn 2-pt	13.86	0.295	17.17	0.95
	Hn 2-pt `BSS5`	0.29	0.208	28.84	0.82
	Hn 2-pt `BSS2`	0.43	0.211	23.39	0.95
	Hn 2-pt `BSS3`	11.71	0.270	17.46	0.99
	Hn 2-pt `BSS` (cont.)	0	0.216	24.17	0.67
Wood ants	Hr `nest.size` + `habitat`	6.29	0.195	21.72	0.89
	Hn 2-pt None	17.34	0.184	15.46	0.96
	Hn 2-pt `habitat`	14	0.188	14.85	0.97
	Hn 2-pt `species`	19.32	0.184	15.48	0.94
	Hn 2-pt `nest.size`	4.37	0.214	15.19	0.76
	Hn 2-pt `habitat` + `species`	15.96	0.186	14.94	0.99
	Hn 2-pt `habitat` + `nest.size`	0	0.179	17.55	0.72
	Hn 2-pt `nest.size` + `species`	4.65	0.210	15.84	0.77
	Hn 2-pt `nest.size` + `species` + `habitat`	1.81	0.178	18.09	0.83
Amakihi	Hr `obs` + `mas`	0	0.319	5.11	0.08
	Hn 2-pt None	28.1	0.283	6.21	0.12
	Hn 2-pt `obs`	1.31	0.279	5.86	0.04
	Hn 2-pt `has`	29.81	0.282	6.95	0.33
	Hn 2-pt `mas`	27.73	0.284	6.52	0.31
	Hn 2-pt `obs`+`has`	5.15	0.283	6.21	0.23
	Hn 2-pt `obs`+`mas`	0.69	0.279	6.1	0.14
	Hn 2-pt `mas`+`has`	31.79	0.282	6.97	0.43
	Hn 2-pt `mas`+`has`+`obs`	7.12	0.282	6.33	0.35

Comparison of results from Williams and Thomas [16], Pike et al. [8], Borkin et al. [17] and Marques et al. [15] with results from fitting mixture model detection functions. In each case the first line for each data set is the model from the original article. Mixture model components were selected by AIC (ΔAIC from best model is listed). In the table "(cont.)" denotes that the covariate was included in the model as continuous, otherwise covariates entered the model as factors (`BBSn` indicates Beaufort sea state with `n` factors). cos(x) indicates a Cosine adjustment of order x. K-S p is the p-value from a Kolmogorov-Smirnov goodness-of-fit test.

Figure 2. Plots of the models used in the simulation. Group A (top row): detection functions for four line transect senarios with no covariates (solid lines) and their constituent mixture components (dashed lines). Group B (second row): pdfs for four point transect simulations with no covariates (solid lines), with associates component pdfs (dashed lines), rescaled so the area under each curve is one; the detection functions are as in the top row. Group C (third row): two 3-point mixture scenarios for non-covariate line transect data. Group D (fourth row): two covariate model scenarios, the first two panels are for a binary covariate scenario, the second two for a continuous covariate scenario; first panels in each pair show the detection function averaged over the covariates (along with the mixture components, similarly averaged) and the second panels show marginal detection functions with the levels (or quartiles) of the detection function. Group E (fifth row): detection functions for two line transect scenarios using 2-point mixtures of hazard rate functions.

Figure 3. Simulation results: boxplots of the estimated average detection probabilities, P_a, for the best mixture model (by AIC score). Layout is as in Figure 2. Grey lines indicate the true value of the average detection probability. Numbers underneath each boxplot give the proportion of AIC best models that were of the same form as the model that the data was simulated from (e.g., Scenario D1 the proportion of AIC best models that were 2-point mixtures that included the covariate in the model).

Figure 4. Plots of the mixture model detection functions fit to the British Columbia marine mammal data. In each case the best mixture model by AIC was a 2-point mixture. Dashed lines show the mixture components.

Figure 5. The (AIC) best model for the long-finned pilot whale data: a 2-point mixture model detection function with Beaufort sea state as a continuous covariate. Left: the average detection function (detection function evaluated over the range $(0, w)$ for each covariate combination and was then averaged point-wise) with components as dashed lines. Right: the marginal detection function with the quantiles (25%, 50% and 75%) of the Beaufort sea state.