Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2011, **2**, 77–85 doi: 10.1111/j.2041-210X.2010.00046.x

Threshold choice and the analysis of protein marking data in long-distance dispersal studies

Frances S. Sivakoff[1]*, Jay A. Rosenheim[1] and James R. Hagler[2]

[1]*Department of Entomology, University of California, Davis, CA 95616, USA; and* [2]*Arid Land Agricultural Research Center, USDA-ARS, Maricopa, AZ 85138, USA*

Summary

1. A valuable technique in the study of insect movement is protein marking, a quantitative method where individuals are categorized as marked or unmarked based on the amount of foreign protein detected by an enzyme-linked immunosorbent assay (ELISA).

2. Whether individuals are considered marked or not is dependent on a threshold value chosen by the experimenter. The traditional method of choosing the threshold accepts some risk of false positives, wherein unmarked individuals are misclassified as marked. The error rate associated with this method, adopted from the rubidium marking literature, relies on assumptions violated by most ELISA data.

3. We critically examined the effect of violating these assumptions on the false positive rate. In long-distance dispersal studies where the ratio of unmarked to marked insects is high, false positives can seriously bias estimates of insect movement abilities.

4. Simulations demonstrated that the conventional method for choosing a threshold (i) masks the presence of false positives, (ii) results in a 10-fold higher than expected false positive rate, and (iii) relies on assumptions of normality that are rarely satisfied; non-normality produces further increases in false positive rates.

5. We provide some solutions by introducing a new procedure for choosing a threshold that decreases the incidence of false positives and allows data to be corrected for anticipated rates of false positives. This methodology should enhance researcher confidence in the data generated from dispersal studies using protein marking techniques.

Key-words: decision threshold, ELISA, false positive, long-distance dispersal, *Lygus hesperus*, protein marking

Introduction

The study of long-distance dispersal has long been considered an uncertain science. Whereas the importance of long-distance dispersal is well recognized (Higgins & Richardson 1999; Nathan 2001; Cain, Nathan & Levin 2003; Trakhtenbrot *et al.* 2005), its study has been impeded by the challenges associated with quantifying rare long-distance dispersal events. In particular, uncertainty is inherent in the study of long-distance dispersal, and the processes associated with long-distance dispersal are highly stochastic (Nathan *et al.* 2003). As proposed by Nathan *et al.* (2003), to estimate long-distance dispersal it is important to reduce the noninherent uncertainties by improving estimation methods.

Dispersal is often studied by quantifying population redistribution through mark–capture techniques (Turchin 1998; Southwood & Henderson 2000). A variety of methods have been employed to mark individuals (e.g., fluorescent dust, trace elements), including recently developed protein markers (Hagler *et al.* 1992; Hagler 1997; Hagler & Jackson 2001). These protein marks are detected with protein-specific enzyme-linked immunosorbent assays (ELISA), the result of which is a continuous variable in the form of an optical density (OD) reading. The experimenter uses a threshold to classify individuals as marked or unmarked. The threshold method commonly used was initially proposed by Stimmann (1974) (henceforth the conventional method) for rubidium marking. This threshold is defined as the mean level of marker in unmarked individuals plus three times the standard deviation of the unmarked distribution.

The extension of protein marking to use common inexpensive proteins (Jones *et al.* 2006) has made protein marking an

*Corresponding author. E-mail: fjsheller@ucdavis.edu
Correspondence site: http://www.respond2articles.com/MEE/

increasingly popular method for studying insect movement on increasingly large scales (Boina *et al.* 2009; Horton, Jones & Unruh 2009). It is therefore important to evaluate the conventional method for choosing a threshold, and to see what modifications might be needed to use this method in the context of ELISA. Specifically, the conventional method explicitly assumes a normal distribution of unmarked individuals and implicitly assumes the evaluation of a large number of unmarked individuals when setting the threshold. One goal of this study is to evaluate the conventional method when these assumptions are violated and to determine whether errors generated affect estimates of dispersal. We demonstrate that when applied to data generated from ELISA, the conventional threshold substantially underestimates the false positive (FP) rate, which can lead to inflated estimates of long-distance dispersal. The second goal is to introduce a new threshold procedure that has a low and quantifiable FP rate.

This study is motivated by our attempts to use protein marking data to study long-distance dispersal. We will first briefly introduce the topic of choosing a threshold and its effect on the ability to discriminate between two groups. We then present a simple simulation to illustrate the effect of FPs in long-distance dispersal data. We will return to this example near the conclusion of the study to demonstrate the proposed threshold procedure's improvement on dispersal estimates.

Discriminating between two groups based on a continuous metric

The binary classification of a continuous variable is accomplished with a threshold that divides cases into mutually exclusive classes (Forbes 1995). When the distributions of values for each group do not overlap the threshold is set at a value between the two distributions. In this case all of the individuals below the threshold will be correctly categorized as unmarked individuals (known as true negatives, TN), and all of the individuals above the threshold will be correctly categorized as marked (known as true positives, TP). More commonly, however, the distributions of the two groups overlap, resulting in classification errors (Fig. 1). The two types of errors are false positives (FP), when cases that are actually unmarked are misclassified as marked, and false negatives (FN), those cases that are actually marked but are incorrectly classified as unmarked. FPs are Type I errors, and FNs are Type II errors.

The optimal threshold for a given discrimination problem depends on the type of error that can be tolerated, which is dependent on the question under investigation. For example, in conservation-based models used to identify habitat areas for protection, FNs are more costly than FPs, because FNs would exclude habitat potentially vital to threatened species, while FPs would merely add protected habitat (Fielding & Bell 1997). Similarly, for a medical test where further tests or treatments are relatively harmless to healthy patients and beneficial to diseased patients, the cost of a FN is greater than the cost of a FP (Metz 1978).

Motivating example: protein marking data, false positives, and the study of long-distance dispersal

In long-distance dispersal studies employing protein marking techniques, FNs are not nearly as problematic as FPs. When an investigation is interested in the shape of the distribution of dispersal distances, the presence of FNs simply decreases the proportion of all captured individuals that are marked (prevalence), thereby reducing the power of the analysis. The importance of this cost is lessened because the marking technique allows for the large-scale field application of protein markers, which increases the prevalence of marked individuals.

The problem posed by FPs can be demonstrated by simulating the instantaneous point release of a marked population that disperses according to simple diffusion. We simulated the release of $N_0 = 10\,000$ marked individuals from a central point ($x = 0$) along a linear array. Marked individuals moved following a Gaussian distribution with mean (μ) 0 and population standard deviation (s) 2. Their distribution is a solution of the simple diffusion equation (Okubo 1980):

$$N(x,t) = \frac{N_0}{\sqrt{4\pi Dt}} \exp(-x^2/4Dt) \qquad \text{eqn 1}$$

where $N(x,t)$ is the density of marked individuals at position x and time t. The diffusion coefficient, D, is a measure of the rate of population spread and is set equal to 2 in this simulation.

Following Kareiva (1982), individual movement was restricted to the linear array ranging $\pm L$ units from the release point. At each distance $-L \le x \le L$, in addition to the marked

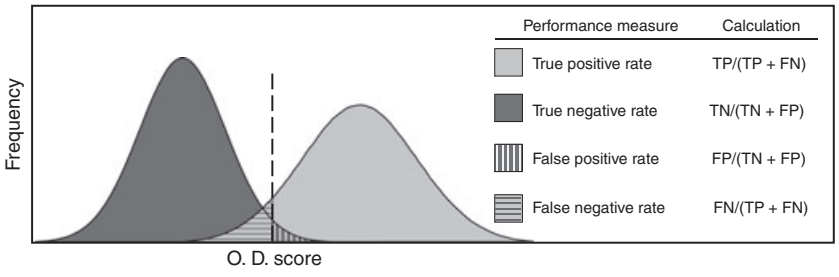

Fig. 1. Hypothetical distributions for unmarked (dark grey) and marked (light grey) populations and one possible decision threshold (dashed line). The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates are dependent on the choice of the threshold value, and each rate is estimated using the calculations listed above (after Metz 1978).

individuals there were $U(x,t) = Y - N(x,t)$ unmarked individuals, where $Y \gg N(x,t)$. A constant number of individuals (Z) was then randomly sampled without replacement from the set of $N(x,t) + U(x,t)$ at each distance, resulting in $\sum_{-L}^{+L} Z$ individuals sampled. The sampled individuals were then randomly assigned biologically realistic OD scores from their respective distributions. Figure 2a demonstrates the distributions of TPs and FPs assuming the use of a threshold that produces a FP rate of 4%. Using eqn 5 in Kareiva (1982) we estimated $D = 2{\cdot}02$ for the TP distribution, but when FPs are added, the combined distribution yielded $D = 3{\cdot}57$. FPs are especially problematic in the tail of the distribution and lead to an incorrect qualitative description of the distribution. Using the nonparametric Kolmogorov–Smirnov test, we found that the TP distribution is not significantly different from the distribution predicted with simple diffusion ($P = 0{\cdot}999$), but the hypothesis of simple diffusion is rejected when tested against the combined distribution ($P = 0{\cdot}011$).

It is clear that a low FP rate is critical for protein marking to be a viable technique for long-distance dispersal studies, where the tail of the distribution is of primary interest. While we aim to reduce the FP rate to zero, when distributions overlap the choice of a threshold always involves a trade-off: placing the threshold higher pushes FP rates down, but at the expense of a higher FN rate. A high FN rate results in fewer marked individuals available for analysis. Because of this trade-off, the optimal choice will often be associated with a small but nonzero FP rate. Thus, it will generally be important to have a way of estimating the FP rate so that the resulting data set can be corrected for this error.

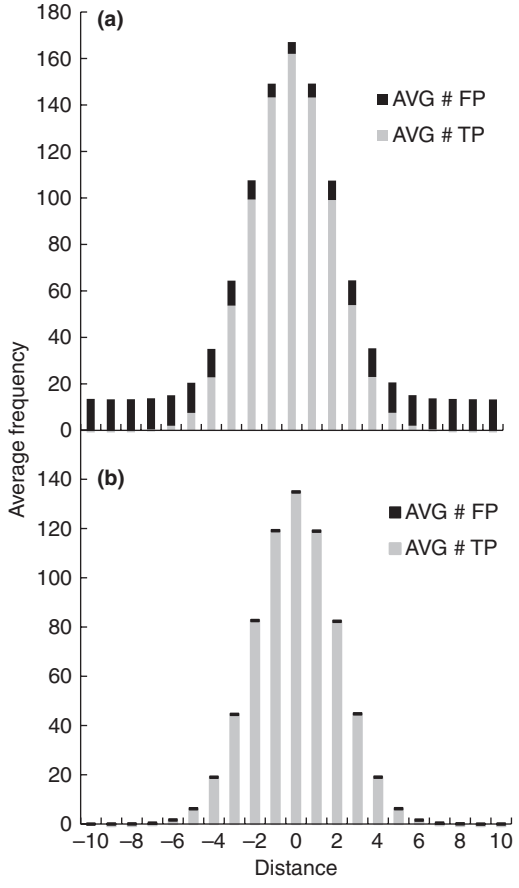

Fig. 2. Recapture distributions for individuals scored as marked using (a) the conventional threshold (FP rate = 4%) or (b) the maximum negative control threshold. The maximum negative control threshold was applied to standard normal variate transformed data.

The conventional method for choosing a threshold, and applications to ELISA

When the conventional threshold method is adopted for the analysis of protein marking data it is critical to examine the method's underlying assumptions to determine whether differences in application affect the method's performance. In its original rubidium context, the expected probability that an unmarked individual will exceed the threshold is 0·0013; we expect 1·3 cases in 1000 unmarked individuals to be erroneously classified as marked. This FP rate is dependent on the assumption that the distribution of unmarked values is normal. In reality, this assumption is rarely met, and distributions are often skewed to the right (Sutula *et al.* 1986). The FP rate also depends on the implicit assumption that a large number of known unmarked samples are considered with setting the threshold. Analysis of rubidium-marked samples is done on an individual basis, and all of the known unmarked samples tested can be considered in the selection of the threshold value. The assay of protein-marked samples, on the other hand, is done on 96-well microplates, which batches samples, and uses a small number of unmarked individuals to determine the threshold for each plate.

ELISA plates typically contain sample(s) known not to contain the protein (negative controls), sample(s) known to contain the protein (positive controls), and samples where the presence of the protein is unknown (e.g. field collected samples). In this study, we evaluate one common plate design, where each plate has eight negative controls and 80 field collected samples (Fig. S1a, Supporting Information). The remaining eight wells typically contain a combination of wells with only a buffer solution (blank wells) and positive controls (the known target protein). While ELISAs are widely used, a serious limitation of the technique is the variability commonly observed between microplates (plate effects; Clark & Adams 1977). Because of this variability between ELISA plates, thresholds have typically been calculated on a plate-by-plate basis. In protein marking studies, the conventional threshold is applied as:

$$\text{Threshold} = \mu_j + 3s_j \qquad \text{eqn 2}$$

where μ_j and s_j are the mean and the sample standard deviation, respectively, of the eight negative controls on plate j. We will now investigate the effect of choosing a threshold based on a sample of eight negative controls on the FP rate associated with the conventional method.

Problems with the conventional threshold

FP RATE IS MASKED

A serious and unexpected problem with the conventional threshold (eqn 2) is that it masks the FP rate. Recall that the expected FP rate for this method is 0·0013. To test this expectation, we simulated a set of 10 000 plates, populating each plate with 88 negative control OD scores drawn from a normal distribution. The first eight OD scores were designated as the plate's 'negative controls' and the plate's threshold was calculated using eqn 2. The plate's eight negative controls were then compared to this threshold, and any case with an OD score greater than the threshold value was considered a FP. Across the 10 000 plates, the observed FP rate was zero (no samples were identified as FPs, instead of the 104 that we expected to observe).

This unexpected failure to observe FPs among the negative controls results from the small number of negative controls on each plate and an unfortunate circularity inherent in the threshold algorithm. Because there are only eight negative controls per plate, each value has a large effect on s. When a plate contains a negative control sample with a very high OD score, the estimate of s is elevated. This in turn increases the threshold to such an extent that the very high OD score (which would have been categorized as a FP had a large number of negative controls been used to calculate the threshold) now falls below the threshold and is classified as a negative. In Appendix S1 (see Supporting Information) we demonstrate analytically that the conventional threshold will entirely mask FPs when there are ≤ 10 negative controls per plate.

We can 'unmask' the FPs within the negative control samples by calculating a threshold using all of the negative control samples on a plate other than the OD score under evaluation ($n = 7$). For each plate we calculated eight different thresholds, with each threshold excluding one of the negative control values on that plate. This method revealed the previously hidden FP rate: the observed average FP rate across the 10 000 plates was 0·016 ± 0·041 (±SD). The substantially elevated FP rate (0·016 vs. the expected 0·0013) is addressed in the following section.

FP RATE IS HIGHER THAN EXPECTED

A second major problem with the conventional method is that it yields a FP rate that is substantially higher than expected. To demonstrate this problem we present a simple extension to the simulation described above. As in the previous simulation, a plate of 88 OD values was drawn from a normal distribution of negative control values and the plate's threshold was calculated from the eight negative controls. The OD score for each of the 80 samples on the plate was compared to the calculated threshold value; any case with an OD score greater than the threshold value was considered to be a FP. The FP rate for each plate was then calculated as (the total number of FPs on that plate)/80. This procedure was replicated 10 000 times.

The simulation revealed a FP rate of 0·013 ± 0·033 (±SD). Thus, 1·3 in 100 individuals drawn from the simulated negative control sample were erroneously categorized as marked. The observed FP rate is an order of magnitude greater than the expected FP rate (0·0013, or 1·3 individuals per 1000 samples). This inflation occurs because the small negative control sample size generates substantial variability in estimates of s. On plates where s is substantially underestimated, the resulting threshold can be quite low – lower than several of the scores drawn from the negative control distribution – yielding FPs. Some plates will have no FPs, while others may have several (see example in Supporting Information, Fig. S1b,c). Across all of the plates, the result will be an average FP rate that is substantially higher than expected.

The hidden FP rate revealed in the previous section (FP rate = 0·016) is slightly larger than then FP rate observed in this simulation (FP rate = 0·013). The former is calculated from $n = 7$ negative controls, while the later is calculated from $n = 8$; this difference illustrates that the smaller the sample size of the negative controls, the greater the escalation of the FP rate.

FP RATE IS FURTHER ELEVATED WHEN DATA ARE NOT NORMALLY DISTRIBUTED

A third problem with the conventional threshold method arises when the assumption of normality is not met; non-normal data will often be associated with still higher FP rates. When a distribution is skewed to the right, as is frequently the case for negative control data, the area under the curve in the right tail of the distribution is larger, resulting in a higher FP rate (Fleischer *et al.* 1986; Hopper & Woolson 1991; Hsu 2007).

We demonstrate this problem with one of our own data sets, a mark–capture study investigating movement of *Lygus hesperus* (Hemiptera: Miridae). In this study, 5·7 hectares of alfalfa were sprayed with a solution of chicken egg whites. We use individuals collected before the spray as negative controls and individuals collected immediately after the spray as positive controls. Each individual was tested for the egg marker with an ELISA (for ELISA methodology see Jones *et al.* 2006).

As seen in Fig. 3, the distributions of *Lygus* negative and positive controls overlap. Furthermore, the distribution of *Lygus* negative controls deviates from normality (Fig. 3a; Shapiro Wilk W statistic = 0·85, $P < 0·0001$). The tail of the distribution is skewed to the right, with kurtosis = 5·59, and thus the distribution is described as leptokurtic (kurtosis > 3·0; Okubo & Levin 2001). For our leptokurtic negative control data set we repeated the simulation described in the previous section (which had revealed a higher than expected FP rate of 0·013). The average FP rate observed after 10 000 runs of the model was 0·044 ± 0·057 (±SD), more than 3× as high as observed for normally distributed data.

We can now summarize the full extent of the problem associated with the conventional threshold method: from an initial expectation of 1·3 FPs in a sample of 1000 negative controls, we see that the conventional approach can generate 44 FPs in a

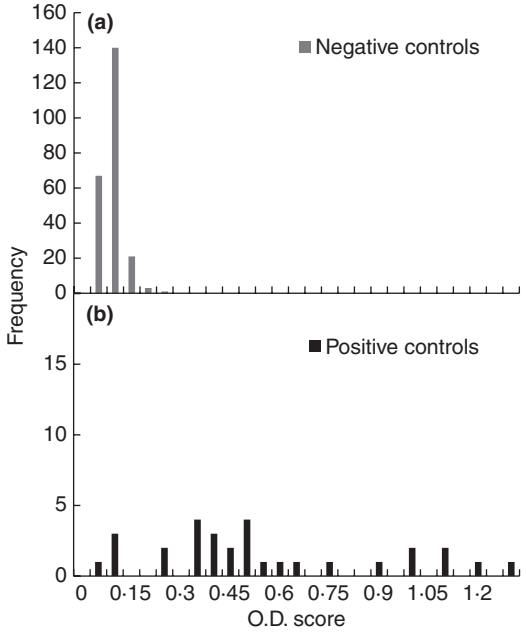

Fig. 3. Distribution of optical density scores of field collected *Lygus* samples tested for egg protein. (a) Negative controls ($n = 232$). The data do not conform to a normal distribution, (Shapiro Wilk W statistic = 0·85, $P < 0·0001$) and the distribution is categorized as leptokurtic (kurtosis = 5·59). (b) Positive Controls ($n = 30$).

sample of 1000 negative controls. At the same time, this error is masked, so that the researcher may be misled into believing that the originally expected FP rate (0·0013), or one still lower (0·00), is in place.

A proposed solution

We now propose a novel algorithm for choosing a threshold that addresses each of the three problems explained above. To solve these issues the method should be one in which (i) the true FP rate is revealed, so that we can estimate it; (ii) the experimenter can choose the FP rate to match the aims of the study; (iii) the algorithm is robust to deviations from a normal distribution of OD scores; and (iv) the algorithm works well for data sets with or without substantial plate effects.

PLATE EFFECTS AND THE STANDARD NORMAL VARIATE TRANSFORMATION

As demonstrated above, the conventional threshold method is problematic in large part because it is applied on a per-plate basis, which introduces issues associated with small sample sizes. This immediately suggests one possible solution: pool all negative control samples across plates to produce a larger sample of negative controls from which s can be calculated.

The pooling of negative controls across plates is not a good solution, because it ignores the plate effects that motivated researchers to calculate thresholds on a plate-by-plate basis in the first place. While steps can be taken to reduce variability between plates, it is rarely possible to eliminate plate effects entirely. The influence of plate effects on s estimated from a pooled sample can be observed by simulating scenarios with varying severities of plate effects. In each scenario we simulated 10 000 plates populated with samples drawn from a normal distribution ($\mu = 0·00$; $s = 1·0$). For each plate we drew a single 'plate effect' value from a second normal distribution ($\mu = 0·00$; σ varied; the larger the value of σ, the larger the plate effect). This plate effect value was then added to each of the 88 plate values, and s was calculated in two ways. In Method 1, s was calculated on a plate-by-plate basis, as is done in the conventional threshold method, and the mean of the 10 000 estimates of s was taken. In Method 2, a subsample ($n = 10\,000$) of the 80 000 negative controls was pooled and s was estimated from these 10 000 negative control values. This procedure was then repeated 10 000 times.

The results are summarized in Fig. 4. Method 1 provides a consistently accurate mean estimate of s as the variability between plates increases, but there is substantial variation in the estimates of s across plates (which leads to the inflation of the FP rate). For Method 2 we observe that increasing the magnitude of plate effects causes increases in the estimate of s. This increase in s will produce a strong increase in the threshold, and thus an undesirable decrease in the TP rate.

To address the problems associated with estimating s of the negative control distribution, we suggest the use of the standard normal variate (SNV) transformation for all OD scores on a plate. The SNV transformation (z) for sample i on plate j is:

$$z_{ij} = \frac{(X_{ij} - \hat{\mu}_j)}{s_j} \qquad \text{eqn 3}$$

where X_{ij} is the OD score of a sample in well i on plate j, and $\hat{\mu}_j$ is an estimate of the mean of the negative controls

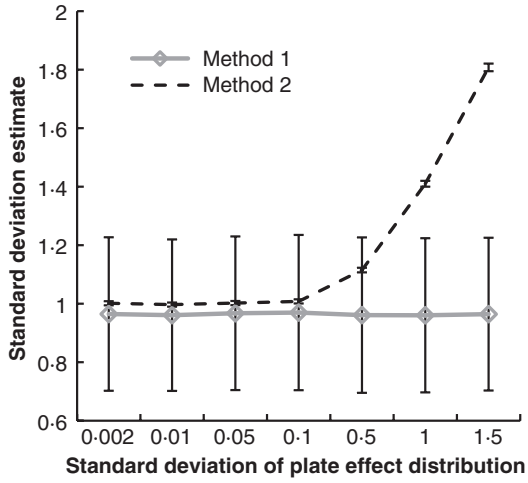

Fig. 4. Change in standard deviation estimate with increasing plate effect, displayed as the mean ± SD. In Method 1, the standard deviation was estimated on a plate-by-plate basis. In Method 2, the standard deviation was estimated from a pooled sample of 10 000 negative controls.

on plate j. When transforming the 80 samples on the plate $\hat{\mu}_j$ is estimated from all eight negative controls, but when transforming the negative controls $\hat{\mu}_j$ is calculated using only the other seven values. If the negative control values are not transformed in this way a problem arises that is similar to the masking problem demonstrated with the conventional threshold (see section FP RATE IS MASKED).

One simple and effective way to obtain a serviceable estimate of s_j is to calculate s_j for the eight negative controls found on a given plate, repeat this across many plates, and average the multiple estimates (see 'Method 1' in Fig. 4). By averaging across multiple estimates, we avoid the problem of estimating s_j from a small sample of observations. We suggest, however, that some improvement in estimating s_j may be obtained by anticipating a likely relationship between $\hat{\mu}_j$ and s_j when viewed across many plates. The relationship between $\hat{\mu}_j$ and s_j can be approximated using a power law function, which averages across values of s_j observed for plates that have similar $\hat{\mu}_j$ values. The power law function should provide a good estimate of s_j even when the estimated slope of the function is not significantly different from zero (in that case, the method essentially reverts to Method 1).

The power law function, originally from Huxley's simple allometry equation ($y = bx^m$), is

$$s_j = c\hat{\mu}_j^k \qquad \text{eqn 4}$$

The parameters k and c are estimated by taking the natural logarithm (ln) of eqn 4

$$\ln(s_j) = \ln(c) + k\ln(\hat{\mu}_j) \qquad \text{eqn 5}$$

and regressing the ln of s on the ln of μ (e.g., Caciagli & Verderio 2003). The values used in the regression are the μ and s of the eight negative controls of each plate.

Substituting eqn 4 into eqn 3, the final equation for the SNV transformation becomes:

$$z_{ij} = \left(\frac{X_{ij} - \hat{\mu}_j}{c\hat{\mu}_j^k}\right) \qquad \text{eqn 6}$$

Because the SNV transformation accommodates the variation introduced by plate effects, transformed data from multiple plates can be pooled prior to setting a threshold. This is true both for samples whose origin (i.e. marked or unmarked) is unknown and for the negative controls, facilitating the selection of a single threshold value.

CHOOSING THE FP RATE

Although the conventional threshold is used widely and is thus viewed by many as the default method for choosing a threshold value, the choice of a threshold remains subjective, and the best approach will depend on the research question and the amount of uncertainty that can be tolerated. For studies of long-distance dispersal, where it is desirable to have a low FP rate, we suggest that one useful approach is simply to set the threshold just above the highest observed negative control z score, as first proposed by Hopper & Woolson (1991).

When the threshold is set just above the highest observed negative control score, it is tempting to infer that the FP rate associated with this threshold is zero – there are no negative control scores greater than the threshold. But for any given distribution of negative control scores, the largest observed value is expected to increase as sample size increases (Hopper 1991). Thus, simply because we do not observe any negative control scores greater than our chosen threshold does not mean that the FP rate is zero.

ESTIMATING THE FP RATE BY BOOTSTRAPPING

One way to estimate the FP rate is to use the resampling technique of bootstrapping. Using this technique, we sample with replacement from a set of negative control scores to generate many test data sets of equal size to the original data set. The overall expectation for the FP rate can then be calculated as the mean of the FP rates observed across a large number of test data sets (Verbyla & Litvaitis 1989).

We demonstrate this procedure with our *Lygus* data set. We first SNV transformed the data by determining the relationship between $\hat{\mu}$ and s of the negative controls on a given plate. As seen in Fig. 5, for the 30 plates that assayed *Lygus* individuals for the egg marker, the slope parameter, $k = 1{\cdot}215 \pm 0{\cdot}388$ ($\pm$SE), and the y-intercept, $\ln(c) = -0{\cdot}968 \pm 1{\cdot}063$ ($\pm$SE). Thus, the SNV transformation was applied to the *Lygus* data as:

$$Z_{ij} = \left(\frac{X_{ij} - \hat{\mu}_j}{\exp(-0{\cdot}968)\hat{\mu}_j^{1{\cdot}215}}\right) \qquad \text{eqn 7}$$

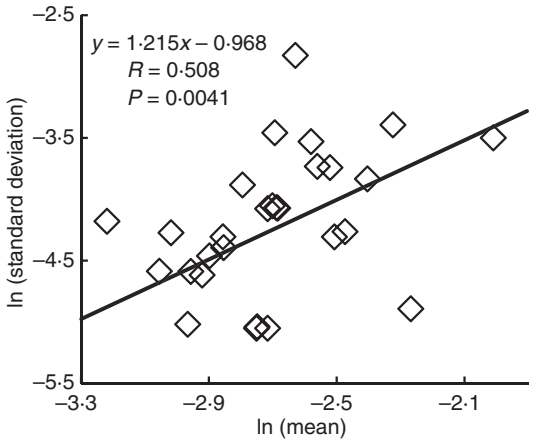

Fig. 5. Linear regression of the natural logarithm (ln) of the standard deviation of the negative controls from each plate ($n = 30$) vs. the ln of their respective means. From the regression equation we estimate the power law parameters: $k = 1{\cdot}215$ and ln (c) $= -0{\cdot}968$.

By bootstrapping the original data set 10 000 times we calculated an expected FP rate of 0·0025 ± 0·0040 (± SD), or an expectation that 2·5 in 1000 unmarked individuals will be incorrectly classified as marked.

THRESHOLD METHOD COMPARISON: POSITIVE PREDICTIVE VALUES

To compare the method proposed in this study (the 'maximum negative control' algorithm) with the conventional approach, we applied both methods to our *Lygus* data set and calculated the TP, TN, FP, and FN rates (Table 1). We also examined the performance of the conventional approach applied to SNV transformed data. For this case study, the maximum negative control method achieves a large improvement in the FP rate (0·0025 vs. 0·0444, a decrease of 94%) at the cost of only a modest increase in the FN rate (0·175 vs. 0·138, an increase of 27%).

Although different methods for choosing a threshold can be evaluated in part by comparing the metrics shown in Table 1, these values are affected by the prevalence of marked individuals. A useful metric that captures how well a given threshold method performs when applied to a set of samples whose state (i.e. marked or unmarked) is unknown is the positive predictive value (PPV; Zweig & Campbell 1993; Hsu 2007). The PPV is the probability that individuals of unknown state will be correctly classified as marked given that their OD scores are higher than the threshold. This metric is dependent on the ratio of the FP rate to the TP rate and, crucially, on the prevalence of marked individuals:

$$\mathrm{PPV} = \frac{1}{\left(\frac{\mathrm{FP\ \ rate}}{\mathrm{TP\ \ rate}}\right)\left(\frac{1\ -\ \mathrm{Prevalence}}{\mathrm{Prevalence}}\right) + 1} \qquad \text{eqn 8}$$

We used the PPV to compare the maximum negative control threshold method to the conventional threshold method for prevalence values ranging from 0·1 to 50% (Fig. 6). The maximum negative control method has a higher PPV than the conventional method or when the conventional threshold is applied to SNV transformed data. The improvement is especially substantial when prevalence values are relatively low (e.g., 1–10%), as expected in studies of long-distance dispersal.

It is important that an experimenter be able to choose a threshold that results in a low FP rate and a high TP rate, reflected in a high PPV value. For the maximum negative control threshold, a large pool of negative controls is needed to have a low FP rate. If the total pool of negative control samples is small, an experimenter might elect to set the threshold one or two standard deviations higher than the highest observed negative control value. Whatever the chosen threshold, the FP rate is then still quantifiable with the bootstrapping technique described in this study. The threshold procedure proposed here thus affords the experimenter substantial flexibility in choosing the FP rate.

Application of the new threshold method to dispersal data

We now return to the 1-dimensional dispersal simulation presented earlier to demonstrate the improvements of the methodology presented here. The distributions presented in Fig. 2a were generated when the simulated samples were randomly placed on ELISA plates and analysed following the conventional threshold method. To demonstrate the improvements gained using the new threshold method, the OD scores from the same simulated ELISA plates were SNV transformed. After the transformation the highest negative control value across all of the plates was chosen as the threshold value, and all of the individuals were compared to that threshold. The average numbers of TPs and FPs at each distance over 1000 runs of the model are shown in Fig. 2b. As in the earlier dispersal simulation, we estimated D for the TP distribution alone (D = 2·02) and when FPs were included (D = 2·12). Further improvement to our estimate of D was gained when we corrected the data for the estimated FP rate = 0·0025; after this correction, D = 2·04. Finally, we tested each distribution using the nonparametric Kolmogorov–Smirnov test and found that neither the TP distribution alone (P = 0·999), nor the distribution with FPs (P = 0·304) was significantly different from the distribution predicted with simple diffusion.

Discussion

Protein markers are appealing as a marking method for studying movement because they are inexpensive, easy to obtain, analysed with a sensitive and specific ELISA, and can be applied to a large number of individuals directly in the field. As a result, the use of this marking technique continues to grow in popularity. While the technique has been used successfully in

Table 1. Comparison of different threshold methods in terms of their performance measures (± SD)

Method	TP rate	TN rate	FP rate	FN rate
Conventional	0·8621 ± 0·0418	0·9556 ± 0·0573	0·0444 ± 0·0573	0·1379 ± 0·0418
SNV transformed	0·8667	0·9845 ± 0·0068	0·0155 ± 0·0068	0·1333
Max Neg Control	0·8247 ± 0·0322	0·9975 ± 0·004	0·0025 ± 0·004	0·1753 ± 0·0322

The 'conventional' threshold calculates the threshold on a plate-by-plate basis using the model: Threshold = $\mu + 3s$. The 'SNV transformed' method uses the same model as the conventional threshold but first SNV transforms the data to remove the plate effect. 'Max Neg Control' chooses the highest SNV transformed negative control value as the threshold.
SNV, standard normal variate; TP, true positive; TN, true negative; FP, false positive; FN, false negative.

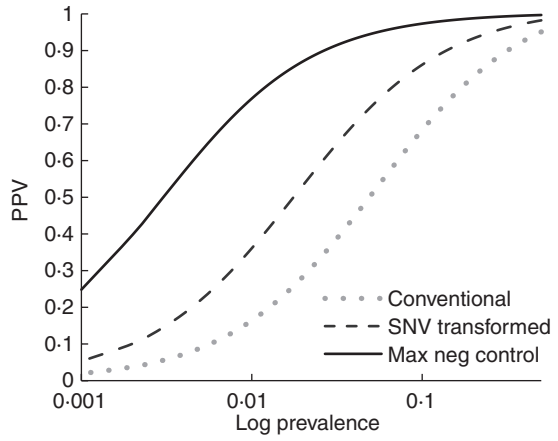

Fig. 6. Comparison of threshold determination methods using the positive predictive value (PPV) across a range of prevalence values (log scale). 'Max Neg Control' is the threshold proposed in this study, where the optical density scores have been standard normal variate (SNV) transformed and the threshold set as the highest observed negative control value. 'SNV transformed' is the threshold obtained by applying the conventional threshold method to SNV transformed data. 'Conventional' is the conventional threshold method applied to untransformed data.

small-scale movement studies (Jones *et al.* 2006; Boina *et al.* 2009; Horton, Jones & Unruh 2009), it has yet to be used to study long-distance dispersal. One difference between studies of local and long-distance dispersal is the prevalence of marked individuals at the furthest sample distances. In long-distance dispersal studies, where the expected prevalence of marked individuals is very low at the furthest sampling distances, the presence of FPs can drastically alter dispersal estimates. Thus, it is important to have a low FP rate. In this study, we determined that the FP rate associated with the conventional threshold is much higher than the expected value (0·0013) when applied to protein marking data analysed using ELISA. We have introduced a new approach that allows the investigator to choose a threshold and estimate the FP rate with bootstrapping, thereby avoiding the need to make assumptions about the underlying distributions. The precision of the FP rate estimate is improved with the use of the SNV transformation, which controls for between-plate variability.

Plate effects are widely recognized as being important in ELISA assays. Previous work has attempted to reduce differences between plates by using microplates from a single manufacturer and by batching all assays in a particular experiment (Clark & Adams 1977; Fenlon & Sopp 1991). Fenlon & Sopp (1991) recognized the importance of combining information across plates to improve threshold estimates. To do this, they used calibration data on all plates to remove between-plate differences. They also suggest that including more negative control samples on each plate (they only had two) has a large effect on the estimate of s. It is clear from the work presented here that the greater the number of negative controls on a plate, the better for estimating s, but increasing the number of negative controls trades off with having wells available for testing experimental samples. The SNV transformation we describe controls for between plate differences without sacrificing a large number of wells on each plate.

Our study underscores the importance of testing a large number of known unmarked and marked individuals (negative and positive controls). As discussed above, when the threshold is set as the maximum negative control value, the resulting FP rate is determined by number of individuals tested as negative controls. Furthermore, it is critical to have an accurate picture of both the unmarked and marked distributions to quantify the amount of overlap between them and the resulting influences on FP and TP rates. Measures can be taken prior to analysis of ELISA data that can further separate the negative and positive control distributions and reduce the classification errors discussed in this study. Such procedures include using a stop solution, lower concentrations of the secondary antibodies, and a dual-wavelength reading of each well (V. Jones, *pers. comm.*).

The ability to quantify the FP rate associated with a particular threshold might lead some observers to suggest that, regardless of the magnitude of the FP rate, we should be able to correct the data for that error. However, the FP rate is always estimated with some uncertainty. If, then, the FP rate is large relative to the prevalence (frequency) of truly marked individuals, the uncertainty in the FP estimate can introduce unacceptable levels of uncertainty in the corrected data. This is basically a case where the signal (i.e., prevalence of marked individuals) to noise (i.e., the FP rate) ratio becomes too small to achieve the desired confidence in the corrected data set. This underscores the importance of a working with a low FP rate in studies of long-distance dispersal. Our hope is that the methodologies presented here will improve researcher confidence in dispersal estimates generated from protein marking data and encourage the use of this technique for the study of long-distance dispersal.

Acknowledgements

We thank the Hagler and Rosenheim laboratory and field crews for their technical assistance. We also thank D. Sivakoff for help with the appendix, and C. Hsu, T. Unruh and V. Jones for helpful discussions. We are grateful to M. Holyoak for helpful comments on an earlier version of this manuscript. This publication was developed under a STAR Fellowship Assistance Agreement No. FP91-6838 by the U.S. Environmental Protection Agency (EPA). It has not been formally reviewed by EPA. The views expressed in this publication are solely those of the authors and EPA does not endorse any products or commercial services mentioned in this publication. Support was also provided by a grant from the USDA (RAMP grant ARZT-358320-G-30-505).

References

Boina, D.R., Meyer, W.L., Onagbola, E.O. & Stelinski, L.L. (2009) Quantifying dispersal of *Diaphorina citri* (Hemiptera: Psyllidae) by immunomarking and potential impact of unmanaged groves on commercial citrus management. *Environmental Entomology*, **38**, 1250–1258.

Caciagli, P. & Verderio, A. (2003) Experimental layout, data analysis, and thresholds in ELISA testing of maize for aphid-borne viruses. *Journal of Virological Methods*, **110**, 143–152.

Cain, M.L., Nathan, R. & Levin, S.A. (2003) Long-distance dispersal. *Ecology*, **84**, 1943–1944.

Clark, M.F. & Adams, A.N. (1977) Characteristics of microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *Journal of General Virology*, **34**, 475–483.

Fenlon, J.S. & Sopp, P.I. (1991) Some statistical considerations in the determination of thresholds in ELISA. *Annals of Applied Biology*, **119**, 177–189.
Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.
Fleischer, S.J., Gaylor, M.J., Hue, N.V. & Graham, L.C. (1986) Uptake and elimination of rubidium, a physiological marker, in adult *Lygus lineolaris* (Hemiptera: Miridae). *Annals of the Entomological Society of America*, **79**, 19–25.
Forbes, A.D. (1995) Classification-algorithm evaluation: five performance-measures based on confusion matrices. *Journal of Clinical Monitoring*, **11**, 189–206.
Hagler, J.R. (1997) Field retention of a novel mark–release–recapture method. *Environmental Entomology*, **26**, 1079–1086.
Hagler, J.R. & Jackson, C.G. (2001) Methods for marking insects: current techniques and future prospects. *Annual Review of Entomology*, **46**, 511–543.
Hagler, J.R., Cohen, A.C., Bradley-Dunlop, D. & Enriquez, F.J. (1992) New approach to mark insects for feeding and dispersal studies. *Environmental Entomology*, **21**, 20–25.
Higgins, S.I. & Richardson, D.M. (1999) Predicting plant migration rates in a changing world: the role of long-distance dispersal. *American Naturalist*, **153**, 464–475.
Hopper, K.R. (1991) Ecological applications of elemental labeling: analysis of dispersal, density, mortality, and feeding. *Southwestern Entomologist*, **14**, 71–83.
Hopper, K.R. & Woolson, E.A. (1991) Labeling a parasitic wasp, *Microplitis croceipes* (Hymenoptera: Braconidae), with trace-elements for mark recapture studies. *Annals of the Entomological Society of America*, **84**, 255–262.
Horton, D.R., Jones, V.P. & Unruh, T.R. (2009) Use of a new immunomarking method to assess movement by generalist predators between a cover crop and tree canopy in a pear orchard. *American Entomologist*, **55**, 49–56.
Hsu, C. (2007) *Spatial distribution of the European Corn Borer,* Ostrinia nubilalis *(Hubner) (Lepidoptera: Crambidae), and response of its parasitoid* Macrocentrus grandii *Goidanich (Hymenoptera: Braconidae) to host spatial heterogeneity*. PhD thesis, University of Minnesota, St. Paul.
Jones, V.P., Hagler, J.R., Brunner, J.F., Baker, C.C. & Wilburn, T.D. (2006) An inexpensive immunomarking technique for studying movement patterns of naturally occurring insect populations. *Environmental Entomology*, **35**, 827–836.
Kareiva, P. (1982) Experimental and mathematical analysis of herbivore movement – quantifying the influence of plant spacing and quality on foraging discrimination. *Ecological Monographs*, **52**, 261–282.
Metz, C.E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8**, 283–298.
Nathan, R. (2001) The challenges of studying dispersal. *Trends in Ecology and Evolution*, **16**, 481–483.
Nathan, R., Perry, G., Cronin, J.T., Strand, A.E. & Cain, M.L. (2003) Methods for estimating long-distance dispersal. *Oikos*, **103**, 261–273.
Okubo, A. (1980) *Diffusion and Ecological Problems: Mathematical Models*. Springer-Verlag, Heidelberg, Germany.
Okubo, A. & Levin, S.A. (2001) *Diffusion and Ecological Problems: Modern Perspectives*, 2nd edn. Springer-Verlag, New York.
Southwood, T.R.E. & Henderson, P.A. (2000) *Ecological Methods*. Blackwell Science, Oxford.
Stimmann, M.W. (1974) Marking insects with rubidium – imported cabbageworm marked in field. *Environmental Entomology*, **3**, 327–328.
Sutula, C.L., Gillett, J.M., Morrissey, S.M. & Ramsdell, D.C. (1986) Interpreting ELISA data and establishing the positive-negative threshold. *Plant Disease*, **70**, 722–726.
Trakhtenbrot, A., Nathan, R., Perry, G. & Richardson, D.M. (2005) The importance of long-distance dispersal in biodiversity conservation. *Diversity and Distributions*, **11**, 173–181.
Turchin, P. (1998) *Quantitative Analysis of Movement: Measuring and Modeling Population Redistribution in Animals and Plants*. Sinauer Associates, Sunderland, MA.
Verbyla, D.L. & Litvaitis, J.A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, **13**, 783–787.
Zweig, M.H. & Campbell, G. (1993) Receiver-operating characteristic (ROC) plots – a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561–577.

Received 7 April 2010; accepted 13 May 2010
Handling Editor: Robert P. Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Mathematical argument that the 'conventional algorithm' for setting a threshold results in a masking of the actual false positive rate.

Figure S1. (a) Microplate layout evaluated in this study, where there are eight empty wells in the first column (Blank), eight negative control samples in the 12th column (Neg), and 80 samples of unknown origin (grey cells). (b) Plate where the standard deviation of the negative controls, s ($s_{\text{plate}}\,1 = 0{\cdot}013$), is similar to that of the original negative control distribution, and the calculated threshold ($\text{Threshold}_{\text{plate}}\,1 = 0{\cdot}086$) is sufficiently large to classify correctly the samples on the plate as unmarked. (c) Plate where s ($s_{\text{plate}}\,2 = 0{\cdot}0042$) is less than that of the original negative control distribution. The resulting threshold ($\text{Threshold}_{\text{plate}}\,2 = 0{\cdot}064$) is too small to encompass the full variation of the negative control distribution, and as a result three of the samples on the plate are incorrectly categorized as positives (FPs highlighted in the shaded cells).

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Appendix 1

Here we present the argument that the 'conventional algorithm' for setting a threshold results in a masking of the actual false positive rate. We show that for a small number of negative controls ($N \le 10$) the threshold is always larger than the largest negative control value.

For a set of negative controls $\{x_i\}_{i=1}^{N}$, assume

$$x_1 \le x_2 \le x_3 \le \ldots \le x_N \qquad [1.1]$$

where N is the number of negative controls.

Let $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the mean of the negative controls. Then

$$\mu \le \max_{i=1,\ldots,N}\{x_i\} = x_N \qquad [1.2]$$

$$\Rightarrow x_N - \mu \ge 0. \qquad [1.3]$$

Let $s^2 = \frac{1}{N-1}\sum (x_i - \mu)^2$ be the variance of the negative controls. Then

$$s^2 \ge \frac{1}{N-1}(x_N - \mu)^2 \qquad [1.4]$$

$$\Rightarrow s \ge \sqrt{\frac{1}{N-1}}(x_N - \mu). \qquad [1.5]$$

From inequality [1.5] and from the definition of the conventional threshold (where the $\text{Threshold} = \mu + 3s$) we see that

$$\text{Threshold} = \mu + 3s \geq \mu + 3\sqrt{\frac{1}{N-1}}(x_N - \mu) \qquad [1.6]$$

$$\text{Threshold} \geq \frac{3}{\sqrt{N-1}}x_N - \left(\frac{3}{\sqrt{N-1}} - 1\right)\mu + \mu. \qquad [1.7]$$

By adding and subtracting x_N to inequality [1.7] and factoring we get

$$\text{Threshold} \geq \left(\frac{3}{\sqrt{N-1}} - 1\right)(x_N - \mu) + x_N \qquad [1.8]$$

From equation [1.3] we know that $(x_N - \mu) \geq 0$, therefore if $\left(\frac{3}{\sqrt{N-1}} - 1\right) \geq 0$ then

$\text{Threshold} \geq x_N$. $\therefore$ If $N \leq 10$ then $\text{Threshold} \geq x_N = \max\{x_i\}$.