ABSTRACT

Aim  Many ecological surveys record only the presence or absence of species in the cells of a rectangular grid. Ecologists have investigated methods for using these data to predict the total abundance of a species from the number of grid cells in which the species is present. Our aim is to improve such predictions by taking account of the spatial pattern of occupied cells, in addition to the number of occupied cells.

Innovation  We extend existing prediction models to include a spatial clustering variable. The extended models can be viewed as combining two macroecological regularities, the abundance–occupancy regularity and a spatial clustering regularity. The models are estimated using data from five tropical forest censuses, including three Panamanian censuses (4, 6 and 50 ha), one Costa Rican census (16 ha) and one Puerto Rican census (16 ha). A serpentine grassland census (8 × 8 m) from northern California is also studied.

Main conclusions  Taking account of the spatial clustering of occupied cells improves abundance prediction from presence–absence data, reducing the mean square error of log-predictions by roughly 54% relative to a benchmark Poisson predictor and by roughly 34% relative to current prediction methods. The results have high statistical significance.

Keywords  Abundance prediction, abundance–occupancy, presence–absence, serpentine grassland, spatial autocorrelation, tropical forest.

INTRODUCTION

Species abundances are central to ecology and conservation planning. Here we study a problem in predicting abundances from incomplete information. Consider survey areas overlaid by rectangular grids. Some surveys record only the presence or absence of species in the cells of the grid. The problem is to predict the abundance of a species in the total area from its presence–absence grid. Predictors in the literature have exploited the cross-species relationship between total grid abundance and number of cells occupied. As reviewed in Holt et al. (2002), plots of abundance against occupancy exhibit a fairly regular pattern, suggesting that abundance could be predicted from occupancy when only presence–absence data are available. Holt and Gaston (2003) suggest that residuals from this abundance–occupancy relationship should be studied as a way of refining the relationship. We follow this advice by showing that prediction can be improved by exploiting the additional regularity that individuals of a species tend to be spatially autocorrelated (Brown, 1995; Ovaskainen and Hanski, 2001; Bahn and McGill, 2007).

Ecologists have studied abundance prediction from presence–absence data at least as far back as Gleason (1920). Predictors have been based largely or solely on the number of occupied cells (the abundance–occupancy relationship), placing little or no emphasis on the spatial clustering of those cells (the autocorrelation regularity). Examples are Gerrard and Chiang (1970), Nachman (1984), Kunin (1998), He and Gaston (2000), and Harte et al. (2005). Predictors have been tested with substantial success against plant and insect data, although Warren et al. (2003) and Tosh et al. (2004) had only limited success testing the fractal method from Kunin (1998) and a negative binomial method from He and Gaston (2000) against insect and large herbivore data.

Since presence–absence-only grids are less labour intensive to record than complete abundance grids, there are many of the former, e.g. for prairie vegetation (Steiger, 1930; Weaver and Albertson, 1943), calcicolous plants (Usher, 1975), heath vegetation (Ivimey-Cook et al., 1975), herbaceous plants (Thompson et al., 1998) and rice stem borers (Torii, 1971). A potentially important but more speculative application, noted by He and Gaston (2000), is to predict abundance from presence–absence grids in...
natural history atlases, such as for North American birds (Unitt, 2005; Cutright et al., 2006) or European mammals (Mitchell-Jones et al., 1999). However, larger areas may incorporate sufficient climate and surface-type heterogeneity to exceed a species’ habitat tolerances. For example, most tree species in the 50-ha Barro Colorado Island census area (used below) have habitat preferences which allow them to live nearly anywhere in the plot. In contrast, most tree or bird species in southern California have climate and surface-type preferences that restrict where they can live (illustrated for birds’ nests in Unitt, 2005). To predict the spatial distribution and abundance of species over large areas, habitat variables would probably need to be included in predictions.

For the census grid of a particular species let \( N \) denote total abundance, \( M \) the number of grid cells and \( m \) the fraction of cells occupied. Under the methods used in the citations above, \( N \) is predicted from \( m \) and \( M \) (sometimes also from \( m' \) and \( M' \), the values at some coarser resolution). Prediction is difficult because of spatial clustering, or aggregation, of individuals of a species. If we knew the average degree to which individuals cluster within occupied cells, we would know \( N/M \) and could calculate \( N \) exactly. Given only a presence–absence grid, however, there is no direct measure of this ‘within-cell clustering’. However, a grid may display spatial clustering of occupied cells (occupied cells near other occupied cells and empty cells near other empty cells), and this ‘cross-cell clustering’ may be an indicator of the missing information on within-cell clustering. Thus, it is hypothesized that a predictor of \( N \) based on a measure of cross-cell occupancy clustering, to be called \( C \), as well as on \( m \) and \( M \), will outperform standard predictors. Various definitions of \( C \) were considered, as detailed in Appendix S1 in Supporting Information. An adjacent-cell occupancy autocorrelation was selected as the definition of \( C \) on the basis of prediction pre-tests with simulated data and prediction tests with real data (Appendix S2). No assumption is made here about whether the autocorrelation is due to a species’ dispersal distance and density dependence (discussed in Molofsky et al., 2002), to landscape heterogeneity (discussed in Overton and Levin, 2003) or to other considerations.

Few methods exist for predicting a species’ abundance from a single presence–absence map for the species, given no other information. Instead, ‘training data’ (including abundance data) are needed to estimate a predictor formula, which can then be applied to the single map. Even when a predictor formula is derived purely from theoretical considerations (as in He and Gaston, 2000), training data are still needed to confirm the predictor’s accuracy.

Training data consist of a sample of data points \((N_i, m_i, M_i, C_i)\), where \( i \) indexes the different observations. Prediction models of the form \( \ln N_i = F(m_i, M_i, C_i) + u_i \) are estimated, where \( F(m_i, M_i, C_i) \) is the predictor of \( \ln N_i \) and \( u_i \) is the prediction error. The ideal training data would be gridded abundance data from multiple earlier surveys of the same species. Unfortunately, such data are almost never available (see two paragraphs below for an exception). Thus, researchers have turned to cross-species training data. The data points \((N_i, m_i, M_i, C_i)\) are for different species \( i \), which may or may not include the species for which prediction is to be made. Given the values \((m_o, M_o, C_o)\) for some new presence–absence grid, the log-abundance prediction for the species is \( F(m_o, M_o, C_o)\), evaluated for the \( F \) estimated from the training data.

In the simplest application of this paper, a user would accept one of the predictor formulae estimated below, and would plug in his or her own values of \((m_o, M_o, C_o)\), as taken from the presence–absence map of a single species. In a more complicated application, a user might rework the training data here, or a different body of training data, estimating his or her own prediction formula.

Although our approach is common statistical methodology – predicting the dependent variable for new values of the independent variables – it raises the ecological question of why prediction for a particular species can sensibly be based on data for other species. A way to look at this question is to think of abundance determination for various species as having some common characteristics and some species-specific characteristics. In \( F(m_o, M_o, C_o) \), the functional form \( F(.,.,) \), which has no \( i \) subscript, represents common characteristics, and the explanatory variables \((m_i, M_i, C_i)\) represent characteristics specific to species \( i \). The question then is whether \((m_i, M_i, C_i)\) capture enough species-specific information to allow reasonable prediction. The empirical answer is yes. Neglecting \( C_o \), scatter diagrams of cross-species data points \((N_i, m_i)\) for given \( M \) show a strong positive relationship between \( N_i \) and \( m_i \) demonstrating a sound reason for predicting abundance from occupancy (see He et al., 2002, Fig. 3; He and Gaston, 2003, Fig. 3; and Fig. 1 of this paper). Results here will show the added benefit of including \( C_i \).
Nachman (1984) illustrated the rare case when training data can be for the same species and site for which predictions are to be made. His grids were greenhouse plants, the cells were leaves and the species were two kinds of mite. A multi-survey design involved initial surveys of mite abundances, from which predictors were estimated, followed by further presence–absence surveys. He (2000) discusses prediction from mixed surveys in which abundance is counted for some cells and presence–absence for the rest. Repeat censuses of plants at a serpentine grassland site in northern California (Green et al., 2003) are being done for a larger area, but, for some of the most abundant species, only on a presence–absence basis (A. Smith, personal correspondence). Gerrard and Cook (1972) analysed prediction accuracy and cost trade-offs between full abundance and presence–absence-only surveys.

Prediction equations will be estimated in this paper. The training data are detailed census data on approximately 950 tropical tree grids at five sites. Some of these data will be held back for testing the estimated predictors, and a serpentine California grassland survey will also be used for testing. The tropical tree grids at five sites. Some of these data will be held back for testing the estimated predictors, and a serpentine California grassland survey will also be used for testing. The addition of C to prediction equations will reduce mean square prediction error by roughly 54% relative to a benchmark Poisson predictor, and by roughly 34% relative to prevailing methods. Results will be compared across sites and scales.

INNOVATION

Model development

Five tropical forest data sets (Barro Colorado Island, Luquillo, San Emilio, Sherman and Cocoli) and one serpentine grassland data set (northern California) were used, covering approximately 950 species, 90 ha and 325,000 individual plants. Each tropical forest plot was divided into two separate non-overlapping areas: an ‘in-sample’ area used to estimate prediction equations and an ‘out-of-sample’ area used to test the prediction equations. The entire serpentine grassland plot was treated as out-of-sample. Appendix S3 describes the data sets and their preparation.

For all but the serpentine grassland site, the data sets give point locations for each plant, implying that we could choose cell sizes small enough to yield perfect prediction. As the grid gets coarser, information is lost and m increases. When the extreme m = 1 is reached, prediction fails completely since there is no information on whether grid cells have few or many individuals. Whether due to large N or small M, larger abundance per cell, NM, makes prediction more difficult. Thus, an ecologist making predictions would use the finest scale (largest M) allowed by the data. For more discussion of how grid resolution affects abundance prediction see Kunin et al. (2000).

A scale coarse enough to present challenging prediction problems is needed to test our models. We chose to focus on 16 × 16 grids (M = 256 cells). Given the great variety of plant sizes and census area sizes across the data sets, 16 × 16 grids faced the prediction methods with a good variety of challenges, and the focus on a particular scale made results easier to present and absorb. Main results are given also for 8 × 8 and 4 × 4 grids.

Current methods for abundance prediction can be illustrated by the scatter of N, the variable to be predicted, against m, the explanatory variable, for a sample of species (see Fig. 1, based on the in-sample species). To form predictors, researchers have fitted curves through such scatters. Note that the points form a clean and tightly packed lower edge, spreading vertically to a ragged upper edge. The lower edge can be explained theoretically by a special case of the He and Gaston (2000) predictor:

\[ N^*(k) = M^k[(1 - m)^{-ak} - 1] \]

This equation for predicting N has optimal properties if grid cells are populated as independent draws from a negative binomial distribution with clustering parameter k (smaller k indicating more within-cell clustering). In fact, empirical cell abundances are typically autocorrelated, depending on the area size, grid resolution and other features of the data set. Hence equation 1 is to be viewed as a starting approximation for prediction, with k an adjustable parameter to be estimated from the data. Note that \( N^*(k) \) is a function of M and m as well as k, despite the streamlined notation \( N^*(k) \).

The limit as k → ∞, namely \( N^*(\infty) = -M \ln(1 - m) \), defines the Poisson case of purely random spatial distribution. The grey line on the figure, which neatly captures the lower edge of the scatter, is the Poisson equation. Similar figures appear in He et al. (2002, Fig. 3) and He and Gaston (2003, Fig. 3). From here on, it will be convenient to express abundances in log form (\( \ln N \)). The Poisson log-predictor \( \ln N^*(\infty) \) will be used as a benchmark for comparisons.

The upward scatter of points from the grey line indicates clustering. For a given species distributed over a grid, clustering causes occupied cells to have, on average, more individuals per cell than under spatial randomness. For grid cell data, clustering can manifest itself in two ways. Within-cell clustering is the tendency for individuals to cluster into fewer cells. Cross-cell clustering is the tendency for well-populated cells to cluster near each other on the grid. The mix between cross-cell and within-cell clustering changes with grid scale; a coarser grid’s within-cell clustering becomes a finer grid’s cross-cell clustering.

He and Gaston (2000) developed predictors based on negative binomial methods. Useful extensions are implicit in the discussions of occupancy–abundance relationships in He et al. (2002) and Holt et al. (2002). In effect, these authors move from the Poisson predictor \( \ln N^{(*)} \), allowing zero clustering, to the negative binomial predictor \( \ln N^*(k) \), allowing within-cell clustering, to an extended version \( a + b \ln N^*(k) \), where a and b as well as k are adjustable parameters. As they show, this formulation includes various special cases from the literature: logistic, power, Nachman, Poisson and negative binomial. (The expression \( a + b \ln N^*(k) \) is the prediction analogue of the logarithm of equation 18 in Holt et al., 2002, where their \( \alpha \) and \( \beta \), along with k and M, are embedded in a and b.)

The novelty in this paper is the incorporation of a cross-cell clustering metric C. Four prediction equations are compared. For each, u is an error term. The predictor itself is the right-hand side without the u.
Poisson predictor \[ \ln N = \ln N^* (\infty) + u \] (2)

Within-cell predictor \[ \ln N = a + b \ln N^*(k) + u \] (3)

Cross-cell predictor \[ \ln N = \ln N^* (\infty) + cC + u \] (4)

Combined predictor \[ \ln N = a + b \ln N^*(k) + cC + u \]. (5)

These predictors will guide our judgement about the importance of cross-cell clustering. The Poisson predictor (equation 2), with no adjustable parameters, is the benchmark. The within-cell predictor (equation 3), with three adjustable parameters, represents a variety of abundance–occupancy relations in the literature described in the Holt et al. (2002) review. The cross-cell predictor (equation 4), with the one adjustable parameter \( c \), is the simplest predictor which includes cross-cell clustering; to our knowledge, it is new. If it can outperform the within-cell predictor, then cross-cell clustering must have predictive importance. The combined predictor (equation 5), with four adjustable parameters, takes the first three predictors as special cases.

Other prediction equations were considered, but none were superior. For example, equations with more terms were considered (e.g. using multiple definitions of \( C \) together), but these added little to accuracy and conflicted with simplicity (equations 13–15 below suggest a possible exception). Simplicity is desirable because users must be comfortable that a predictor fit to particular data sets has legitimacy when extrapolated to other data sets. Complicated equations raise more doubts. Equations were considered with \( N \) (out of logs) as a linear function of \( N^*(k) \) and of a term involving \( C \) (either by itself or interacting with \( N^*(k) \)). Species with low abundances (for example, \( N = N^*(k) = 1 \)) fix the left end of the \( N \) against \( N^*(k) \) curve near the origin, leaving the relatively few high-abundance species at the right end to dominate the estimate of the slope. However, high-abundance species need not be of dominating importance to abundance prediction, since researchers will often have the most information about the most common species. Logging \( N \) and \( N^*(k) \) mitigates this dominance. It also mitigates error term heteroskedasticity.

In principle, questions of variable definition and predictor specification could be evaded by deriving a maximum likelihood estimator of abundance. Appendix S4 explains why this approach is intractable.

Parameters in equations 3–5 are estimated by least squares regression using training data on \( (N, m, M, C) \) from the in-sample species. Since the parameter \( k \) enters \( \ln N^*(k) \) nonlinearly in equations 3 and 5, and since heteroskedasticity of error terms is apparent, nonlinear least squares with heteroskedasticity correction is used. In particular, the commercial application RATS (Estima, 2003, Version 5) is used.

The accuracy of each prediction equation is measured by the familiar regression \( s^2 \):

\[ s^2 = \frac{\sum_{\text{species}} (\ln N - \text{predicted } \ln N)^2}{\text{no. of observations} - \text{no. of parameters}}. \] (6)

This measure is readily interpretable as a mean square error, and is a common model-ranking criterion since it rewards predictors with fewer parameters. Other model-ranking criteria, such as the Akaike information criterion (AIC), could be used. However, when the number of parameters is small relative to the number of observations, as here, standard criteria give the same, or nearly the same, rankings. For example, in the many model rankings below, there are only three small reversals between \( s^2 \) and AIC rankings (noted under Tables 1 & 2). The reason is that AIC is related to \( s^2 \) by \( \text{AIC} = \ln(s^2) + \ln(1 - r) + 2r \), where \( n \) denotes the number of observations and \( r \) is the ratio of parameters to \( n \). For small \( r \), AIC/n \( \approx \ln(s^2) \), leading to the same rankings.

For each prediction equation, the standard \( R^2 \) will be reported, defined by \( R^2 = 1 - [\bar{s}/\text{var} (\ln N)] \), where \( \text{var}(\ln N) \) is the ‘unbiased’ sample variance of the log-abundance being predicted. Though \( R^2 \)’s are reported, we emphasize that \( R^2 \) is not a measure of predictive accuracy since it only reveals the ratio of \( s^2 \) to \( \text{var}(\ln N) \), not \( s^2 \) itself. High variability in \( \ln N \) can lead to an \( R^2 \) near 1 despite high prediction error \( s^2 \). The scatter (Fig. 1) illustrates this. For mid-range values of \( m \), the vertical variability of points is large, indicating substantial prediction error, even though the scatter as a whole has high \( R^2 \).

The cross-cell clustering metric \( C \) used in the statistical fits below is the standard Moran autocorrelation for adjacent cells (Legendre & Legendre, 1998, Chapter 13). To define it, let the grid cells be labelled in some order 1, 2, ..., \( M \), and let \( m \) be a binary variable equal to 1 if cell \( i \) is occupied and 0 otherwise. The average of the \( m_i \) over cells equals the fraction of cells occupied, which is \( m \). The Moran autocorrelation is

\[ C = \frac{S^2 \sum_{i \neq j} (m_i - m)(m_j - m)}{M^2 \sum (m_i - m)^2} \] (7)

In the denominator, \( \Sigma \) denotes the sum over individual cells \( i \). In the numerator, \( \Sigma_{i \neq j} \) denotes the sum over pairs of cells \((i,j)\) for which cell \( i \) is horizontally or vertically adjacent to cell \( j \). \( S \) denotes the number of such pairs; if the grid is \( R \times D \), then \( S = R(D-1) + D(R-1) \). Definition (7) gives equal weight to all adjacent cell pairs and zero weight to all other pairs. To give a sense of the variable \( C \), Fig. 2 plots \( C \) against \( \ln N \) for the 955 in-sample observations.

In addition to \( C \), several other clustering metrics (based on scale, commonality and related considerations) were investigated, as detailed in Appendix S2. Rather than use only our limited real data in choosing among metrics, which would risk a data-dredging bias, we tested the metrics using two sets of 5000 simulated landscapes, each landscape representing a species. The first set of 5000 was based on the ‘quadrisection’ spatial model from Conlisk et al. (2007), and the second set of 5000 was based on the ‘Poisson cluster’ model of Plotkin et al. (2000). These models and simulations are described in Appendix S2. Equations 2–5 were estimated for each clustering metric for each simulated data set, and also for the real in-sample data set. The Moran metric \( C \) performed as well as or better than the other metrics in reducing \( s^2 \) on moving from equations 2 and 3 to equations 4 and 5. Estimates of equations 2–5 for the simulated data sets were broadly consistent with estimates for the real data (presented next), providing some confidence in the two simulation models.
Predicting abundance from presence–absence

Table 1 Estimated predictors by site. For each panel of the table (each site), the results for the four predictors (Poisson, within-cell, cross-cell, combined) are site-specific estimates of the four models defined by text equations 2–5. Data are for the ‘in-sample’ areas, as described in the text and detailed in Appendix S3. In the last two columns, the values of the ‘$s^2$-test’ and ‘$s^2$-cells’ are computed using parameters from the all-sites regression equations 8–10 and equations 11–13, respectively. For the San Emilio site, the near-tie in model performance between the cross-cell and combined models would reverse direction if the Akaike criterion was used instead of the $s^2$ criterion. Otherwise, the two criteria rank models the same. Numbers in parentheses next to coefficients are their estimated standard deviations.

<table>
<thead>
<tr>
<th>Site</th>
<th>Predictor</th>
<th>k</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>$R^2$</th>
<th>$s^2$</th>
<th>$s^2$-test</th>
<th>$s^2$-cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barro Colorado Island</td>
<td>Poisson</td>
<td>13.2</td>
<td>0.140</td>
<td>1.049</td>
<td>0.019</td>
<td>0.920</td>
<td>0.3254</td>
<td>0.3254</td>
<td>0.3254</td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>2.376</td>
<td>0.241</td>
<td>0.967</td>
<td>0.1347</td>
<td>0.1382</td>
<td>0.1363</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>2.150</td>
<td>0.410</td>
<td>0.968</td>
<td>0.1285</td>
<td>0.1378</td>
<td>0.1352</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1.029</td>
<td>0.028</td>
<td>0.964</td>
<td>0.1955</td>
<td>0.2577</td>
<td>0.2356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luquillo n = 121</td>
<td>Poisson</td>
<td>0.906</td>
<td>0.5065</td>
<td>0.5065</td>
<td>0.5065</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>3.197</td>
<td>0.255</td>
<td>0.966</td>
<td>0.1838</td>
<td>0.2257</td>
<td>0.2203</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>0.970</td>
<td>0.031</td>
<td>0.980</td>
<td>0.1074</td>
<td>0.1634</td>
<td>0.1660</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1.398</td>
<td>0.176</td>
<td>0.986</td>
<td>0.0403</td>
<td>0.0569</td>
<td>0.0475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Emilio n = 131</td>
<td>Poisson</td>
<td>0.959</td>
<td>0.1152</td>
<td>0.1152</td>
<td>0.1152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>2.5</td>
<td>0.014</td>
<td>1.065</td>
<td>0.012</td>
<td>0.0754</td>
<td>0.0727</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>1.364</td>
<td>0.292</td>
<td>0.986</td>
<td>0.0398</td>
<td>0.0497</td>
<td>0.0443</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>6.1</td>
<td>0.011</td>
<td>0.992</td>
<td>0.008</td>
<td>0.0143</td>
<td>0.0157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sherman n = 193</td>
<td>Poisson</td>
<td>1.6</td>
<td>0.045</td>
<td>0.998</td>
<td>0.010</td>
<td>0.0281</td>
<td>0.0281</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>1.024</td>
<td>0.122</td>
<td>0.994</td>
<td>0.0165</td>
<td>0.0284</td>
<td>0.0182</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>0.988</td>
<td>0.0344</td>
<td>0.993</td>
<td>0.0262</td>
<td>0.0262</td>
<td>0.0265</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>2.284</td>
<td>0.030</td>
<td>0.992</td>
<td>0.008</td>
<td>0.0146</td>
<td>0.0153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cocoli n = 237</td>
<td>Poisson</td>
<td>6.1</td>
<td>0.011</td>
<td>1.034</td>
<td>0.011</td>
<td>0.0260</td>
<td>0.0260</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>1.339</td>
<td>0.160</td>
<td>0.993</td>
<td>0.0146</td>
<td>0.0202</td>
<td>0.0153</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>1.330</td>
<td>0.208</td>
<td>0.994</td>
<td>0.0143</td>
<td>0.0185</td>
<td>0.0157</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Out-of-sample prediction results. Each cell gives the mean square prediction error $s^2$ for the corresponding predictor (column) and site (row). The four predictors are as defined by text equations 2–5. Data are for ‘out-of-sample’ areas, as described in the text and detailed in Appendix S3. For the two serpentine rows, with their small n values, the near tie in model performance between the cross-cell and combined models would reverse direction if the Akaike criterion was used instead of the $s^2$ criterion. Otherwise, the two criteria rank models the same.

<table>
<thead>
<tr>
<th>Site</th>
<th>Predictor</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poisson</td>
<td></td>
</tr>
<tr>
<td>All (n = 912)</td>
<td>0.2660</td>
<td></td>
</tr>
<tr>
<td>Barro Colorado Island (n = 269)</td>
<td>0.2559</td>
<td></td>
</tr>
<tr>
<td>Luquillo (n = 116)</td>
<td>0.3866</td>
<td></td>
</tr>
<tr>
<td>San Emilio (n = 138)</td>
<td>0.0691</td>
<td></td>
</tr>
<tr>
<td>Sherman (n = 202)</td>
<td>0.0248</td>
<td></td>
</tr>
<tr>
<td>Cocoli (n = 137)</td>
<td>0.0164</td>
<td></td>
</tr>
<tr>
<td>Serpentine 1998 (n = 23)</td>
<td>2.2849</td>
<td></td>
</tr>
<tr>
<td>Serpentine 2005 (n = 27)</td>
<td>2.2009</td>
<td></td>
</tr>
</tbody>
</table>

Model testing

First consider estimates of the four predictors using the 955 in-sample observations on tropical tree species (as described in Appendix S3). Grids are 16 × 16 (M = 256). Since prediction fails when m is close to 1, only species with m < 0.98 were included. The prediction mean square error for the benchmark Poisson predictor $\ln N = \ln N^*(\infty)$ is $s^2 = 0.1872$ (with $R^2 = 0.9538$). The regression estimates for the within-cell, cross-cell and combined predictors are:

$$\ln N = 0.049 + 1.047 \ln N^*(3.24) \quad (s^2 = 0.1151, R^2 = 0.972) \quad (8)$$

$$\ln N = \ln N^*(\infty) + 2.052C \quad (s^2 = 0.0863, R^2 = 0.979) \quad (9)$$

$$\ln N = 0.071 + 0.969 \ln N^*(2.10) + 1.841C \quad (0.016) (0.010) (0.338) (0.193) \quad (10)$$

$$\ln N = 0.0754, R^2 = 0.981.$$ 

Here and below, numbers in parentheses under regression coefficients are their estimated standard deviations. The conventional within-cell predictor (equation 8) performs well. It has a prediction error variance ($s^2 = 0.1150$) that is 39% less than the benchmark Poisson value ($s^2 = 0.1872$). The two predictors involving C perform better. The error variances of the cross-cell

© 2008 The Authors

Global Ecology and Biogeography, 18, 1–10, Journal compilation © 2008 Blackwell Publishing Ltd
predictor \( (s^2 = 0.0863) \) and the combined predictor \( (s^2 = 0.0754) \) are 54% and 60% less than the benchmark Poisson value, and 25% and 34% less than the standard within-cell predictor, respectively. A simple description of the cross-cell predictor (equation 9) is ‘Poisson plus twice the autocorrelation’. The variable \( C \) has high statistical significance in both equations 9 and 10. Thus, the data confirm the main hypothesis of the paper. A cross-cell clustering variable improves abundance prediction.

Since prediction is difficult at high \( m \) values, heteroskedasticity of errors was expected. For the combined predictor (equation 10), the residual variance for observations with \( m > 0.8 \) is more than three times the overall residual variance \( (s^2 = 0.0754) \). Therefore, all regressions are computed with a heteroskedasticity-robust technique.

A visual sense of the improvement in prediction gained by adding \( C \) is given by Fig. 3(a) and (b). They plot predicted versus observed log-abundances (a) for the within-cell model of equation 8 and (b) for the combined model of equation 10. The points cluster more closely about the 45° line for the latter plot than for the former plot, hence the 34% drop in \( s^2 \) between the two equations.

For the combined predictor, we expected the intercept parameter \( a \) to be approximately zero. When only one cell is occupied \( (m = 1/M) \), \( N^*(k) = 1 \) for any \( k \); the data usually show \( N = 1 \); and \( C \) will be close to 0. Substituting these values \( (N, N^*(k), C) = (1, 1, 0) \) in the combined prediction equation (5) yields \( a = 0 \). We also expected the slope parameter \( b \) to be approximately 1. Since \( \ln N^*(k) \) is a predictor of \( \ln N \), a unit increase in \( \ln N^*(k) \) should produce approximately a unit increase in \( \ln N \). In fact, \( a \) and \( b \) were estimated near 0 and 1 in equation 10.

To see whether inclusion of highly abundant species in the estimation distorted predictions for the broad intermediate range of \( m \) values, the three regressions 8–10 were re-estimated with the \( m \) cut-off reduced from \( m < 0.98 \) to \( m < 0.8 \). The number of observations went from 955 to 919. These new regressions had very similar coefficients and improvement in \( s^2 \) values on moving through the models (hence the results are not presented).

The prediction equations 8–10 were recalculated for the five sites separately (Table 1). For each site, the variable \( C \) is significant, and \( s^2 \) declines from the within-cell to the cross-cell to the combined predictor. The differences in \( s^2 \) across sites probably reflect differences across sites in average \( m \). However, parameter estimates for \( c \) and \( k \) are somewhat unstable across sites. This is potentially troublesome to users who must predict abundance at a new site. The seriousness of the problem depends on the

Figure 2  Plot of the adjacent-cell autocorrelation \( C \) against \( \ln N \) for the 955 in-sample species. The mean of \( C \) is 0.090, the standard deviation 0.126 and the skewness 1.430.

Figure 3  Plots of observed against predicted \( \ln N \) for: (a) the within-cell model, which excludes the adjacent-cell autocorrelation \( C \), and (b) the combined model, which includes \( C \). The points cluster more closely about the 45° line for plot (b) \( (s^2 = 0.0754) \) than for plot (a) \( (s^2 = 0.1151) \), indicating the usefulness of the cross-cell clustering measure \( C \) as a predictor variable.
sensitivity of the prediction error $s^2$ to parameter instability. If $s^2$ is a flat function of parameters in the vicinity of the $s^2$-minimizing values, then a non-minimizing parameter selection may yield good predictions. To check this, the ‘$s^2$-test’ column of the table presents $s^2$ values computed using parameter values from the all-sites regressions 8–10. Comparison of the ‘$s^2$’ and ‘$s^2$-test’ columns shows how much the error measure $s^2$ increases in moving from site-specific parameter values to all-sites values. For sites other than Sherman, predictions based on the all-sites parameters still provide substantial improvement over the Poisson benchmark. Thus, parameter instability appears to be troublesome but not fatal. The most unstable parameter in Table 1 is $k$, but this is misleading since $\ln N^*(k)$ is not very sensitive to $k$. For example, the correlation coefficient between $\ln N^*(1)$ and $\ln N^*(\infty)$ over the 955 in-sample observations is 0.9985. Instability of $c$ is the bigger problem. (The last column of Table 1 is discussed shortly.)

Addition of site-descriptive variables might mitigate parameter instability. The site-specific $C$ coefficients in Table 1 are roughly in order of the site cell sizes (see Appendix S3), suggesting that the prediction equations should include an interaction between $C$ and a cell-size variable. Let $H$ denote cell size in hectares per cell. The following regressions add the variable $\ln H$ to the within-cell, cross-cell and combined predictors:

$$\ln N = 0.323 + 1.034 \ln N^*(3.6) + 0.059 \ln H$$  \hspace{1cm} (11)

$$s^2 = 0.1104, R^2 = 0.973$$

$$\ln N = \ln N^*(\infty) + (3.610 + 0.452 \ln H)C$$  \hspace{1cm} (12)

$$s^2 = 0.0801, R^2 = 0.980$$

$$\ln N = (0.074) + 0.970 \ln N^*(2.3) + (3.088 + 0.357 \ln H)C$$  \hspace{1cm} (13)

$$s^2 = 0.0717, R^2 = 0.982.$$  

Although the addition of $\ln H$ lowers $s^2$ only slightly relative to the original three regressions 8–10, its coefficient in each equation is statistically significant by a standard $t$-test. More important, the inclusion of $\ln H$ slightly mitigates the problem of parameter instability across sites. The last column of Table 1 shows $s^2$ values by site, labelled ‘$s^2$-cells’, when predictions are based on equations 11–13. The inclusion of $\ln H$ brings predictor quality closer to the quality of site-specific parameters, especially for the Sherman site. However, since we only added $\ln H$ after seeing Table 1, there is a flavour of data-dredging here. More important, a proper treatment of cell size effects would consider the size of the species as well as the size of the cell. The data sets include information on diameter at breast height (d.b.h.) for each individual, from which average d.b.h. for each species can be estimated. However, experiments with average d.b.h. as a predictor variable were disappointing. It reduced prediction error only very slightly and was typically not statistically significant. Nonetheless, a better measure of species size might provide better results.

To investigate scale effects, predictors were estimated for $8 \times 8$ and $4 \times 4$ grids as well as the $16 \times 16$ grid. Again the $m < 0.98$ cut-off was used, resulting in fewer observations (907 and 773 for the $8 \times 8$ and $4 \times 4$ grids, respectively) since $m$ runs larger at coarser scales. Table 3 compares $16 \times 16, 8 \times 8,$ and $4 \times 4$ results. Since coarser scales carry less information, $s^2$ values increase substantially, roughly doubling from the $16 \times 16$ scale to the $8 \times 8$ scale, and roughly doubling again from the $8 \times 8$ scale to the $4 \times 4$ scale. Since coarser scales have larger cells, some of the cross-cell clustering at a finer scale becomes within-cell clustering at a coarser scale. As a result, $C$ declines in importance as a predictor variable; in particular, the estimates of the coefficient $c$ get smaller as the scale gets coarser (see the table). Nonetheless, $C$ maintains statistical significance even at the $4 \times 4$ scale. For a check on the importance of coefficient changes across scales, let us focus on the combined prediction model. When the estimated coefficients ($k, a, b, c$) for the $16 \times 16$ scale are applied to the

<table>
<thead>
<tr>
<th>Scale</th>
<th>Predictor</th>
<th>$k$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$R^2$</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$16 \times 16$</td>
<td>Poisson</td>
<td>3.24</td>
<td>0.049</td>
<td>1.047</td>
<td>0.954</td>
<td>0.1872</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>0.087</td>
<td>0.017</td>
<td>0.008</td>
<td>0.972</td>
<td>0.1150</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>2.10</td>
<td>0.071</td>
<td>0.969</td>
<td>2.052</td>
<td>0.0863</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>0.338</td>
<td>0.026</td>
<td>1.103</td>
<td>2.041</td>
<td>0.0754</td>
<td></td>
</tr>
<tr>
<td>$8 \times 8$</td>
<td>Poisson</td>
<td>0.061</td>
<td>0.015</td>
<td>0.010</td>
<td>0.981</td>
<td>0.3449</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>0.061</td>
<td>0.026</td>
<td>1.103</td>
<td>0.939</td>
<td>0.2038</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>0.086</td>
<td>0.025</td>
<td>1.111</td>
<td>2.141</td>
<td>0.2035</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>0.74</td>
<td>0.045</td>
<td>1.153</td>
<td>1.4790</td>
<td>0.1629</td>
<td></td>
</tr>
<tr>
<td>$4 \times 4$</td>
<td>Poisson</td>
<td>0.140</td>
<td>0.041</td>
<td>0.040</td>
<td>0.936</td>
<td>0.3634</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within-cell</td>
<td>0.140</td>
<td>0.045</td>
<td>1.064</td>
<td>0.936</td>
<td>0.2772</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cross-cell</td>
<td>0.185</td>
<td>0.045</td>
<td>0.936</td>
<td>0.853</td>
<td>0.5656</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>0.73</td>
<td>0.045</td>
<td>0.936</td>
<td>0.853</td>
<td>0.3634</td>
<td></td>
</tr>
</tbody>
</table>

© 2008 The Authors

*Global Ecology and Biogeography, 18, 1–10, Journal compilation © 2008 Blackwell Publishing Ltd*
16 × 16 values of \((m, M, C)\), the error measure is \(s^2 = 0.0754\). When the estimated coefficients \((k, a, b, c)\) for the coarser 8 × 8 scale are applied to the same 16 × 16 values of \((m, M, C)\), the error measure rises only to \(s^2 = 0.0840\). This suggests that a predictor estimated from data for a coarse scale may be reasonably accurate in prediction for a fine-scale. Put differently, there can be substantial cross-scale robustness in an estimated predictor. However, when the estimated coefficients \((k, a, b, c)\) for the still coarser 4 × 4 scale are applied to the same 16 × 16 values of \((m, M, C)\), the error measure jumps to \(s^2 = 0.1616\), illustrating that, at some degree of coarseness, cross-scale robustness must start to break down. Even so, the corresponding error calculation for the within-scale predictor \((4 × 4\) coefficients applied to \(16 × 16\) m and \(M\) values\) is \(s^2 = 0.2523\), suggesting that the inclusion of \(C\) in predictors may be useful even after cross-scale robustness has started to break down.

As detailed in Appendix S3, roughly half the data were held back for out-of-sample tests. A sixth site was added, a serpentine grassland site for northern California, with a square 16 × 16 grid. For all six sites, only species with \(m < 0.98\) were included. The serpentine species are typically annuals, with rapid turnover of plants, justifying the inclusion of two censuses (1998 and 2005). To see how well the in-sample predictors (equations 8–10) perform in predicting out-of-sample abundances, \(s^2\) values were computed (Table 2). For all out-of-sample observations pooled (912 species), the pattern of predictor accuracy agrees well with in-sample results. The predictors in order of decreasing \(s^2\) are Poisson, within-cell, cross-cell and combined. This ranking of models also applies to the BCI, San Emilio, Sherman and serpentine sites individually, and it almost applies to Luquillo (the exception is that the within-cell predictor outperforms the cross-cell predictor). The serpentine site has the largest percentage decline in \(s^2\) between the Poisson and combined predictors. For the Cocoli site, however, the pattern fails. The Cocoli predictions are typically too large, leaving the naive Poisson predictor as the best predictor. Since Cocoli has smaller cells than the sites above it on Table 2, this may be a cell-size issue. Overall, the out-of-sample results reinforce the in-sample results.

### DISCUSSION

Abundance information helps to determine at-risk populations, guide conservation decisions and further knowledge of species natural histories. In the literature, abundance prediction from gridded presence–absence data is usually based on a single predictor variable, the number of occupied cells. It was hypothesized that the addition of a second predictor variable – the adjacent-cell autocorrelation of binary occupancy variables – would improve predictor accuracy. A larger autocorrelation indicates a higher degree of spatial clustering for the species in question; and more clustering, leading to more individuals per occupied cell on average, predicts greater abundance. Since spatial clustering is common, the new predictor variable has potential relevance to a variety of species, adding to the existing macroecological literature on abundance–occupancy relationships.

Four predictors were emphasized. (1) The ‘benchmark Poisson predictor,’ which assumes pure spatial randomness and thus no clustering. (2) A ‘within-cell predictor’ which extends the Poisson model to allow for within-cell (but not cross-cell) clustering. This predictor, based on the work of Nachman (1984), He and Gaston (2000) and others, represents a prevailing approach in the literature. (3) A ‘cross-cell predictor’ which extends the Poisson model to include cross-cell clustering as represented by the autocorrelation variable. (4) A ‘combined predictor’ which includes both within-cell and cross-cell clustering.

The autocorrelation variable did improve prediction. The primary estimates (equations 8–10) were based on 955 tree species (the in-sample data) over five sites using a 16 × 16 grid. Relative to the Poisson benchmark the prevailing within-cell predictor reduced the mean square error of log-prediction \(s^2\) by 39% using three parameters; the new cross-cell predictor reduced \(s^2\) by 54% using one parameter; and the combined predictor reduced \(s^2\) by 60% using four parameters. Relative to the prevailing within-cell predictor the addition of the cross-cell autocorrelation \(C\) reduced \(s^2\) by 34%. These \(s^2\) reductions were highly significant statistically.

Estimated predictor parameters were relatively stable over wet and dry tropical sites with very different cell sizes (156.25 m² for the Cocoli 16 × 16 grid to 15,625 m² for the BCI 4 × 4 grid); and the value of \(s^2\) was fairly insensitive to moderate changes in parameters (Table 1). Estimated parameters were also relatively stable under major changes in scale (Table 3). These facts suggest fairly good predictor robustness. The out-of-sample tests were supportive. In particular, the serpentine site, which was not used in the predictor estimation, is a good robustness check. It has a different climate (northern California), much smaller species (grasses and forbs), much smaller grid cells (0.25 m² for the 16 × 16 grid) and much larger \(s^2\) values (perhaps due to the small grid cells, or to heavy clustering of individuals in one fertile corner of the plot, or to an abundance distribution skewed by a few highly abundant species). Despite these major differences, the serpentine site showed the largest proportional \(s^2\) reductions in moving from the Poisson through the other predictors.

If a user has the presence–absence grid for a species, prediction equations 8–10 can be used, as we did for the serpentine site. The training data did not include serpentine species, but the predictions were quite accurate. The Poisson predictor is a good lower bound (recall Fig. 1), and our results suggest a high probability that the within-cell, cross-cell and combined predictors will improve on the Poisson benchmark. If one of the sites for our calculations seems a better match to the user’s site, prediction can be based on the site-specific parameters of Table 1. If a user has training data more relevant than the training data used here, new predictors can be estimated. For example, a user might do repeat surveys. The first survey could be a costly full abundance census from which a predictor is estimated. Successive surveys could be less costly presence–absence surveys, possibly for larger areas.

A problem is that the detailed abundance censuses used as training data, due to their expense, cover relatively small areas, perhaps limiting the applicability of the estimated predictors. Further work studying the interactions among species size, cell
size and area size might give a better sense of the range of applicability of the methods here (equations 11–13 are a small start). Atlas data illustrate the cell size problem in extreme form. The San Diego County Bird Atlas (Unitt, 2005) has a fairly fine grid by atlas standards, yet each individual cell is approximately 50 times the entire area of the largest site we use (BCI).

In conclusion, there is substantial empirical support for the hypothesis that a measure of the clustering of occupied cells can substantially improve abundance prediction from presence–absence data, making these data more useful to ecologists. This conclusion improves our understanding of the occupancy–abundance relationship.

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of many individuals and organizations in gathering the data sets used here and for granting permission to use them: the Center for Tropical Forest Science of the Smithsonian Tropical Research Institute and their supporters, the National Science Foundation and the MacArthur Foundation, for the Barro Colorado Island, Sherman, and Cocoli data sets; the Luquillo Long-Term Ecological Research Program and their contributors, the National Science Foundation, the University of Puerto Rico, and the International Institute of Tropical Forestry, for the Luquillo data set; the Area de Conservacion Guanacaste, Costa Rica, the site of the San Emilio data set; and Jessica Green and Adam Smith for permission to use the serpentine data set. We also thank the National Science Foundation for direct financial support. Finally, we thank Craig Moritz, Perry de Valpine, referees and editors for many helpful comments and suggestions.

REFERENCES


**SUPPORTING INFORMATION**

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

- **Appendix S1** Cross-cell clustering metrics.
- **Appendix S2** Pre-testing the metrics.
- **Appendix S3** Data preparation.
- **Appendix S4** Intractability of maximum-likelihood prediction.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

**BIOSKETCHES**

**Erin Conlisk**, a recent PhD from the Energy and Resources Group at the University of California, Berkeley, is an associate scientist at the Pesticide Research Institute in Berkeley, California. She studies the spatial patterns of plants and the environmental hazards of pesticides.

**John Conlisk** is Professor Emeritus in Economics at the University of California, San Diego.

**Brian Enquist** is Professor in Ecology and Evolutionary Biology at the University of Arizona, Tucson. He investigates how functional and physical constraints on individuals influence larger scale ecological and evolutionary patterns.

**Jill Thompson** is an associate scientist at the University of Puerto Rico working on the Luquillo Long Term Ecological Research Project. She studies the effects of human land use and natural disturbances on population dynamics and community structure.

**John Harte** is a professor in the Energy and Resources Group at the University of California, Berkeley. He has published widely on biogeochemistry, climate change, environmental pollution, and macroecology.

Editor: Jack Lennon