The importance of accounting for overdispersion in site-occupancy estimations

By Adam Eric Miller

Colorado State University

Abstract

Understanding species distributions across landscapes can help set management goals to protect vital areas for species persistence. Areas of high species occurrence may represent a crucial habitat that should be protected against anthropogenic and environmental impacts. Recently, a new model has been developed to estimate the probability that a site is occupied, given cryptic individuals that are imperfectly detected, allowing experts to infer how habitat use varies across the landscape. However, ecologists and conservationists often make the mistake of trying to choose the model that best fits the data, which may not necessarily accurately model the environment being studied. In this study, we review occupancy modeling, and techniques to account for overdispersion in data sets, and then examine a data set on Northern Colorado birds to test its importance. We found that the estimates of the overdispersion parameter varied from species to species. Not accounting for overdispersion in some species caused the bias of multi-model inference, and could lead to misleading recommendations by conservationists. As environmental agencies face limited resources, drawing inference from poorly fitting misleading models could be extremely costly. We demonstrate the importance of assessing overdispersion in site-occupancy models, as it can change the inference made and the knowledge drawn from monitoring.

Introduction

In the face of many potential negative impacts on wildlife populations, understanding species distributions across landscapes has become an important consideration in many ecological studies.[1,2] Because some species and individuals are cryptic (hard to see or find), recent advancements in wildlife research have emphasized the importance of accounting for detection probability in estimating population and landscape level parameters.[3,4] Tied with this advancement in ecological statistics, is the use of presence/absence data, which is often less time and cost intensive than count and mark/re-capture studies to estimate population abundance and occurrence.[4]

Recent population models have been developed that take into account imperfect detection probabilities ($p<1.0$) while estimating the probability that a site is occupied by a species.[3,5] Although similar models have been developed in the past, the model developed by MacKenzie et. al. has proven to be more robust, and has become widely used in monitoring programs, ecological studies, and conservation biology.[3,6,7]

MacKenzie et. al. use a sampling design where a certain number of sites (N) are sampled a specific number of times (T). Researchers visit the sites multiple times (T), and detect species using various methods (e.g., point counts, traps). Then for each site (i) a detection history can be observed (Xi) and a combination of site and sample level covariates can be used to explain variation.[3,5]

In basic occupancy models, the two main parameters of interest are the proportions of N sites occupied (Ψ) and the probability of detecting species Z (p). A detection history (Xi) can then be used to estimate a site specific Ψ at site i. For example, if site i was sampled six times, and species Z at site i had a detection history of "001100". This denotes that the species was not detected at the site during the first two or last two sampling occasions, but was detected in the third and fourth sampling occasion. The probability of observing the above outcome can be described as:

(1.0)

$$\Pr(X_1=001100)=\Psi_1(1-p_1)(1-p_2)(p_3)(p_4)(1-p_5)(1-p_6)$$

A history of all 1s denotes that a species was detected at each sampling occasion, while a detection history of all 0s denotes a history where the species was never detected. Given the number of sampling occasions, there are a variety of combinations of detections and non-detections at a site. This can be described as:

(2.0)

$$L(X_1, X_2, \ldots X_n|\psi, p) = \prod_{i=1}^{N} \Pr(X_i)$$

This likelihood can then be used to calculate the maximum likelihood estimates for model parameters of interests (i.e., site occupancy). However, MacKenzie et. al. express occupancy and detection probabilities based on various site and sample level covariates (x).[3] Therefore, a logistic regression is used to estimate the parameter of interests (θi) by the following equation:

(3.0)

$$\theta = \frac{e^{(B_0+B_1x)}}{1+e^{(B_0+B_1x)}}$$

The goal of using models to estimate individual, population, and community-level parameters is to fit a set of models that represents the researchers' hypothesis to a given data set to choose the most accurate model(s).[8] A popular approach in ecological studies is to use model selection techniques such as Akaike's Information Criterion (AIC) to choose the model that best fits the data.[8] However, an important assumption in using such model selection criteria, is that at least one model adequately fits the data.[8,9] This raises concerns that without other methods to assess model fit, researchers could make the mistake of choosing the statistical model that best fits the data, rather than an ecological model that accurately describes the environment being studied.

A good ecological model of species distributions may hinge on accurate estimates of overdispersion.[11] Although the method for calculating the overdispersion parameter for occupancy models has been developed recently, few studies have investigated its accuracy with respect to the numerous effects it may have on different data sets.[12] Additionally, the application of this method is still not directly included in popular population modeling software such as MARK.[13] However, the program does have the ability to call on other programs to assess model fit.

Despite this recent advancement in assessing model fit for site occupancy studies, few demonstrated studies have evaluated the potential effects of overdispersion on model results. We had two primary objectives. First, we review the method for assessing model fit used in our study.[12] Second, we validate its use in ecological studies. Specifically, we test the importance of assessing model fit on a data set of detection/non-detection data on winter bird habitat use in Northern Colorado to show the importance of accounting for model fit. With the recent development of powerful statistical software that has increased the access to modelling, studies that evaluate the importance of different methodologies within these programs are crucial to keep scientists aware of ways to evaluate the efficacy and reliability of their model sets.

Methods

Assessing model fit

MacKenzie and Bailey assess model fit by obtaining a test statistic for a model calculated as:

$$X^2 = \sum_{h=1}^{2^T} \frac{(O_h - E_h)^2}{(E_h)}. \tag{4.0}$$[12]

By acquiring a test statistic the over dispersion parameter can be calculated as described by White et. al., and modified for occupancy studies in MacKenzie and Bailey:

$$\hat{c} = \frac{X^2_{obs}}{X^2_c}. \tag{5.0}$$[14]

In (5.0), the denominator represents the average test statistic obtained from the bootstrap procedure described above.[12,14] This estimate can then be used to adjust variation and model selection criteria (i.e., switch from AIC to QAIC).

Field data collection

We collected data on winter bird habitat use from December 2012-March 2013. A total of 53 sites were sampled in Northern Colorado with four each site being sampled four times on average. Sites were sampled along a residential housing gradient built in ArcGIS 10 using a fixed kernel density estimator map that scaled the home range size of the coyote.[15] This study was conducted simultaneously with a study investigating the effects of exurban development on mammalian habitat use. However, the home range size of the coyote is appropriate for this study as it is large enough to encompass the winter home range size of small passerines which were the focus of this study. This was then used as a resistance layer in a least-cost distance map. For the purposes of this paper, this methodology will not be covered as it does not directly relate to assessing model fit. We mean to make no commentary based on the importance of these covariates, but rather use them to draw inference on how accounting for overdispersion can affect model inference.

Fixed-radius point counts were used at each of the 53 sites to assess winter bird habitat use. Each site was visited, then all birds were noted during the duration of a seven-minute count within a 100 meter radius. Additionally, we used ArcGIS to calculate the percent of privately owned land in a 40 meter radius around each point. To account for other anthropogenic covariates, we used a handheld palm pilot from the National Park Service Sounds and Night Sky Division to conduct 15-minute surveys at each point. These surveys were then used to calculate the percent of audible non-natural noise. Finally, in a 25 meter radius around the center of each point, we recorded the ocular proportion of canopy cover, and percent understory cover (to the nearest 5%), and the average understory height of vegetation (to the nearest 0.1 meter). These data were used as site covariates in combination with the residential gradient, percent of land privately owned, and percent audible non-natural noise (Table 12). We chose to collect environmental covariates on a micro scale (25 meter radius around each point), and a macro scale (using ArcGIS) in order to attempt to accurately describe the possible variables that may affect species occurrence. Information on the time of day, temperature (average high for day in degrees C), wind (assessed qualitatively on a scale from 0-5),

Site Name	X Actual	Y Actual	Site Name	X Actual	Y Actual
B2-1	483867	4510303	G1-10	464200	4509073
B2-2	483615	4509717	G1-11	464294	4508333
B2-3	482956	4511847	G1-12	465458	4506880
B2-4	482416	4511103	G1-13	464549	4507127
C1-1	465868	4523248	G1-14	466535	4509776
C1-2	467140	4523731	G1-15	467205	4509168
C1-3	465948	4522843	G1-3	463982	4509791
C1-4	463235	4523412	G1-4	463348	4510539
C1-5	462516	4524985	G1-5	465144	4508427
C1-6	461007	4523529	G1-6	465288	4509670
C1-7	469975	4522413	G1-7	464643	4510563
C1-8	468295	4525369	G1-8	462950	4509562
C1-9	463482	4524075	G1-9	465185	4510246
C2-1	461618	4530621	G2-1	464477	4509977
C2-2	457867	4530278	G2-10	465407	4511262
C2-3	459928	4531678	G2-11	461741	4510321
C2-4	460574	4531915	G2-2	464815	4509810
C2-5	461067	4528801	G2-3	463724	4510151
C2-6	458701	4532193	G2-4	463936	4510501
C2-7	460237	4528897	G2-5	462689	4510645
DM-1	472006	4509257	G2-6	462535	4510406
DM-2	469541	4508237	G2-7	465378	4508030
DM-3	471298	4509557	G2-8	462803	4511118
G1-1	464493	4509513	G2-9	462505	4509727
U1-4	469009	4531360	U1-1	464366	4537487
U2-1	462181	4536457	U1-2	465027	4537177
U2-2	464785	4537422	U1-3	467735	4534961

Table 1. Coordinates for points located in Northern Colorado where data was collected

and percent cloud cover (to nearest 10%) was taken at each point. Furthermore, some point counts were only conducted a few minutes apart and therefore lacked temporal variation. To account for this in our model set, we used a "trap response covariate" as used by many mark-recapture studies.[16] This will not be discussed extensively for the purposes of this study.

Data analysis

Program PRESENCE was used to fit occupancy models, calculate maximum likelihood estimations, and investigate the importance of accounting for overdispersion in occupancy models.[5] For each species, a global model (i.e., most parameterized) was run with 10,000 parametric bootstraps.

We fit the occupancy parameter as constant with six site-level covariates, and fit detection probability as constant with six sample-level covariates (Table 1). For each species, we parameterized the occupancy to its most general state) [Ψ(gradient + % canopy cover + own + % audible non-natural noise + % understory cover)] and fit it with each of the possible covariates for detection probability (Table 2) to come up with the covariates that most impacted detection.[3]

Occupancy (Ψ)	**Detection Probability (p)**
Constant	Constant
Percent Canopy Cover	Time of Day
Residential Gradient	Wind (1-5 scale)
Percent Understory Vegetation Cover	Temperature (C)
Average Understory Height	Percent Cloud Cover
Percent Audile Non-Natural Noise	Percent Canopy Cover
Percent of Land Private in 40m radius	Trap Response

Table 2. The covariates used to estimate site specific occupancy and detection probability.

Species	**c-hat**
Dark-eyed Junco	2.5
Black-billed Magpie	1.3
Townsend's Solitaire	0.9
American Robin	0.5
Mountain Chickadee	1.0
Pygmy Nuthatch	1.3
Steller's Jay	7.0

Table 3. Estimates of the overdispersion parameter as a result of 10,000 parametric bootstraps for each of the global models for the seven species.

Model Set 1	**AIC**	**ΔAIC**	**AIC weight**	**K**	**-2LogLike**
Ψ (ownership), p(wind)	227.1	0	0.36	4	219.1
Ψ (.), p(wind)	227.47	0.37	0.21	3	221.5
Ψ (US height), p(wind)	228.9	1.88	0.10	4	221.0
Model Set 2	**QAIC**	**ΔQAIC**	**QAIC weight**	**K**	**-2LogLike**
Ψ (.), p(wind)	170.25	0	0.24	3	221.5
Ψ (.), p(wind)	170.49	0.24	0.22	4	219.1
Ψ (.), p(wind)	171.88	1.63	0.11	4	221.0

Table 4. The most accurate models (AIC weight>0.10) for the black-billed magpie. The first set of models represent models that were not adjusted by the c-hat parameter, and second set represents models adjusted for over dispersion.

Model Set 1	**AIC**	**ΔAIC**	**AIC weight**	**K**	**2LogLike**
Ψ (US height), p(.)	107.64	0	0.16	3	101.6
Ψ (US height).p(%cloud)	107.74	0.1	0.15	4	99.7
Model Set 2	**QAIC**	**ΔQAIC**	**QAIC weight**	**K**	**2LogLike**
Ψ (.), p(.)	107.74	0	0.15	2	105.9
Ψ (US height), p(.)	46.21	0.33	0.13	3	101.6

Table 5. The most accurate models (AIC weight>0.10) for the dark-eyed junco. The first set of models represent models that were not adjusted by the c-hat parameter, and second set represents models adjusted for over dispersion.

Results

We fit single-season occupancy models on the American Robin (*Turdus migratorius*) the Black-billed Magpie (*Pica hudsonia*), the Steller's Jay (*Cyanocitta stelleri*), the Townsend's Solitaire (*Myadestes townsendi*), the Pygmy Nuthatch (*Sitta pygmaea*), the Dark-eyed Junco (*Junco hyemalis*), and the Mountain Chickadee (*Poecile gambeli*).

Estimates of c-hat for the global model (i.e., most parameterized) varied by species. For example, the American Robin had an estimate of <one overdispersion, indicating the most global model is accurately modeling reality (Table 3). However, for the Steller's Jay, the overdispersion factor was extremely high (Table 2; c=7.0).

The examination of two species yielded different overdispersion parameters, but both estimates were >1.0 reveals the effect of the c-hat parameter on model inference. The c-hat parameter for the Black-billed Magpie was 1.3. Accounting for overdispersion in the model set of the Black-billed Magpie caused model inference to change (Table 4). The most accurate model set (AIC weight>0.10), while not accounting for overdispersion revealed that occupancy was most impacted by land ownership and that detection probability was most impacted by wind (Table 3). However, after accounting for overdispersion within the data set (i.e., model set 2), the null model [Ψ (.), p(wind)], became the top model (i.e. model with highest QAIC weight). In both cases the third model in the top model set (AIC weight>0.10) showed that occupancy was impacted by understory vegetation height (US height).

The c-hat parameter for the Dark-eyed Junco was 2.5 (Table 2). In model set 1, which did not account for overdispersion, the first model (AIC weight=0.16) revealed that occupancy was impacted by understory vegetation height (US height), and that detection probability was constant. In model set 2, which did account for overdispersion, the most accurate model (AIC weight=0.15) was the null model [Ψ(.), p(.)], revealing that occupancy and detection probabilities were constant across the landscape.

Discussion

For our selected species, and for this study area in Northern Colorado, we showed the importance of assessing overdispersion in population models. First, no two species had the same estimated overdispersion parameter. Although all species are common throughout their range and passerines (i.e., common songbirds), the range of values calculated for the overdispersion parameter differed greatly between species. This stresses the importance of calculating the overdispersion parameter for each species separately in conservation and management based research that include multiple species. Although the case could be made that species are generally similar in taxonomy and habitat use, even small differences can result in variation in the data set resulting a range of c-hat estimates. These results are similar to those reached by MacKenzie and Bailey, who found that accounting for overdispersion in occupancy studies was important to properly draw model inference, especially when assumptions of the models were violated.[12]

For the Black-billed Magpie and the Dark-eyed Junco, we examined the effect of accounting for overdispersion in the model set. In both species, adjusting for overdispersion greatly affected the inference drawn from the possible model set.[5,12] Without this calculation, managers and conservationists could come to inaccurate misleading conclusions on species distributions, possibly leading to the improper use of resources. For example, if managers were concerned with Dark-eyed Junco conservation, and based their efforts on a model set that did not account for overdispersion, they might over emphasize restoring understory vegetation where in actuality species persistence was affected little by this covariate. This is a considerable concern when budgets and funding are limited in most agencies.

Adjusting for overdispersion has been more properly documented and accounted for in count data and mark-recapture studies which share many similarities to site-occupancy estimations.[17] Richards suggests there are four ways most studies deal with overdispersion in count data.[17] First, is to estimate it, but to ignore it. Based on our results, failing to estimate the c-hat parameter for each should be greatly discouraged. Second, is to collect additional data (when data are available), to off-set for unexplained variation in the data set. Third, is to model the causes of variation (e.g., non-independence) directly in the model set by inclusion of covariates. Finally, overdispersion can be addressed by modifying model selection methods (e.g., switching from AIC to QAIC). Richards concluded that although in some species AIC and QAIC revealed similar results, and provided one of the first studies with quantitative support for QAIC.[17]

Despite the stressed importance of this, no method currently exists in some popular software used in mark/re-capture studies. For example, program MARK, one of the most popular programs used for population level research with an emphasis on presence/absence and mark/re-capture data, does not currently have the ability to estimate the c-hat parameter for occupancy studies.[13] This suggests that a number of studies underutilize this method, and many may not be aware of its use in occupancy analysis, even though it is more commonly used in count data.

Acknowledgements

I would like to thank Erica Goad and Dr. Liba Pejchar, for their development of this project and help with its implementation. I would like to thank Dr. Larissa Bailey for first introducing me to occupancy modeling, and help in understanding its complexities. I also want to extend a thank you to all the private land owners who allowed us to collect data on their land, and the Center for Collaborative Conservation for funding this project.

References

[1] Morrison, M. L., Marcot, B, G. and Mannan, R. W. (2006) Wildlife-Habitat Relationships: Concepts and Applications. Island Press.

[2] Gardner, T. A., Barlow, J. and Peres, C. A. (2007) “Paradox, presumption and pitfalls in conservation biology: the importance of habitat change for amphibians and reptiles.” Biological Conservation. 138. Pg 166-179.

[3] MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A. and Langtimm, C. A. (2002) “Estimating site occupancy rates when detection probabilities are less than one.” Ecology. 83.8. Pg 2248-2245.

[4] Gu, W. and Swihart, R. K. (2004) “Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models.” Biological Conservation. 116. Pg 195-203.

[5] MacKenzie, D. I., Nicholls, J. D., Royle, J. A., Pollock, K. A., Bailey, L. L. and Hines, J. E. (2006) Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Academic Press.

[6] Geissler, P. H. and Fuller, M. R. (1987) “Estimating of the Proportion of Area Occupied by an Animal Species.” In Proceedings of the Section on Survey Research Methods of the American Statistical Association. American Statistical Association. Pg 553-538.

[7] Azuma, D. L., Baldwin, J. A. and Noon, B. R. (1990) “Estimating the occupancy of spotted owl habitat areas by sampling and adjusting bias.” USDA Forest Service General technical report. PSW-124.

[8] Burnham, K. P. and Anderson, D. R. (2002) Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer-Verlag.

[9] Anderson, D. R., Burnham, K. P. and White, G, C. (1998) “Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies.” Journal of Applied Statistics. 25.2. Pg 263-282.

[10] Guisan, A. and Thuiller, W. (2005) “Predicting species distribution: offering more than simple habitat models.” Ecology Letters. 8. Pg 993-1009.

[11] Potts, J. and Elith, J. (2006) “Comparing species abundance models.” Ecological Modeling. 199. Pg 153-163.

[12] MacKenzie, D, I. and Bailey, L, L. (2004). “Assessing the fit of site-occupancy models.” Journal of Agricultural, Biological, and Environmental Studies. 9.3. Pg 300-318.

[13] White, G. C. and Burnham, K. P. (1999) “Program MARK: survival estimation from population of marked animals.” Bird Study Supplement. 46. Pg 120-138.

[14] White, G. C., Burnham, K. P. and Anderson, D. R. (2002) “Advanced Features of Program Mark.” In Integrating People and Wildlife for a Sustainabile Future: Proceedings of the Second International Wildlife Management Congress. Fields, R., ed. The Wildlife Society.

[15] ESRI (2011) ArcGIS Desktop: Release 10. Environmental Systems Research Institute.

[16] Riddle, J, D., Mordecai, R. S., Pollock, K. H. and Simons, T. R. (2010) “Effects of prior detections on estimates of detection probability, abundance, and occupancy.” The Auk. 127.1. Pg 94-99.

[17] Richards, S, A. (2008) “Dealing with overdispersed count data in applied ecology.” Journal of Applied Ecology. 45.1. Pg 218-227.