A note on investigating cooccurrence patterns and dynamics for many species, with imperfect detection and a log-linear modelling parameterisation

Darryl I. MacKenzie

Proteus, PO Box 7, Outram 9062, New Zealand, and,

Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin 9054, New Zealand

Jason V. Lombardi

Caesar Kleberg Wildlife Research Institute, Texas A& M University-Kingsville, Kingsville, USA

Michael E. Tewes

Caesar Kleberg Wildlife Research Institute, Texas A& M University-Kingsville, Kingsville, USA

Summary:

1. Patterns in, and the underlying dynamics of, species cooccurrence is of interest in many ecological applications. Unaccounted for, imperfect detection of the species can lead to misleading inferences about the nature and magnitude of any interaction. A range of different parameterisations have been published that could be used with the same fundamental modelling framework that accounts for imperfect detection, although each parameterisation has different advantages and disadvantages.

2. We propose a parameterisation based on log-linear modelling that does not require a species hierarchy to be defined (in terms of dominance), and enables a numerically robust approach for estimating covariate effects.

3. Conceptually the parameterisation is equivalent to using the presence of species in the current, or a previous, time period as predictor variables for the current occurrence of other species. This leads to natural, 'symmetric', interpretations of parameter estimates.

4. The parameterisation can be applied to many species, in either a maximum-likelihood or Bayesian estimation framework. We illustrate the method using camera trapping data collected on three mesocarnivore species in South Texas.

Keywords: bobcat (*Lynx rufus*), coyote (*Canis latrans*), imperfect detection, log-linear model, multiple season, ocelot (*Leopardus pardalis*), single season, species cooccurrence

#Introduction Examining patterns of species cooccurrence has a long history in ecology. One of the earliest examples of statistical analysis in modern-day ecology was examining the independence of fish species in Illinois streams using a simple two-way contingency table (Forbes, 1907). Since then there have been a large number of publications devoted to the development, and application, of statistical methods to evaluate the level of independence of species occurrence in an

area of interest (e.g., Dice, 1945; Pielou, 1977; Connor & Simberloff, 1979; Diamond & Gilpin, 1982; Manly, 1995), and investigating possible covariate relationships (e.g., Kelt *et al.*, 1995; Peres-Neto *et al.*, 2001). Prior to the mid-2000's, little attention had been devoted to the practical sampling issue of imperfect detection with species cooccurrence assessments, i.e., species may occur at a surveyed location, yet be undetected by the field methods employed (but see Cam *et al.*, 2000). This will lead to 'false absences' that may result in misleading inferences about species cooccurrence patterns. MacKenzie *et al.* (2006) demonstrated that when the probability of species detection is unaffected by the presence of other species, the direction of any association between the two species (i.e., positive or negative effect on cooccurrence) may be correctly estimated using methods that do not account for imperfect detection, but the magnitude of the dependence will be underestimated. Whereas, when detection probability of one species is different depending on the presence of the second species (e.g., due to behavioural changes in the presence of a competing species), using methods that ignore imperfect detection may not even estimate the direction of any association correctly.

MacKenzie *et al.* (2004) developed a modelling approach to investigate cooccurrence patterns between two species, while accounting for imperfect detection. An important basis of their method is recognising that with two species of interest, a surveyed location may be in one of four possible states defined by the presence or absence of each species (i.e., species A and B present, only species A present, only species B present, or neither species present). MacKenzie *et al.* (2004) parameterised the cooccurrence component of their model in terms of the joint probability of both species occurring at a unit (ψ^{AB}) and the marginal, or overall, probabilities of each species occupying a unit (i.e., ψ^A and ψ^B). They suggested the level of cooccurrence could be quantified in terms of:

$$\varphi = \frac{\psi^{AB}}{\psi^A \psi^B},$$

where a value of 1 would imply independence. They used a similar parameterisation for the detection component, noting that which species could be detected in a survey of a unit would depend on the 'true' state of the location. Potential covariate relationships with any of the parameters could be explored, however it was found to be numerically unstable because of the constraints im-

posed upon possible parameter values (MacKenzie *et al.*, 2006). Richmond *et al.* (2010) and Waddle *et al.* (2010) independently implemented an alternative parameterisation (hereafter referred to as the RW parameterisation) of the MacKenzie *et al.* (2004) model that was more numerically robust, particularly with covariates. The RW parameterisation requires identifying a hierarchy between species where species A is defined as the 'dominant' species and species B is the 'subordinate' species, where the 'subordinate' species is the focal species in an analysis (i.e., how is the occurrence of species B affected by the presence/absence of species A). The model is parameterised in terms of the marginal occurrence probability of species A, and the occurrence probability for species B conditional on species A being either present or absent from the unit (denoted here as $\psi^{B|A}$ and $\psi^{B|a}$, respectively; with the lowercase 'a' indicating absence of species A). A similar conditional parameterisation was also implemented for the detection component of the model. The RW parameterisation could be regarded as 'asymmetric' as a direction to the interaction between species is assumed, while the MacKenzie *et al.* (2004) parameterisation is 'symmetric' as no direction is assumed. While both the MacKenzie *et al.* (2004) and RW models were initially presented in the context of cooccurrence between two species, they generalise to situations with a greater number of species, with the number of possible parameters to estimate increasing exponentially with the number of species (although constraints could be applied to reduce the number of parameters in the model).

Rota *et al.* (2016) developed a species cooccurrence model using a 'multivariate Bernoulli distribution', which has one Bernoulli random variable per species. However this is essentially the same general approach used by earlier authors, where possible states are defined in terms of the combinations of which species are present or absent. Therefore, the Rota *et al.* (2016) model can be considered as another parameterisation, which, for the two-species situation, is in terms of the conditional probabilities $\psi^{A|b}$ and $\psi^{B|a}$, and the odds-ratio of cooccurrence ν (MacKenzie *et al.*, 2018). The odds-ratio ν indicates how the odds of occurrence for one species is different given the presence or absence of the other species, and is the same for either species. The Rota *et al.* (2016) parameterisation is therefore symmetric (as with the MacKenzie *et al.* (2004) model), with the numerical robustness of the RW parameterisation.

The underlying dynamic processes of species cooccurrence are also of interest to many ecologists, although methods to quantify them have received much less attention than those examining

3

cooccurrence patterns, particularly while also accounting for the imperfect detection of the target species (although see MacKenzie *et al.*, 2006; Miller *et al.*, 2012; Haynes *et al.*, 2014; Yackulic *et al.*, 2014; Fidino *et al.*, 2019). As in the static cooccurrence situation, there are numerous ways in which such a model could be parameterised to quantify the level of interaction between species in terms of cooccurrence dynamics (e.g., MacKenzie *et al.*, 2006, 2018; Fidino *et al.*, 2019).

In this paper, we first note the link between the 'multivariate Bernoulli distribution' used by Rota *et al.* (2016), and the well-known statistical method of log-linear modelling. Understanding this connection improves our ability to formulate, and interpret, models for more than two species. We also detail how a dynamic multi-species model could be defined using the log-linear framework, with a simple example application. In the following, we focus on how the models can be parameterised in terms of log-linear models, and do not supply the details of the underlying modelling procedure, as that has been suitably described elsewhere (e.g., MacKenzie *et al.*, 2004; MacKenzie *et al.*, 2009; Richmond *et al.*, 2010; Waddle *et al.*, 2010; Rota *et al.*, 2016; MacKenzie *et al.*, 2018; Fidino *et al.*, 2019).

#Material and Methods ## General sampling situation Throughout this paper we assume a situation where s sampling units (e.g., grid cells, ponds, habitat patches) have been selected from the wider population of units of interest for surveying, ideally using a probabilistic sampling scheme (to extrapolate to unsurveyed units). Units are surveyed for the presence of each species of interest, possibly at systematic points in time when cooccurrence dynamics are of interest. At each of the T survey periods (that shall be referred to as *seasons* henceforth), it is assumed that the species' distributions are static or stable, therefore the pattern of cooccurrence is assumed to be stable in each season. Changes in the distributions, and cooccurrence, is allowed between seasons. Due to imperfect detection, multiple surveys of each unit are conducted each season. The number of surveys may vary spatially and temporally (MacKenzie *et al.*, 2004; MacKenzie *et al.*, 2018).

Log-linear models

4

Table 1: Example of cell probability (π_i) structure for 2×2 contingency table, using the corner-point constraint. U and V are the factors of interest, each with 2 levels. The binary indicator variables (z_i^U and z_i^V) for the second level of each factor are also presented.

U	V	z_i^U	z_i^V	$\log(\pi_i)$	π_i
1	1	0	0	$0 - \log(K)$	$1/K$
2	1	1	0	$\alpha^U - \log(K)$	$\exp(\alpha^U)/K$
1	2	0	1	$\alpha^V - \log(K)$	$\exp(\alpha^V)/K$
2	2	1	1	$\alpha^U + \alpha^V + \alpha^{UV} - \log(K)$	$\exp(\alpha^U + \alpha^V + \alpha^{UV})/K$

Log-linear models are used to analyse count data, particularly to assess the independence of factors used to construct contingency tables, and possibly other predictor variables. Analyses can be conducted on the counts in each cell of the table, or on the underlying cell probability structure (i.e, the probability an observation has a particular combination of factor values). It is not possible to separately estimate parameter values for all combinations of factor levels, and constraints must be applied. One option is the 'corner-point constraint' where the values for parameters associated with one row and one column are set equal to 0, with either the first, or last, row and column typically being used. For example, consider a 2×2 contingency table for factors U and V, and let i index the row and column of the table (i.e., $i = \{u, v\}$, where $u = 1, 2$ and $v = 1, 2$). The log-linear model for the cell probability π_i could be defined as:

$$\log(\pi_i) = \alpha^U + \alpha^V + \alpha^{UV} - \log(K),$$

where K is a normalizing constant such that the π_i's sum to 1.0. The α^U parameter defines the effect of level 2 of factor U on the probability when $v = 1$, the α^V parameter defines the effect of level 2 of factor V on the probability when $u = 1$. The α^{UV} parameter defines the level of interaction, or dependence, between factors U and V on the probability structure. The two factors are independent when $\alpha^{UV} = 0$, and in many applications it is the nature of the interaction between the factors on the cell probabilities (or counts) that is of interest. The cell probabilities for a 2×2 table are given in more detail in Table 1, where $K = 1 + \exp(\alpha^U) + \exp(\alpha^V) + \exp(\alpha^U + \alpha^V + \alpha^{UV})$.

An equivalent approach to using the corner-point constraint, is to define the log-linear model in terms of binary indicator variables representing the levels of each factor of interest. For example, if a factor contains M levels select one level to use as a reference category, then define $M-1$ binary indicator variables for observations from the other levels for that factor. In the 2×2 contingency table case, using the first level of factors U as V the 'reference' levels, then the indicator variables z_i^U and z_i^V can be defined, which equal 1 if the observed factor level was 2, and equal 0 otherwise (Table 1). The log-linear model can then be expressed as:

$$\log(\pi_i) = \alpha^U z_i^U + \alpha^V z_i^V + \alpha^{UV} z_i^U z_i^V - \log(K).$$

Hence, in a regression context, the indicator variables are predictor variables representing the combination of factor levels for an observation, and the α terms are regression coefficients quantifying the magnitude of the effect for each factor level. Coefficients associated with an interaction between two (or more) factors, e.g., the parameter α^{UV} for the $z_i^U z_i^V$ interaction, quantifies how the effect of one factor is different depending on the value of the other factor(s).

When there is more than 2 levels for a factor, then the log-linear model generalises in the obvious manner. For example, if factor U had 2 levels and factor V contained 3, the indicator variables z_i^{V2} and z_i^{V3} could be defined to equal 1 if the observed factor level was 2 or 3, respectively. The log-linear model would then be:

$$\log(\pi_i) = \alpha^U z_i^U + \alpha^{V2} z_i^{V2} + \alpha^{V3} z_i^{V3} + \alpha^{UV2} z_i^U z_i^{V2} + \alpha^{UV3} z_i^U z_i^{V3} - \log(K).$$

Similarly, the approach easily generalises to a greater number of factors. For example, with three factors (U, V and W) with two levels each, then:

$$\log(\pi_i) = \alpha^U z_i^U + \alpha^V z_i^V + \alpha^W z_i^W + \alpha^{UV} z_i^U z_i^V + \alpha^{UW} z_i^U z_i^W + \alpha^{VW} z_i^V z_i^W + \alpha^{UVW} z_i^U z_i^V z_i^W - \log(K).$$

In all cases K would be defined differently to ensure that the cell probabilities sum to one.

155 Species cooccurrence data, assuming perfect detection, can be represented as a contingency ta-

156 ble. Each factor is a species, and the absence/presence case there are two levels for each species

157 (henceforth denoted with lowercase and uppercase characters, respectively). The structure of the

158 possible observations for two species (species *A* and *B*), indicator variables and associated cell

159 probability structure is given in Table 2. The log-linear model, expressed in terms of the indicator

160 variables, would therefore be:

$$\log(\pi_i) = \alpha^A z_i^A + \alpha^B z_i^B + \alpha^{AB} z_i^A z_i^B - \log(K),$$

161 where z^A and z^B are the binary-valued variables indicating the presence of each species. While

162 covariates have not been considered here, the general cell probability structure is the same as that

163 used Rota *et al.* (2016) where the set of indicator variables represent their 'multivariate Bernoulli

164 distribution', with α^A, α^B and α^{AB} being equivalent to the f_1, f_2 and f_{12} parameters defined by

165 Rota *et al.* (2016).

Table 2: Example of cell probability (π_i) structure for a 2-species (*A* and *B*) cooccurrence application.

Sp. A	Sp. B	State (*i*)	z_i^A	z_i^B	$\pi_{u,v}$
Absent	Absent	ab	0	0	$1/K$
Present	Absent	Ab	1	0	$\exp(\alpha^A)/K$
Absent	Present	aB	0	1	$\exp(\alpha^B)/K$
Present	Present	AB	1	1	$\exp(\alpha^A + \alpha^B + \alpha^{AB})/K$

166 As shown by Rota *et al.* (2016), the model parameters are directly interpretable in terms of the

167 probability of each species being present, conditional upon the presence or absence of the other

168 species. That is:

$$\text{logit}(\psi^{A|b}) = \alpha^A,$$

$$\text{logit}(\psi^{A|B}) = \alpha^A + \alpha^{AB},$$

$$\text{logit}(\psi^{B|a}) = \alpha^B,$$

$$\text{logit}(\psi^{B|A}) = \alpha^B + \alpha^{AB}.$$

169 Therefore, α^A and α^B determine the probability of occupancy (on the logit-scale) for each species

170 given the absence of the other species, and α^{AB} is the effect that the presence of one species has on

171 the other. Hence, α^{AB} parameter is a symmetric measure of cooccurrence between the two species,

172 where $\alpha^{AB} = 0$ indicates the species cooccur independently, while a negative value indicate some

173 form of exclusion or avoidance, and a positive value indicate the species tend to occur together.

174 Inferences about the level of cooccurrence between species could be based on estimates of α^{AB}

175 (e.g., by considering confidence intervals), or one could 'test' for independence of the species by

176 comparing the fit of a model where α^{AB} is estimated, to the fit of a model with the constraint

177 $\alpha^{AB} = 0$. Note that the level of association, can also be expressed as an odds ratio:

$$v = \exp(\alpha^{AB})$$
$$= \frac{\psi^{A|B}/(1 - \psi^{A|B})}{\psi^{A|b}/(1 - \psi^{A|b})}$$
$$= \frac{\psi^{B|A}/(1 - \psi^{B|A})}{\psi^{B|a}/(1 - \psi^{B|a})}.$$

178 Therefore, this is similar to the RW parameterisation, but the interaction between species is mod-

179 elled as a symmetric relationship.

180 Heuristically, the presence or absence of one species is being used as a covariate on the proba-

181 bility of occurrence of the other species.

182 The extension to more than two species is therefore straightforward. For example, with three

183 species a third indicator variable can be defined (z^C) and the model for the contingency table cell

184 probabilities becomes:

$$\log(\pi_i) = \alpha^A z_i^A + \alpha^B z_i^B + \alpha^C z_i^C + \alpha^{AB} z_i^A z_i^B + \alpha^{AC} z_i^A z_i^C + \alpha^{BC} z_i^B z_i^C + \alpha^{ABC} z_i^A z_i^B z_i^C - \log(K).$$

185 The parameters α^{AB}, α^{AC} and α^{BC} quantify the two-way interactions between species, and α^{ABC}

the three-way interaction. As noted by Rota *et al.* (2016), and also MacKenzie *et al.* (2018), it is not always necessary to estimate higher order interaction terms between many species, and in fact very large sample sizes may be required to obtain reliable parameter estimates. Furthermore, complex interactions between many species will be difficult to interpret biologically. Therefore some higher-order interaction terms may be set equal to zero. In the log-linear modelling literature, this is known as conditional independence. For example, the occurrence of species A and B may appear to be not independent, but that is because both species have a non-independent cooccurrence relationship with species C. Given the presence or absence of species C, species A and B occur independently of each other (i.e., species A and B are conditionally, upon species C, independent). This hypothesis could be fit by constraining $\alpha^{ABC} = 0$ and $\alpha^{AB} = 0$.

Covariates

The effect of potential covariates on the occurrence, or cooccurrence, for each species can be easily incorporated in the log-linear modelling framework, where the effect of such covariates may be the same, or different for each species. For example, if a covariate x_1 is thought to affect the occurrence of species A, the covariate x_2 affect the occurrence of species B, but the level of cooccurrence interaction is unaffected by either covariate, the following model could be fit to the data:

$$\log(\pi_i) = (\alpha^A + \beta_1^A x_1)z_i^A + (\alpha^B + \beta_2^B x_2)z_i^B + \alpha^{AB}z_i^A z_i^B - \log(K).$$

If covariate x_1 is also thought to affect the level of interaction between species, then another model could be fit:

$$\log(\pi_i) = (\alpha^A + \beta_1^A x_1)z_i^A + (\alpha^B + \beta_2^B x_2)z_i^B + (\alpha^{AB} + \beta_1^{AB} x_1)z_i^A z_i^B - \log(K).$$

Interpretation of the covariate effects would proceed exactly as normal.

Extension to multiple-seasons

To examine how species cooccurrences change over time, it is neccesary to have data from multiple seasons, preferably at equally-spaced intervals. A general approach to analysing such data is to model how the combination of species present at each unit changes over time. A transition

9

probability matrix (TPM) can be defined, which provides the probability structure for which combination of species are present in season $t+1$, given that combination of species present at a unit in season t (MacKenzie *et al.*, 2018). For example, in the two-species case, the TPM would be of the form:

$$
\phi_t =
\begin{bmatrix}
ab \to ab & ab \to Ab & ab \to aB & ab \to AB \\
Ab \to ab & Ab \to Ab & Ab \to aB & Ab \to AB \\
aB \to ab & aB \to Ab & aB \to aB & aB \to AB \\
AB \to ab & AB \to Ab & AB \to aB & AB \to AB
\end{bmatrix}
$$

where X $\to$ Y denotes the probability of transitioning from occupancy state X in season t to state Y in season $t+1$ (where the states are denoted as above). Importantly, the elements of each row must sum to 1, as a unit must be of one of the four states by the next season. When there are l species of interest, then the dimension of the TPM will be $2^l \times 2^l$.

As noted by MacKenzie *et al.* (2018), there are a range of possible parameterisations that could be used to estimate the parameters associated with the transition probabilities. Building on the log-linear parameterisation outlined above for the single-season situation, the expected cell probabilities could be defined in terms of the binary indicator variables for the presence/absence of each species at both times t and $t+1$ (Table 3).

Table 3: Binary variable coding for 2-species multi-season cooccurrence model.

Row	Column	State t (i)	State $t+1$ (j)	z_i^A	z_i^B	z_j^A	z_j^B
1	1	ab	ab	0	0	0	0
1	2	ab	Ab	0	0	1	0
1	3	ab	aB	0	0	0	1
1	4	ab	AB	0	0	1	1
2	1	Ab	ab	1	0	0	0
2	2	Ab	Ab	1	0	1	0
2	3	Ab	aB	1	0	0	1
2	4	Ab	AB	1	0	1	1
3	1	aB	ab	0	1	0	0

Row	Column	State t (i)	State $t+1$ (j)	z_i^A	z_i^B	z_j^A	z_j^B
3	2	aB	Ab	0	1	1	0
3	3	aB	aB	0	1	0	1
3	4	aB	AB	0	1	1	1
4	1	AB	ab	1	1	0	0
4	2	AB	Ab	1	1	1	0
4	3	AB	aB	1	1	0	1
4	4	AB	AB	1	1	1	1

Let z_i^X denote the presence of species X in given state in season t, z_j^X denote the presence of the species in season $t+1$. The general structure for the cell probability in row i and column j could be defined as:

$$
\begin{aligned}
\log(\pi_{i,j}) =& \beta^A z_j^A + \beta^B z_j^B + \beta^{AB} z_j^A z_j^B \\
&+ \left(\gamma^A z_j^A + \gamma^B z_j^B + \gamma^{AB} z_j^A z_j^B \right) z_i^A \\
&+ \left(\delta^A z_j^A + \delta^B z_j^B + \delta^{AB} z_j^A z_j^B \right) z_i^B \\
&+ \left(\xi^A z_j^A + \xi^B z_j^B + \xi^{AB} z_j^A z_j^B \right) z_i^A z_i^B \\
&- \log(K)
\end{aligned}
$$

where K is a normalising constant defined to ensure the probabilities for each row of the TPM sum to 1.

This is a very general formulation, allowing complex relationships about the dynamic cooccurrence processes to be evaluated, providing sufficient data. However the model can be simplified by applying constraints to some parameters. For example, the γ, δ and ξ parameters are all associated with the effects of the presence of each species in the previous season (season t), on which combination of species are present in the current season (season $t+1$). This represents a situation where changes in occurrence (and cooccurrence) can be represented as a Markov process.

Constraining all these parameters to equal 0 represents a model where the probability of which species are present in season $t + 1$ is independent of the combination of species that were present in season t (i.e., non-Markovian, or a random process). Alternatively, one may set only the ξ parameters to 0, representing a situation where the presence of each species in season t has an effect on the cooccurrence structure in season $t + 1$, but only as additive effects. If the constraints $\beta^{AB} = \gamma^{B} = \gamma^{AB} = \delta^{A} = \delta^{AB} = 0$ are also enforced, that represents a model where the occurrence of each species changes as a Markov process, but changes are independent for each species. Finally, in the model where $\delta = \xi = 0$, the γ parameters indicate how the presence of species A in season t affects the cooccurrence between the species in the next season. Specifically, the parameters γ^{B} and γ^{AB} quantify what effect the presence of species A in season t has on the probability of species B being present in season $t + 1$. One could make a-priori predictions about the expected direction of such effects based on whether the species are considered to exclude one another, or not.

Generalising to a greater number of species is achieved by defining the respective set of binary indicator variables for the presence of each species in seasons t and $t + 1$, with potentially a large number of parameters associated with the full model (including all interaction terms amongst species). Regardless of whether it is possible to estimate many of those parameters for a given data set, interpretation of the effects may be challenging. Hence, it is recommended that practitioners limit the number of interaction terms they include in a model when analysing data, and carefully consider the biological interpretation of the estimates.

Modelling the detection component

An important consideration for modelling the detection component is that the possible number of categories, or types of detection, will vary depending on which combination of species are present at a unit. For example, if only one species of interest is present at a unit then there are two types of detections (nondetection/detection of that species), while if two of the target species are present there are four possible detection outcomes from a survey. This is demonstrated in Table 4 for the two-species case. The number of possible observations can be accounted for by defining the detection component to be both a function of the true (but unknown) presence/absence of the species (z_i^{X} indicator variables) and binary indicator variables based on the observed outcomes of

12

each survey, which will be defined as h_k^X.

Table 4: Possible observations admitting imperfect detection. Lowercase characters for the true state or survey observation (Obs) indicate the absence or nondetection of that species, respectively, while uppercase characters indicate the presence or detection of that species. z_i^X are the binary indicator variables for the presence or absence of species X and h_k^X are the binary indicator variables for the detection or nondetection of species X in a survey.

True State (i)	z_i^A	z_i^B	Obs (k)	h_k^A	h_k^B
ab	0	0	ab	0	0
Ab	1	0	ab	0	0
Ab	1	0	Ab	1	0
aB	0	1	ab	0	0
aB	0	1	aB	0	1
AB	1	1	ab	0	0
AB	1	1	Ab	1	0
AB	1	1	aB	0	1
AB	1	1	AB	1	1

Detection probability can therefore be defined using a log-linear modelling framework as:

$$
\begin{aligned}
\log(p_{i,k}) =\ & \eta^A h_k^A z_i^A \\
& + \eta^B h_k^B z_i^B \\
& + \left(\zeta^A h_k^A + \zeta^B h_k^B + \zeta^{AB} h_k^A h_k^B \right) z_i^A z_i^B \\
& - \log(K)
\end{aligned}
$$

where,

$$
\begin{aligned}
K =\ & 1 + (\exp(\eta_1))\, z_i^A (1 - z_i^B) \\
& + (\exp(\eta_2))\, (1 - z_i^A) z_i^B \\
& + (\exp(\eta_1 + \eta_3) + \exp(\eta_2 + \eta_4) + \exp(\eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5))\, z_i^A z_i^B.
\end{aligned}
$$

264 The motivation for developing this parameterisation of the multi-season cooccurrence model was
265 a 7-year camera trap dataset of bobcats (*Lynx rufus*), ocelot (*Leopardus pardalis*) and coyote (*Canis*
266 *latrans*) collected in South Texas (Lombardi *et al.*, 2020). This dataset is part of a long-term ocelot
267 monitoring study on the East Foundation's El Sauz Ranch in Willacy and Kenedy counties, Texas.
268 Although ocelot share a geographic overlap with bobcats and coyotes from South Texas to Central
269 Mexico (Sánchez-Cordero *et al.*, 2008; Horne *et al.*, 2009; Hody & Kays, 2018), interactions among
270 this community are poorly understood in this region.

271 From 8 May 2011 to 24 March 2018, 56 camera traps (Cuddeback® white-flash Expert Scout-
272 ing Cameras and Cuddeback® X-Change Color cameras (NonTypical, Isanti, WI, USA) were de-
273 ployed at 28 paired camera stations in the northwestern and southwestern regions of the El Sauz
274 Ranch. Camera traps were set in forests containing live oak (*Quercus virginiana*), honey mesquite
275 (*Prosopis glandulosa*), and thornshrub (lime prickly ash [*Zanthoxylum fagara*], huisache [*Acacia far-*
276 *nesiana*], and spiny hackberry [*Celtis pallida*]). Camera stations were spaced 1 km apart, which was
277 based on the mean minimum distance moved for ocelots in the region (M. Tews, unpub. data). At
278 a station, cameras were placed facing each other and offset 1-2 meters, with each camera attached
279 to a tree or wooden stake about 30 cm above the ground. Camera stations were maintained all
280 year and cameras were replaced if they malfunctioned (Lombardi *et al.*, 2020).

281 A sampling season was defined to be a 20-week period, either 8 May to 23 September (hot
282 season) or 8 November to 24 March (cool season). A survey was defined to be a 4-week period,
283 i.e., a species was detected ($h_k^X = 1$) if it was photographed at least once at a station during the
284 4-week period, and undetected ($h_k^X = 0$) otherwise. Hence each season comprised of 5 surveys.
285 Surveys were defined to be a 4-week period such that detections of bobcats and coyotes within a
286 survey period could be assumed independent (Lombardi *et al.*, 2020).

287 The log-linear parameterisation discussed above provides a great deal of flexibility for exam-
288 ining the patterns and dynamics of cooccurrence between multiple species, especially given the
289 ability to incorporate spatial and temporal covariates. However, given the number camera stations
290 deployed (i.e., 28 surveyed units), only relatively simple models are fit to the data here to illustrate
291 some key concepts. Lombardi *et al.* (2020) conduct a fuller analysis of the data set examining the

292 effect of covariates.

293   Five models were fit to the data-set, each representing a different set of hypotheses about cooc-
294 currence patterns and dynamics (Table 5). While model parameters could be season-specific, they
295 have been assumed to be season invariant. Additional information about the exact parameter-
296 isation is supplied in the Supplemental Material. The same detection component was assumed
297 for all models, where a separate detection probability was estimated for each species, which was
298 assumed to be independent of both the presence and detection of other species. Model 1 assumes
299 species occur near camera trap stations independently of each other, and the probability of occur-
300 rence is the same each season and independent of the species being present near a station in the
301 previous season. Model 3 also assumes species occur independently of each other, although the
302 probability of occurrence after season 1 depends on the presence of the species in the previous sea-
303 son. This is equivalent to modelling the occurrence of each species as independent single-species
304 multi-season models (MacKenzie *et al.*, 2003), where changes in occurrence is assumed to be a
305 first-order Markov process.

Table 5: Summary of effects included in each model fit to the Texas cam-
era trapping data. '2-way Interaction' is interaction effects between pairs of
species, 'Depends on z_i^X' and 'Depends on z_i^Y' indicates whether occurrence
in the current season depends on the presence of the focal (X), or other (Y)
species in the previous season.

Model	2-way Interactions	Depends on z_i^X	Depends on z_i^Y
1	N	N	N
2	Y	N	N
3	N	Y	N
4	Y	Y	N
5	Y	Y	Y

306   The species cooccurrence models were fit using maximum likelihood techniques (e.g., MacKen-
307 zie *et al.*, 2004; MacKenzie *et al.*, 2009; Richmond *et al.*, 2010; Waddle *et al.*, 2010; MacKenzie *et al.*,
308 2018) using custom-written R code, although Bayesian methods could also be used (e.g., Rota *et al.*,
309 2016; Fidino *et al.*, 2019). Models were compared on the basis of Akiake's Information Criterion

(AIC).

Results

Example – mesocarnivores in Texas

Table 6 presents a summary of the five models fit to the mesocarnivore data. On the basis of AIC, Model 4 had the majority of the support with 79% of the AIC model weight, and Model 5 also has some support with 21% AIC model weight. The results provide strong evidence the probability of a species occurring near a station is dependent on the presence of the species near the station in the previous seasons (given ranking of Models 3-5), and affected by the presence of other species in the same season (Models 4 and 5 ranked highest). The is some indication that occurrence may also depend on the presence of other species in the previous season (Model 5 ranked second).

Table 6: Summary of model comparison process. Given is the relative difference in AIC (ΔAIC), AIC model weight (w), number of estimated parameters (K) and two times the negative log-likelihood value ($-2ll$).

Model	ΔAIC	w	K	$-2ll$
1	175.20	0.00	6	6298.15
2	66.11	0.00	9	6183.06
3	104.62	0.00	12	6215.57
4	0.00	0.79	15	6104.95
5	2.65	0.21	21	6095.59

From Model 4, the estimated probability of detecting ocelots, bobcats and coyotes during 4 weeks of camera trapping was estimated to be 0.43 (0.02), 0.49 (0.01) and 0.51 (0.01), respectively (standard error in parentheses). For each of the three species, the probability of occurrence in the current season is estimated to be higher if they were present in the previous season, particularly for ocelots, although the effect is small for bobcats (Table 7; parameters γ^O, δ^B and ξ^C). Note that under the parameterisation used here, the β parameters determine the probability of occurrence given the absence of the species in the previous season, i.e., the probability of colonisation. Therefore, the γ^O, δ^B and ξ^C parameters are the difference between the colonisation and persis-

tence probabilities (on the logit-scale) for the respective species. The estimated 2-way interaction terms (parameters α^{OB}, α^{OC} and α^{BC}) are all positive, indicating that if one species is present, the other species are more likely to be also present. The odds-ratio for the cooccurrence of ocelots and bobcats is estimated to be 4.16, 5.31 for ocelots and coyotes, and 5.88 for bobcats and coyotes. The confidence intervals for each of the odds ratios are relatively wide, which is a reflection of the number of surveyed stations, although the intervals are all greater than 1.0 suggesting strong evidence of a positive correlation.

Table 7: Parameter estimates from Model 4. Given are the estimates, associated standard errors, estimated odds ratio (OR) with associated lower and upper limits of 95% confidence intervals.

Parameter	Est	SE	OR	Lower	Upper
α^{O}	-2.10	0.63	0.12	0.04	0.42
α^{B}	-1.16	0.55	0.31	0.11	0.92
α^{C}	-1.29	0.56	0.28	0.09	0.82
α^{OB}	1.43	0.37	4.16	2.03	8.53
α^{OC}	1.67	0.48	5.31	2.07	13.60
α^{BC}	1.77	0.34	5.88	2.99	11.56
β^{O}	-3.72	0.51	0.02	0.01	0.07
β^{B}	-0.89	0.34	0.41	0.21	0.81
β^{C}	-0.56	0.34	0.57	0.30	1.11
γ^{O}	2.11	0.30	8.24	4.62	14.69
δ^{B}	0.06	0.31	1.06	0.57	1.96
ξ^{C}	0.55	0.36	1.74	0.87	3.49

Discussion

The log-linear parameterisation outlined here for the multi-season, multi-species cooccurrence model is not unique, and other parameterisations are possible (e.g., MacKenzie *et al.*, 2006, 2018; Fidino *et al.*, 2019). The log-linear parameterisation provides the ability to directly estimate, and interpret, how the presence of species is affected by the presence of other species in either the

current, or previous, season. With this structure, the presence of each species is essentially being used as a predictor variable for the presence of other species, although the general framework that accounts for imperfect detection allows for the fact that the presence of any species may not be known with certainty. Furthermore, the parameterisation can also be applied to the detection process, to allow for non-independent detections of each species.

Complexity breeds complexity. As practitioners attempt to address more complex questions of ecological data, more complex methods of analysis are generally required to provide quantitative inspections of that data. Such is the case with multi-season, multi-species cooccurrence models. Irrespective of the preferred parameterisation to be used, proper analysis should involve careful consideration of hypotheses of interest, which species interactions should be included and whether such interactions change over time, effect of potential covariates for cooccurrence- and detection-related parameters. Proper analysis will require time, and some degree of skill in fitting and interpreting model results. While tools can be developed to simplify certain aspects of the process, practitioners should have a realistic expectation that such analysis require a substantial investment of time and effort.

Practitioners are strongly encouraged to gain a realistic expectation of the type, and quantity, of data required to achieve their objectives, before embarking on any data collection. Complex models, with a large number of biologically relevant parameters to estimate, will require relatively large datasets to produce accurate estimate with suitable levels of precision. Simulation studies are an incredibly useful approach to evaluating the expected quality of the results from a proposed study design. The outcome will often be enlightening, and sometimes, sobering. While the exact outcome will depend on the specifics of the situation, in general we suggest that typically the number of sampling units required to be survey will be in the 100's rather than the 10's of units. This is based on our experience with similar models, and on the simple premise that there is not a lot of information in binary observations, and therefore a large number of them tend to be required to obtain adequate precision of parameter estimates.

Log-linear modelling can be used in situations where a factor of interest has m levels (with $m \geq 2$), by defining $m - 1$ indicator variables. In this paper we have focused on situations where $m = 2$ (i.e., species presence or absence), although as alluded to above, this parameterisation extends naturally to situations where the occurrence of species may be defined using

18

a greater number of categories (e.g., absent, present without breeding, present with breeding). The log-linear modelling parameterisation therefore provides a framework for assessing relevant questions about cooccurrence patterns and dynamics for more these more complex situations, in combination with multi-state occupancy models (e.g., Royle & Link, 2005; Nichols *et al.*, 2007; MacKenzie *et al.*, 2009).

This parameterisation of a many-species cooccurrence model is currently being incorporated into Program PRESENCE and the `RPresence` R package.

Authors' contributions

MT led the design of the camera trapping study and initial data collection; JL was responsible for the data collection in some years, processed the data into a usable format for analysis, and developing initial questions about species cooccurrence; DM conceptualised the parameterisation, developed the computer code, analysed the data, and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

Data used in this manuscript (i.e., three-species camera data, and high resolution spatial data) are accessible in the repository Dryad. Please see https://doi.org/10.5061/dryad.931zcrjgp.

References

Cam, E., Nichols, J.D., Hines, J.E. & Sauer, J.R. (2000) Inferences about nested subsets structure when not all species are detected. *Oikos*, **91**, 428–434.

Connor, E.F. & Simberloff, D. (1979) The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Diamond, J.M. & Gilpin, M.E. (1982) Examination of the "null" model of Connor and Simberloff for species co-occurrences on islands. *Oecologia*, **52**, 64–74.

Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Fidino, M., Simonis, J.L. & Magle, S.B. (2019) A multistate dynamic occupancy model to estimate local colonization–extinction rates and patterns of co-occurrence between two or more interacting species. *Methods in Ecology and Evolution*, **10**, 233–244.

Forbes, S.A. (1907) On the local distribution of certain Illinois fishes: An essay in statistical ecology. *Bulletin of the Illinois State Laboratory of Natural History*, **7**, 273–303.

Haynes, T.B., Schmutz, J.A., Lindberg, M.S., Wright, K.G., Uher-Koch, B.D. & Rosenberger, A.E. (2014) Occupancy of yellow-billed and Pacific loons: evidence for interspecific competition and habitat mediated co-occurrence. *Journal of Avian Biology*, **45**, 296–304.

Hody, J.W. & Kays, R. (2018) Mapping the expansion of coyotes (Canis latrans) across North and Central America. *ZooKeys*, p. 81.

Horne, J.S., Haines, A.M., Tewes, M.E. & Laack, L.L. (2009) Habitat partitioning by sympatric ocelots and bobcats: implications for recovery of ocelots in southern Texas. *The Southwestern Naturalist*, **54**, 119–127.

Kelt, D.A., Taper, M.L. & Mesevre, P.L. (1995) Assessing the impact of competition on community assembly: a case study using small mammals. *Ecology*, **76**, 1283–1296.

Lombardi, J.V., MacKenzie, D.I., Tewes, M.E., Perotto-Baldivieso, H.L., Mata, J.M. & Campbell, T.A. (2020) Co-occurrence of bobcats, coyotes, and ocelots in Texas. *Ecology and Evolution*.

MacKenzie, D.I., Bailey, L.L. & Nichols, J.D. (2004) Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, **73**, 546–555.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, **84**, 2200–2207.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy estimation and modeling : inferring patterns and dynamics of species occurrence*. Elsevier, Amsterdam ;;Boston.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2018) *Occupancy estimation and modeling : inferring patterns and dynamics of species occurrence*. Elsevier, Amsterdam ;;Boston, 2nd edition.

MacKenzie, D.I., Nichols, J.D., Seamans, M.E. & Gutiérrez, R.J. (2009) Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology*, **90**, 823–835.

Manly, B.F.J. (1995) A note on the analysis of species co-occurrences. *Ecology*, **76**, 1109–1115.

Miller, D.A., Brehme, C.S., Hines, J.E., Nichols, J.D. & Fisher, R.N. (2012) Joint estimation of habitat dynamics and species interactions: disturbance reduces co-occurrence of non-native predators with an endangered toad. *Journal of Animal Ecology*, **81**, 1288–1297.

Nichols, J.D., Hines, J.E., Mackenzie, D.I., Seamans, M.E. & Gutiérrez, R.J. (2007) Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology*, **88**, 1395–1400.

Peres-Neto, P., Olden, J.D. & Jackson, D.A. (2001) Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, **93**, 110–120.

Pielou, E.C. (1977) *Mathematical ecology*. Wiley, 2nd edition.

Richmond, O.M., Hines, J.E. & Beissinger, S.R. (2010) Two-species occupancy models: a new parameterization applied to co-occurrence of secretive rails. *Ecological Applications*, **20**, 2036–2046.

Rota, C.T., Ferreira, M.A., Kays, R.W., Forrester, T.D., Kalies, E.L., McShea, W.J., Parsons, A.W. & Millspaugh, J.J. (2016) A multispecies occupancy model for two or more interacting species. *Methods in Ecology and Evolution*, **7**, 1164–1173.

Royle, J.A. & Link, W.A. (2005) A general class of multinomial mixture models for anuran calling survey data. *Ecology*, **86**, 2505–2512.

Sánchez-Cordero, V., Stockwell, D., Sarkar, S., Liu, H., Stephens, C.R. & Giménez, J. (2008) Competitive interactions between felid species may limit the southern distribution of bobcats Lynx rufus. *Ecography*, **31**, 757–764.

Waddle, J.H., Dorazio, R.M., Walls, S.C., Rice, K.G., Beauchamp, J., Schuman, M.J. & Mazzotti, F.J. (2010) A new parameterization for estimating co-occurrence of interacting species. *Ecological Applications*, **20**, 1467–1475.

Yackulic, C.B., Reid, J., Nichols, J.D., Hines, J.E., Davis, R. & Forsman, E. (2014) The roles of competition and habitat in the dynamics of populations and species distributions. *Ecology*, **95**, 265–279.