Logistic regression for clustered data from environmental monitoring programs


A R T I C L E   I N F O

Keywords:
Bryoria
Cluster-specific model
Complex sampling design
Correlated data
Logistic regression
National forest inventory
Population-averaged model

A B S T R A C T

Large-scale surveys, such as national forest inventories and vegetation monitoring programs, usually have complex sampling designs that include geographical stratification and units organized in clusters. When models are developed using data from such programs, a key question is whether or not to utilize design information when analyzing the relationship between a response variable and a set of covariates. Standard statistical regression methods often fail to account for complex sampling designs, which may lead to severely biased estimators of model coefficients. Furthermore, ignoring that data are spatially correlated within clusters may underestimate the standard errors of regression coefficient estimates, with a risk for drawing wrong conclusions. We first review general approaches that account for complex sampling designs, e.g. methods using probability weighting, and stress the need to explore the effects of the sampling design when applying logistic regression models. We then use Monte Carlo simulation to compare the performance of the standard logistic regression model with two approaches to model correlated binary responses, i.e. cluster-specific and population-averaged logistic regression models. As an example, we analyze the occurrence of ephiphytic hair lichens in the genus Bryoria; an indicator of forest ecosystem integrity. Based on data from the National Forest Inventory (NFI) for the period 1993–2014 we generated a data set on hair lichen occurrence on > 100,000 Picea abies trees distributed throughout Sweden. The NFI data included ten covariates representing forest structure and climate variables potentially affecting lichen occurrence. Our analyses show the importance of taking complex sampling designs and correlated binary responses into account in logistic regression modeling to avoid the risk of obtaining notably biased parameter estimators and standard errors, and erroneous interpretations about factors affecting e.g. hair lichen occurrence. We recommend comparisons of unweighted and weighted logistic regression analyses as an essential step in development of models based on data from large-scale surveys.

1. Introduction

Long-term monitoring programs, such as national forest inventories (Fridman et al., 2014; Lawrence et al., 2014) play a key role in providing ecological and environmental data for decision making and research (Lindenmayer and Likens, 2010). The need for efficiency in field studies often leads to a spatial structure in the sampling design of such programs. Travel between sampling locations and other logistic aspects usually imply that several of the units in question are sampled at each location (Berglund et al., 2009; Nordén et al., 2013). These circumstances result in hierarchical data sets, where units collected at the same location tend to be similar to one another, i.e., their properties are correlated. Although this may be a practical necessity it implies a statistical “cost”, i.e. selecting an additional unit from the same location adds less new information than a completely independent selection.

Data from monitoring programs often include binary responses, such as presence/absence of species, which can be modeled through logistic regression to assess the relationship between the probability of a response (e.g. occurrence of the species) and a set of covariates. A standard logistic (SL) regression model relies on the assumption that observations are independent (Hosmer et al., 2013), and ignoring existing correlations in the data may result in substantially biased standard errors of logistic regression coefficient estimators – they are typically underestimated but may be overestimated in some cases (Fieberg et al., 2009; Fitzmaurice et al., 1993; Wilson and Lorenz, 2015).

Corresponding author.
E-mail address: Magnus.Ekstrom@umu.se (M. Ekström).

https://doi.org/10.1016/j.ecolinf.2017.10.006
Received 12 June 2017; Received in revised form 11 October 2017; Accepted 18 October 2017
Available online 18 October 2017
1574-9541/ © 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).
Two common approaches to model correlated binary responses are cluster-specific (CS) and population-averaged (PA) models (Fieberg et al., 2009; Hosmer et al., 2013; Muff et al., 2016). Both of them extend the SL model for independent observations to correlated data, but they address the problem of correlated data differently. In addition, estimated regression coefficients obtained from these two approaches are numerically different, as are their interpretations (Hu et al., 1998). These differences are not always well understood in ecological research (Fieberg et al., 2009). Both Fieberg et al. (2009) and Muff et al. (2016) evaluated the usefulness of the CS and PA approaches in an ecological context, with illustrative examples of temporally clustered data from a 3-year study of mallard nest structures. Here, we analyze the same approaches, but for spatially correlated large-scale survey data collected using complex sampling designs that involve geographical stratification and units selected in groups (cluster sampling; Lumley, 2010). Survey data may be viewed as the outcome of two processes: the process (e.g. the habitat suitability for the species in question) that generates the values of units in the population, often referred to as the superpopulation model (Heeringa et al., 2010), and the process of selecting the sample units from the population. Regression methods like SL that do not account for complex sampling designs may have biased estimates of coefficients and standard errors (Pfeffermann and Sverchkov, 2009; Wilson and Lorenz, 2015).

In this study, we compare the performance of SL, CS, and PA logistic regression estimates under complex sampling designs using Monte Carlo simulation. We first review the SL, CS, and PA models. We then discuss commonly used approaches in regression modeling that account for informative sampling designs; a design is said to be informative if the inclusion probabilities are related to the response even after taking the covariates into account, where the inclusion probability of a population unit refers to the chance that it becomes part of a sample. We base the Monte Carlo simulation models on e.g. data for occurrence of epiphytic hair lichens (Bryoria spp.) on Norway spruce (Picea abies) collected throughout Sweden in the National Forest Inventory (NFI). Hair lichens are widespread in boreal and temperate forests and are useful indicators for monitoring integrity of forest ecosystems (Esseen et al., 2016; Will-Wolf et al., 2002). Using NFI data for studies of this kind can be expected to be more common in the future due to the increased application of such data in decision-making and research (Tomppo et al., 2010). Thus it is important to assess how data from complex NFI designs should be utilized in modeling to avoid erroneous conclusions.

2. Materials and methods

2.1. Theory

A binary response variable is usually coded as a 1 or 0, and can be used to represent binary outcomes such as yes/no and presence/absence. The aim is to model a relationship between the probability that a response $Y$ is equal to 1 and a set of covariates. The ratio of the probability of $Y = 1$ to the probability of $Y = 0$ is known as the odds.

2.1.1. The standard logistic regression model

Let $\mathbf{\beta} = (\beta_0, ..., \beta_p)^\top$ denote a vector of $p + 1$ coefficients and $\mathbf{x} = (x_0, ..., x_p)$ a set of values of $p$ covariates and a constant, $x_0 = 1$. Let $p_1$ be the conditional probability that $Y = 1$ given $\mathbf{x}$. The SL regression model is then given by

$$\logit(p_1) = \mathbf{x} \cdot \mathbf{\beta},$$

(1)

where $\logit(p_1) = \log \left( \frac{p_1}{1 - p_1} \right)$, $\log()$ is the natural logarithm, and $p_1/(1-p_1)$ is the odds (of, e.g., occurrence). The odds ratio for two different groups is the ratio of the odds for the two different groups, and the coefficients $\beta_r$, $r = 1, ..., p$, in the SL model (1) are commonly interpreted in terms of odds ratios (Hosmer et al., 2013). If we are interested in the odds ratio for a one-unit increment in the $r$th covariate, while holding the other covariates fixed, then the odds ratio is $exp(\beta_r)$.

Eq. (1) may be rewritten as

$$p_1 = p_1(\mathbf{x}) = \frac{\exp(\mathbf{x} \cdot \mathbf{\beta})}{1 + \exp(\mathbf{x} \cdot \mathbf{\beta})},$$

(2)

The unknown $\mathbf{\beta}$ needs to be estimated from observed data. In SL regression, observations are assumed to be independent and $\mathbf{\beta}$ is estimated using maximum likelihood (ML). Given a sample of $n$ observations, with observed values of $x_i$ and $y_i$, $i = 1, ..., n$, the ML estimate of $\mathbf{\beta}$ is obtained by solving the likelihood equations

$$\frac{\partial L(\mathbf{\beta})}{\partial \mathbf{\beta}} = \sum_{i=1}^{n} y_i (1 - p_i) \mathbf{x}_i \cdot \mathbf{\beta} = 0,$$

(3)

$r = 0, ..., p$, with respect to $\mathbf{\beta}$ (Hosmer et al., 2013). Thus, in SL regression the effect of any clustering in the data is ignored.

2.1.2. The cluster-specific model

Suppose we have observations from $m$ clusters (e.g., sample plots), with $n_i$ observations (e.g., trees) per cluster. For observation $k$ in cluster $j$, let $Y_{jk}$ be the binary response and $x_{jk} = (x_{0jk}, ..., x_{pk})$ the corresponding covariate vector with $x_{0jk} = 1$. In the CS model, a random effect $\alpha_j$, specific to cluster $j$, is added to the logit. The CS model is thus defined as

$$\logit(p_{1jk}^{CS}) = \alpha_j + \mathbf{x}_{jk} \cdot \mathbf{\beta}^{CS},$$

where $\mathbf{\beta}^{CS} = (\beta_{1j}, ..., \beta_{pj})$ and $p_{1jk}^{CS}$ is the probability that $Y_{jk}$ is equal to 1 given the random effect $\alpha_j$ and $x_{jk}$, where $\alpha_1, ..., \alpha_m$ are assumed to be independent and normally distributed with mean zero and variance $\sigma^2$. We may regard $\alpha_j$ as the effect of being in cluster $j$, and an increase in $\alpha_j^2$ increases both the within-cluster correlation and the between-cluster heterogeneity (Hosmer et al., 2013). The odds ratio for a one-unit increment in the $r$th covariate, while holding the random effect and the other covariates fixed, is $exp(\beta_{rj}^{CS})$. Since each cluster has its own cluster-specific random effect, the interpretation of the effect of that covariate applies to a specific cluster, i.e. it is a within-cluster effect.

That random effects are unobservable leads to complications when considering ML estimation of $\mathbf{\beta}^{CS}$. This complication may be solved by numerically integrating out the random effects from the likelihood and then maximize it with respect to $\mathbf{\beta}^{CS}$ (Hosmer et al., 2013).

2.1.3. The population-averaged model

The PA probabilities are, in a sense, obtained by averaging the cluster-specific probabilities, $p_{1jk}^{CS}(\mathbf{\beta}^{CS}, \alpha_j, x_{jk})$, over all possible realizations of the random effects. More specifically, the PA probability is obtained as $p_{1jk}^{PA} = p_{1jk}^{PA}(x_{jk}) = E(p_{1jk}^{CS}(\mathbf{\beta}^{CS}, \alpha_j, x_{jk}))$, where the expectation is taken over the distribution of $\alpha_j$ (Hardin and Hilbe, 2012). Because the logit is a non-linear function, the PA probability, considered as a function of $x_{jk}$, will have a different shape than the CS probabilities (in fact, it does not exactly follow the logistic formula Agresti (2002)), and

$$p_{1jk}^{PA}(x_{jk}) = E\left(\frac{\exp(\mathbf{x}_{jk} \cdot \mathbf{\beta}^{CS})}{1 + \exp(\mathbf{x}_{jk} \cdot \mathbf{\beta}^{CS})}\right) = \frac{\exp(\mathbf{x}_{jk} \cdot \mathbf{\beta}^{CS})}{1 + \exp(\mathbf{x}_{jk} \cdot \mathbf{\beta}^{CS})}.$$

This is illustrated in Fig. 1 for the case of a single covariate, $x_{jk}$. Note that the effect of $x_{jk}$ on the PA probability is smaller than its effect on the CS probabilities, and that the population median probability (i.e., the median of the CS probabilities) is identical to the CS probability evaluated at the median of the random effect $\alpha_j$ (i.e., at $\alpha_j = 0$). The larger variance of the random effects, the more heterogeneity among the clusters and the less steep is the PA probability as a function of $x_{jk}$ (or more generally, of the respective covariates $x_{rjk}$, $r = 1, ..., p$).

In the PA approach, let $\mathbf{\beta}^{PA} = (\beta_{0k}, ..., \beta_{pk})$ and assume that

$$
The coeﬃcients in a PA model are typically estimated using generalized estimating equations (GEEs; Liang and Zeger, 1986). In addition to the working correlation, it is assumed that the correlation between diﬀerent observations is time-invariant. Under this assumption, the GEEs have the form

$$
\text{logit} \left( p_{jk}^{PA} \right) = x \beta^{PA},
$$

where we may interpret \( p_{jk}^{PA} \) as the proportion of units in the (assumed infinite) population with \( Y_{jk} = 1 \) among units with covariates equal to \( x_{jk}, \ldots, x_{jk} \) (Hosmer et al., 2013). The odds-ratio interpretation of the coeﬃcient vector \( \beta^{PA} \) in a PA model is analogous to the odds-ratio interpretation of a coeﬃcient from a SL regression model (Hosmer et al., 2013) and describes the eﬀect of the \( r \)th covariate in broad groups of units (cf. the CS model where we instead have cluster-speciﬁc interpretations).

The coeﬃcients in a PA model are typically estimated using generalized estimating equations (GEEs; Liang and Zeger, 1986). In addition to the working correlation, it is assumed that the correlation between \( Y_{jk} \) and \( Y_j \) is equal to a constant \( \rho \) for all \( k \neq l \) and all \( j \). Under an exchangeable correlation structure, the GEEs have the form

$$
\sum_{j=1}^{m} D_i V_{ji} S_j = 0,
$$

where \( D_i \) is an \( n_i \times (p + 1) \) matrix containing the elements \( x_{ijk}, \ldots, x_{ijk} \), \( V_j \) is an \( n_i \times n_i \) matrix with the diagonal elements \( p_{ijk}^{PA} (1 - p_{ijk}^{PA}) \) and the oﬀ-diagonal elements \( \sqrt{p_{ijk}^{PA} (1 - p_{ijk}^{PA})} \sqrt{p_{ilk}^{PA} (1 - p_{ilk}^{PA})} \), \( k \neq l \), and \( S_j \) is a vector containing the elements \( y_{jk} - p_{ijk}^{PA} \), where \( y_{jk} \) is the observed value of \( Y_{jk} \). Liang and Zeger (1986) suggest a moment estimator for \( \rho \), and to compute the estimates of \( \rho \) and \( \beta^{PA} \) one may iterate between solving the GEEs for \( \beta^{PA} \) and the moment estimation of \( \rho \) until convergence.

Under the assumption that the random effects are normally distributed, Zeger et al. (1988) (see also Hosmer et al., 2013) show that

$$
\hat{\beta}^{PA} \approx \frac{\hat{\beta}^{CS}}{\sqrt{1 + \left( \frac{16}{15} \right)^{1/2} \frac{\hat{\sigma}^2}{\tau^2}}}.
$$

where \( \hat{\beta}^{PA} \) and \( \hat{\beta}^{CS} \) are estimated coeﬃcient vectors and \( \hat{\sigma}^2 \) is the estimated random eﬀects variance. The implication of Eq. (6) is that the diﬀerence between \( \hat{\beta}^{PA} \) and \( \hat{\beta}^{CS} \) increases with the size of the random eﬀects variance.

Estimates from GEEs are known to be robust to choice of correlation structure (Liang and Zeger, 1986), i.e., the asymptotic normality of \( \hat{\beta}^{PA} \) and a so-called “sandwich-type” estimator of its asymptotic variance matrix are robust to misspeciﬁcation of the working correlation. Hosmer et al. (2013) recommend exchangeable correlation unless the nature of the data clearly suggests another choice; three other possible working correlation structures that can be speciﬁed in statistical packages are independent, auto-regressive, and unstructured (Wilson and Lorenz, 2015). Under an independent correlation structure, \( \rho = 0 \), and the GEEs (Eq. (5)) reduce to the likelihood equations (Eq. (3)) for SL regression. One simple approach to analyzing correlated binary responses is therefore to use SL regression, followed by a robust sandwich-type estimate of the variance matrix of the coeﬃcients. This approach can be expected to work well when the correlation between the responses is not too high (Zeger, 1988).

### 2.2. Survey sampling

The general principle of sampling is to select a subset of units (e.g., trees) from a population, and to draw inference from the sample to the entire population. Many widely used sampling methods assign varying inclusion probabilities to the individual units. The chance for a unit to be included in the sample is often determined by the values of several design variables, which may include cluster and stratum indicators and variables measuring characteristics of sample units (but usually not the response variable itself). The informativeness of a sampling design is the extent to which inclusion probabilities are dependent upon the response variable after conditioning on all other variables in the regression model. Ignoring an informative sampling design may yield severely biased estimates of regression coefficients, possibly leading to false inference in, e.g., SL regression (Nordberg, 1989; Pfeﬀermann and Sverchkov, 2009). Three commonly used approaches that account for informative sampling designs can be identiﬁed (Pfeﬀermann and Sverchkov, 2009):

- using design variables as additional covariates, or other variables associated with the design variables, which are highly correlated with the inclusion probabilities;
- using inclusion probabilities as additional covariates (which is an alternative when it is not possible to include all important design variables in the model due to parsimony, lack of data, or because the resulting model may not be of interest); and
- methods using probability weighting.

Probability-weighted estimators have the disadvantage of being typically more variable than unweighted estimators, notably when there is a large variation in the weights (Nordberg, 1989; Pfeﬀermann, 1993; Pfeﬀermann and Sverchkov, 2009). For complex sampling schemes where the inclusion probabilities are determined by the values of design variables, the informativeness of the sampling design is not always apparent. In this case, hypothesis testing can be used to test whether the sampling design can be ignored or not. Let \( \pi_i \) be the inclusion probability of the \( ith \) unit in the population. Testing that the sampling design can be ignored for estimating the coeﬃcient vector \( \beta \) in a generalized linear model (GLM), which includes the SL regression model as a special case, can be made using Nordberg’s (1989) test: i) augment the covariate vector by bringing in \( \pi_{ri}, r = 0, \ldots, p \), as additional covariates; ii) ﬁt the unweighted SL regression model using the augmented covariate vector; and iii) use a partial likelihood ratio test (Hosmer et al., 2013) to test the null hypothesis of jointly zero coeﬃcients of \( \pi_{ri}, r = 0, \ldots, p \).

Let us consider a probability-weighted SL regression. For a population of \( N \) units and by using the same notation as in Eq. (3), we deﬁne the population likelihood equations,

$$
\sum_{j=1}^{N} x_{ji}(y_{ij} - p_{ij} | \beta^{CS} x_{ij})) = 0,
$$

where \( p_{ij} = \text{logit}^{-1}(x_{ij} \beta^{CS}) \) is the population-average probability of \( y_{ij} \) and \( x_{ij} \) is the \( j \)th covariate for the \( i \)th unit in the population.
\[ r = 0, \ldots, p, \quad \text{where } p_1 (\beta | x_i) \text{ is given by Eq. (2); cf. Eq. (3). Let } \beta_{SE}, \text{ denote the solution of Eq. (7). Based on a survey sample } s, \text{ the unknown population coefficient vector } \beta_0 \text{ is estimated by a coefficient vector that solves the probability-weighted estimate of Eq. (7),}
\]
\[
\sum_{i \in s} w_i x_i (y_i - p_1 (\beta | x_i)) = 0,
\]
(8)
\[ r = 0, \ldots, p, \quad \text{where } w_1 = 1/\pi_i^h (\text{Heeringa et al., 2010). The left-hand side of Eq. (8) is design-unbiased for the left-hand-side of Eq. (7) for any sampling design, including informative ones, and the solution of Eq. (8) is approximately unbiased for the corresponding population quantity } \beta_{SE}, (\text{Heeringa et al., 2010). That is, the mean of the solutions of Eq. (8), calculated from repeated independent samplings of the population, approximately equals } \beta_{SE}.
\]

As an example of a complex sampling design, let us consider a so-called stratified two-stage sampling design, where the population of units is partitioned into L strata. In stratum \( h \), the units are partitioned into \( M_h \) disjoint clusters (e.g., sample plots), denoted by \( C_1, \ldots, C_{M_h} \), and in the first sampling stage, a sample \( s_h \) of clusters is drawn. In the second stage, a sample \( s_{kh} \) of units (e.g., trees) is drawn from each cluster \( j \in S_h \). If \( \pi_{kj} \) is the probability that cluster \( j \in S_h \) and \( \pi_{shj} \) the conditional probability that unit \( k \in S_h \) given that cluster \( j \in S_h \), then \( \pi_{kj} = \pi_{kj} \pi_{shjk} \) is the inclusion probability of unit \( k \) in cluster \( j \) of stratum \( h \) (Särndal et al., 1992). It is assumed that every time \( j \in S_h \), exactly the same sampling design is used for sampling units within that cluster, and that sampling in cluster \( j \in S_h \) is carried out independently of sampling in any other cluster.

Consider the PA model under the stratified two-stage sampling design. For \( k \in S_{sh} \), let \( y_{shk} \) be a binary variable with possible values 0 and 1 and \( x_{shk} = (x_{shk1}, \ldots, x_{shkF}) \) a vector of covariates. The population GEEs are defined as \( \sum_{h=1}^{H} \sum_{j=1}^{M_{sh}} \mathbf{D}_{shj} \mathbf{V}_{shj}/\pi_{shj} = 0 \), where \( \mathbf{D}_{shj} \), \( \mathbf{V}_{shj} \), and \( \pi_{shj} \) are defined as in Eq. (5), but with \( y_{shk}, x_{shk}, \pi_{shk}, \) and \( \eta \) replaced by \( y_{shk}, x_{shk}, \pi_{shk}, \) and \( N_{sh} \), respectively, where the latter is the size of cluster \( j \) in stratum \( h \). Under an exchangeable correlation structure with a given \( \rho \), let \( \beta_{PA} \) denote the solution of the population GEEs. Based on a stratified two-stage sample, the population coefficient vector \( \beta_{PA} \) is estimated by solving the probability-weighted estimate of the population GEEs, \( \sum_{h=1}^{H} \sum_{j=1}^{M_{sh}} w_{shj} \mathbf{D}_{shj} \mathbf{V}_{shj}/\pi_{shj} = 0 \), where \( w_{shj} = w_{shj} w_{shjk}/\pi_{shjk} \). \( w_{shj} = 1/\pi_{shj} \) and \( w_{shjk} = 1/\pi_{shjk} \). Under related settings and an exchangeable correlation structure, (Rao 1998) and Lipsitz et al. (2014) suggest the unknown \( \rho \) in the aforementioned GEEs to be estimated using an estimated coefficient vector obtained under an independent correlation structure.

In contrast to probability-weighted estimation of a PA model under a stratified two-stage sampling design, the CS model requires knowledge of not only the products \( w_{shjk} = w_{shj} w_{shjk}/\pi_{shjk} \), but also the individual factors \( w_{shj} \) and \( w_{shjk} \). A (suitably) scaled version of \( w_{shjk} \) is required when integrating out the random effects from the likelihood, and \( w_{shj} \) weights the \( j \)th cluster contribution to the log-likelihood in stratum \( h \); see Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) for details.

The above reasoning for PA and CS models can easily be modified to other designs than the stratified two-stage sampling design. To our knowledge, there are currently no packages in R (R Core Team, 2016; the software used in our Monte Carlo simulations, see below) for fitting PA and CS models in logistic regression settings under complex sampling designs such as the stratified two-stage sampling design.

2.3. Data for the Monte Carlo simulations

Data for the Monte Carlo simulations were obtained from the Swedish NFI, which provides information about forests for regional, national and international policy, planning, and reporting (Fridman et al., 2014). The NFI has been operating since 1923 and at present > 200 variables are recorded, including detailed forest data, soil characteristics, presence and cover of different forest-floor plants as well as occurrence and length of hair lichens on Picea abies, recently being used in several ecological studies (e.g. Esseen et al., 2016; Hedwall and Brunet, 2016). The NFI covers all forests in Sweden (55–69°N) except subalpine birch forests in the Scandinavian mountains. The design includes both stratification (Fig. 2) and clustering of sample plots into square-formed tracts. The length of tract side varies from 300 to 1800 m among regions.

Data were compiled from 27,989 permanent plots (10 m’s radius, area 314 m²) from the period 1993–2014 (more than two full inventory cycles). We included plots in productive forest land (site quality as mean volume increment \( \geq 1 \text{ m}^3 \text{ ha}^{-1} \text{ yr}^{-1} \)) and excluded plots classified as unstocked, thicket or young stands to simplify model construction (cf. Esseen et al., 2016). This resulted in 21,168 remaining plots. The occurrence of the lichen genus Bryoria (Fig. 3) was recorded.

Fig. 2. The strata used in the Swedish NFI. Strata roughly correspond to the vegetation zones: 1) northern boreal; 2) mainly middle-northern boreal; 3) southern-middle boreal; 4) mainly hemiboreal; and 5) temperate.
Therefore, for each of the 115,887 only for the Poisson sample trees or some of the Poisson sample trees. In the 21,168 plots, we generated values of the variables crown limit as covariates. Details are given in the supplementary material (Appendix B), which includes formulas for calculating $\beta_{PA}$. Design 2 is essentially the one used in the Swedish NFI since 2013, but we ignore that plots are grouped into tracts. Design 2 is identical to design 1, except that occurrence of Bryoria is registered for all trees in the candidate set. Thus, for design 2, $\beta_{PA}$ is given by Eq. (9).

In addition to logistic regression models of Bryoria occurrence, we also considered such models for the binary response variable crown limit above/below a height of 5 m. In each Monte Carlo simulation, 10,000 samples were drawn according to designs 1 and 2, respectively, and for each sample, model coefficients and corresponding standard errors were estimated. In each sample, 20% of all sample plots were drawn in each of the five strata. For comparison, we also fitted logistic regression models using all 115,887 $P$. abies trees in the population. Whenever a PA logistic regression model was fitted, an exchangeable correlation structure was used.

All Monte Carlo simulations have been conducted using the R software (R Core Team, 2016), and the R packages survey (Lumley, 2014), geepack (Højsgaard et al., 2006), and lme4 (Bates et al., 2015). As earlier mentioned, to our knowledge there are currently no packages in R for fitting PA and CS models in logistic regression settings under complex sampling designs. Consequently, probability-weighted estimates were computed only in SL settings. Unweighted SL models were fitted using the glm function in R, and for the probability-weighted ones we used the svyglm function from the survey package. The function syntax of svyglm is almost identical to that of glm, except that the data argument of glm is replaced by a sampling design argument, which is created using the svydesign function. Our PA models were fitted using the geeglm function from geepack and for CS models we used the glmer function from lme4. To specify logistic regression, rather than the default of linear regression, one uses the option family = binomial in the glm, svyglm, geeglm, and glmer functions. For data examples with complete R code, we refer to Lumley (2010, Chapter 6) for svyglm examples, Zuur et al. (2009, Chapter 12) for geeglm examples, Bates (2010, Chapter 6) for glmer examples, and Wilson and Lorentz (2015, Chapters 5, 6, and 9) for svyglm, geeglm, and glmer examples.

3. Results

3.1. Example 1. Occurrence of Bryoria spp.

The obtained population coefficient vectors from the CS, PA, and SL regression models are denoted by $\beta_{CS}$, $\beta_{PA}$, and $\beta_{SL}$, respectively (Table 1a). In each model we used the same covariates as in the superpopulation (Table A.2, Appendix A, supplementary material). In the CS model, the estimated variance $\sigma^2_j$ of the random effects is 8.55, and if we compare with Table A.2 we see, as expected, that $\beta_{CS}$ is close to the corresponding superpopulation coefficient vector. In Table 1a, $\beta_{PA}$ is

\[
q_{DBH_k} = \min \left\{ 1, \frac{a + b \cdot DBH_k^2}{10,000} \right\},
\]

where DBH$_k$ (in mm) is the diameter at 1.3 m of tree $k$, and

\[
a \approx \begin{cases} 
-3.9 & \text{in strata 1 and 2}, \\
40.5 & \text{in strata 2 and 3}, \\
63.3 & \text{in strata 4 and 5}, 
\end{cases}
\]

and

\[
b \approx \begin{cases} 
0.0029 & \text{in strata 1 and 2}, \\
0.0022 & \text{in strata 2 and 3}, \\
0.0019 & \text{in strata 4 and 5}. 
\end{cases}
\]

If the set of candidate trees is non-empty, then a single sample tree is selected with equal probability from the candidate set, and occurrence of Bryoria is registered for the sample tree. This design may be regarded as stratified two-stage sampling; see the supplementary material (Appendix B).
close to $\beta_{SL}$, and the coefficients in $\beta_{CS}$ are about twice as large as those in $\beta_{PA}$, which agrees well with formula (6) in which the square root expression on the right-hand side is 1.99.

**Design 1:** For each sample, we obtained the unweighted SL regression estimate $\hat{\beta}_{SL}$ and its estimated standard error $\hat{\sigma}(\hat{\beta}_{SL})$ as well as the corresponding probability-weighted estimate $\hat{\beta}_{PA}$ and its estimated standard error $\hat{\sigma}(\hat{\beta}_{PA})$. The respective averages of these estimates are given in Table 1b, together with standard deviations of $\hat{\beta}_{CS}$ and $\hat{\beta}_{SL}$, respectively (averages and the standard deviations are taken over the 10,000 replications). Table 1a and b show that $\hat{\beta}_{PA}$ and $\hat{\beta}_{CS}$ give nearly unbiased estimates of $\beta_{PA}$ and $\beta_{CS}$, respectively. If we compare $\text{ave}(\hat{\beta}_{CS})$ to $\text{std}(\hat{\beta}_{CS})$ and $\text{ave}(\hat{\sigma}(\hat{\beta}_{CS}))$ to $\text{std}(\hat{\sigma}(\hat{\beta}_{CS}))$, we conclude that the estimated standard errors are approximately unbiased and that $\hat{\beta}_{CS}$ is less variable (i.e., more efficient) than $\hat{\beta}_{CS}$. For each replication, we tested if the sampling design can be ignored for estimating the coefficient vector $\beta_{CS}$ using Nordberg's (1989) test. A histogram revealed that the obtained P-values of Nordberg's test follow, approximately, a uniform distribution on the interval (0, 1), which suggests that the sampling design is not ignorable for estimating $\beta_{CS}$. Note, in this example, the covariates in the SL models were the same as those in the superpopulation model that generated the population of responses, implying that the sampling design is non-informative (with respect to the superpopulation model).

**Design 2:** From Table 1a and c it follows that not only the probability-weighted estimator $\hat{\beta}_{PA}$ gives a nearly unbiased estimate, but that also $\hat{\beta}_{PA}$ and $\hat{\beta}_{CS}$ give approximately unbiased estimates of their respective population coefficients in Table 1a. The estimator $\hat{\beta}_{CS}$ is somewhat biased, most notably for the intercept and the coefficient for temperature. In Table 1c (and other tables to follow), we add the subscript r to $\hat{\sigma}$ in cases where a robust sandwich-type estimator of standard error has been used for an unweighted estimator of coefficients. In the table, note that $\text{ave}(\hat{\sigma}(\hat{\beta}_{PA})) < \text{std}(\hat{\beta}_{PA})$, i.e., the standard errors of the coefficients in $\beta_{PA}$ are all underestimated by the classical estimator. In comparison, we do not see this tendency for the robust estimator $\hat{\sigma}(\hat{\beta}_{PA})$. The estimated standard errors for $\beta_{CS}$, $\beta_{PA}$, and $\beta_{SL}$ are all approximately unbiased. It should be noted that $\text{std}(\hat{\beta}_{PA}) \approx \text{std}(\hat{\beta}_{SL})$, indicating that the PA estimator is (slightly) more efficient than the SL estimator.

### Table 1

Coefficients for logistic models fit to the lichen occurrence data: (a) Population coefficients; in the cluster-specific (CS) model, $\beta_{CS} \approx 8.55$, and in the population-averaged (PA) model, $\beta_{PA} \approx 0.43$; (b) Monte Carlo estimates for design 1; and (c) Monte Carlo estimates for design 2. Ave denotes sample average and std standard deviation.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\hat{\beta}_{CS}$</th>
<th>$\hat{\beta}_{PA}$</th>
<th>$\hat{\beta}_{SL}$</th>
<th>Covariate</th>
<th>$\text{Ave}(\hat{\beta}_{CS})$</th>
<th>$\text{Ave}(\hat{\sigma}(\hat{\beta}_{CS}))$</th>
<th>$\text{Std}(\hat{\beta}_{CS})$</th>
<th>$\text{Ave}(\hat{\beta}_{PA})$</th>
<th>$\text{Ave}(\hat{\sigma}(\hat{\beta}_{PA}))$</th>
<th>$\text{Std}(\hat{\beta}_{PA})$</th>
<th>$\text{Ave}(\hat{\beta}_{SL})$</th>
<th>$\text{Ave}(\hat{\sigma}(\hat{\beta}_{SL}))$</th>
<th>$\text{Std}(\hat{\beta}_{SL})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.89</td>
<td>5.03</td>
<td>5.00</td>
<td>Intercept</td>
<td>5.00</td>
<td>0.41</td>
<td>0.39</td>
<td>5.02</td>
<td>0.50</td>
<td>0.49</td>
<td>5.00</td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td>(TEMP + 3.3)/5$^2$</td>
<td>$-2.67$</td>
<td>$-1.33$</td>
<td>$-1.34$</td>
<td>(TEMP + 3.3)/5$^2$</td>
<td>$-1.34$</td>
<td>0.07</td>
<td>0.07</td>
<td>$-1.36$</td>
<td>0.08</td>
<td>0.08</td>
<td>$-1.36$</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>BA/10</td>
<td>$-1.03$</td>
<td>$-0.54$</td>
<td>$-0.53$</td>
<td>BA/10</td>
<td>$-0.53$</td>
<td>0.10</td>
<td>0.09</td>
<td>$-0.53$</td>
<td>0.12</td>
<td>0.12</td>
<td>$-0.53$</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>(AGE/100)$^{-1}$</td>
<td>$-1.45$</td>
<td>$-0.75$</td>
<td>$-0.75$</td>
<td>(AGE/100)$^{-1}$</td>
<td>$-0.76$</td>
<td>0.11</td>
<td>0.10</td>
<td>$-0.75$</td>
<td>0.12</td>
<td>0.12</td>
<td>$-0.75$</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>log(CL + 0.1)/10</td>
<td>$-1.36$</td>
<td>$-0.69$</td>
<td>$-0.74$</td>
<td>log(CL + 0.1)/10</td>
<td>$-0.76$</td>
<td>0.33</td>
<td>0.33</td>
<td>$-0.76$</td>
<td>0.41</td>
<td>0.42</td>
<td>$-0.76$</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>(log(CL + 0.1)/10)$^2$</td>
<td>$-0.69$</td>
<td>$-0.35$</td>
<td>$-0.37$</td>
<td>(log(CL + 0.1)/10)$^2$</td>
<td>$-0.38$</td>
<td>0.15</td>
<td>0.14</td>
<td>$-0.38$</td>
<td>0.18</td>
<td>0.18</td>
<td>$-0.38$</td>
<td>0.18</td>
<td>0.18</td>
</tr>
</tbody>
</table>

$^a$ TEMP = temperature, BA = basal area, AGE = stand age, CL = crown limit.

#### 3.2. Example 2. Crown limit above/below 5 m.

Based on a sample according to design 1, the best model in terms of AIC (Akaike information criterion) was the one with stand height, DBH, volume of *P. abies*, as well as the interaction term of the latter two covariates (the model was not improved by adding the inclusion probability $\pi_{ijk}$ as a covariate or any interaction terms involving $\pi_{ijk}$, and Nordberg’s test of whether the sampling design can be ignored for estimating the coefficient vector $\hat{\beta}_{CS}$ gave a P-value equal to 0.28.) The same covariates were selected based on a sample according to sampling design 2, where the drawn sample plots were identical to those in the former sample. For sake of comparison, we also consider models without the design variable DBH, i.e., models with only stand height and volume of *P. abies* as covariates. The obtained population coefficients are given in Tables 2a and 3a for the models with and without the design variable DBH, respectively.

**Design 1:** For each sample, we obtained the estimates $\hat{\beta}_{CS}$ and $\hat{\beta}_{SL}$, and their corresponding estimated standard errors; see Tables 2b and 3b for the models with and without DBH, respectively. For both models, we see that $\hat{\beta}_{SL}$, in contrast to $\hat{\beta}_{CS}$, gives a somewhat biased estimate of the corresponding population coefficient $\beta_{SL}$, which indicates that the sampling design is not ignorable. Nordberg’s test rejects the null hypothesis, which states that the sampling design can be ignored for estimating the coefficient vector $\beta_{CS}$. For each sample, we obtained the unweighted SL regression estimate $\hat{\beta}_{SL}$ using Nordberg’s (1989) test. A histogram revealed that the obtained P-values of Nordberg’s test follow, approximately, a uniform distribution on the interval (0, 1), which suggests that the sampling design is ignorable for estimating $\beta_{CS}$.
are smaller than those of \( \hat{\beta}_{SL} \), and in the population-averaged (PA) model, \( \hat{\beta}_{PA} \sim \beta_{Ave} \). In the population-averaged (PA) model, \( \hat{\beta}_{PA} \sim \beta_{Ave} \), and in the population-averaged (PA) model, \( \hat{\beta}_{PA} \sim \beta_{Ave} \).

\[ \hat{\beta}_{SL} \approx \beta_{Ave} \]

To study the effect of the strength of correlation between responses within clusters, we generated binary responses for each tree in the population according to the fitted CS model in Table 2a, but with a random effects variance equal to 8.39 (rather than 0.98 as specified in Table 2a). Based on these new and more strongly correlated responses, we performed Monte Carlo simulations using the same setup as was used in Tables 2–3. The corresponding results for the more strongly correlated responses are presented in Tables C1–C2 in the supplementary material (Appendix C). The most striking difference when increasing the correlation between responses is that the coefficients in the fitted models, except those in the CS models, are shrunk towards zero. The conclusions drawn from Tables 2–3 remain valid for Tables C1–C2, with the following exceptions: i) In contrast to the unweighted estimators in Table 2b and c the corresponding estimators in Tables C1b and C1c are at most mildly biased (cf. Tables 3b, c, 3b, and C2c, where the design variable DBH is missing in the model, resulting in biased unweighted estimators); ii) With the new response, Nordberg’s test rejects the null hypothesis in 4% and 66% of the replications for the unweighted estimators; iii) With the new response, Nordberg’s test rejects the null hypothesis in 4% and 66% of the replications for the unweighted estimators; iv) With the new response, Nordberg’s test rejects the null hypothesis in 4% and 66% of the replications for the unweighted estimators; v) The standard errors of the coefficients in \( \hat{\beta}_{SL} \) are underestimated by the classical estimators, and our examples have illustrated that ignoring such spatial correlations may result in underestimation of standard errors of coefficient estimates. For unit-level (tree-level) covariates, Fitzmaurice et al. (1993) argue that the standard errors may tend to be overestimated, but we have not seen this happen in our study. For avoiding the problem of biased standard errors, PA and CS logistic regression models that take the spatial correlation into account are useful. We have, as expected, seen that the absolute values of the CS coefficient estimates are generally larger than the corresponding PA estimates, and that the differences increase with the heterogeneity of the random

\[ \hat{\beta}_{CS} \text{ are smaller than those of } \hat{\beta}_{SL}. \]

4. Discussion

We have investigated the need for taking a complex sampling design into account when analyzing spatially correlated binary response variables from large-scale surveys using logistic regression models. Our study has been supported by illustrative Monte Carlo simulations of, e.g., occurrence of Bryoria spp. in the lower canopy of P. abies. These simulations show that if a complex sampling design is used to draw a sample from a population, then unweighted estimators can be notably biased, whereas weighted estimators of regression coefficients and standard errors are approximately unbiased. Furthermore, in ecological studies spatially close measurements are often more similar than distant ones, and our examples have illustrated that ignoring such spatial correlations may result in underestimation of standard errors of coefficient estimates. For unit-level (tree-level) covariates, Fitzmaurice et al. (1993) argue that the standard errors may tend to be overestimated, but we have not seen this happen in our study. For avoiding the problem of biased standard errors, PA and CS logistic regression models that take the spatial correlation into account are useful. We have, as expected, seen that the absolute values of the CS coefficient estimates are generally larger than the corresponding PA estimates, and that the differences increase with the heterogeneity of the random

Table 3

Coefficients for logistic models fit to the crown limit data (without the covariate DBH): (a) Population coefficients; in the (CS) cluster-specific model, \( \hat{\beta}_{CS} \approx 0.98 \), and in the population-averaged (PA) model, \( \hat{\beta} \approx 0.15 \); (b) Monte Carlo estimates for design 1; and (c) Monte Carlo estimates for design 2. Ave denotes sample average and std standard deviation.

(a)                        (b)                        (c)

| Covariate | \( \hat{\beta}_{CS} \) | \( \hat{\beta}_{PA} \) | \( \hat{\beta}_{SL} \) | Ave(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) | Ave(\( \hat{\beta}_{PA} \)) | Ave(\( \hat{\beta}_{SL} \)) | Std(\( \hat{\beta}_{CS} \)) |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Intercept     | −5.96          | −5.10          | −4.84          | −4.93          | 0.32           | −4.64          | 0.23           | 0.22           | −4.64          | 0.21           | 0.23           | 0.22           | −4.85          | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           | 0.27           |
| H/10          | 1.85           | 1.57           | 1.50           | 1.46           | 0.17           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           | 1.51           | 0.21           | 0.21           |
| VOLPA/100     | 1.04           | 0.92           | 0.87           | 0.89           | 0.07           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           | 0.88           | 0.09           | 0.09           |

H = stand height, VOLPA = volume of P. abies.
effects. CS coefficients also have different interpretations than corresponding PA and SL coefficients.

The CS model is most useful when one wants to provide inferences for covariates that change within clusters (Hosmer et al., 2013). For such a covariate, its coefficient has a cluster-specific interpretation. Thus, the effect is a “within-cluster” one. A CS model is less useful for covariates that never change within clusters/plots (cf. Feiberg et al., 2009). An example of the latter is a comparison of montane and non-montane forests using a dummy covariate. For such a covariate, we cannot make a “within-cluster” odds-ratio interpretation of its coefficient. Instead, the odds-ratio interpretation represents a comparison of a montane forest plot and a non-montane forest plot with the same value of their random effects, i.e. the coefficient has a “between-cluster” interpretation. However, as two plots with exactly the same random effect do not typically exist, this “between-cluster” interpretation is not likely to be useful in practice. By contrast, effects in PA models are averaged over all clusters, so those effects do not refer to a comparison at a fixed value of a random effect. Thus, for covariates that are constant within clusters, a PA model is likely to be more useful (Hosmer et al., 2013).

An important property of PA (and SL) regression is that the coefficients can be estimated consistently, even if the dependence structure is not properly modeled (Rabe-Hesketh and Skrondal, 2012). In addition, if the degree of dependence between binary responses within clusters is small enough, then SL regression coefficient estimators are about as efficient as PA estimators that take the correlation into account (Tables 2c and 3c). Additional simulations, not presented here, suggest that this remains valid for correlations up to about \( \rho = 0.3 \). However, even if the degree of dependence is weak, we recommend estimation of standard errors of SL regression coefficients that take the dependence into account (otherwise we may draw wrong conclusions about the significance of the covariates). As illustrated in our Monte Carlo simulations, this may be accomplished by using a robust sandwich-type estimator of variance. In SL and PA models, this robustness property is, however, an asymptotic property, meaning that the sandwich estimator may be unreliable unless there are a large number of sampled clusters. In our simulations, the number of sampled clusters was \( > 4000 \), so this was not an issue. In addition, the sandwich estimator is implicitly relying on the assumption that there are many replications of the response associated with each distinct set of covariate values (Rabe-Hesketh and Skrondal, 2012). Since all our regression models included several continuous covariates this assumption is not quite true. However, the sandwich estimator gave nearly unbiased estimates of standard error in all our simulation examples.

A CS model more fully describes the structure of the correlated binary data, and is usually a good choice unless the research question (i.e., the desired coefficient interpretations) requires a PA model (cf. Diggle et al., 1994; Lee and Nelder, 2004; Muff et al., 2016). For example, an advantage of the CS model is that PA effects can be estimated using marginalization (Lee and Nelder, 2004), and the “marginal coefficients” of fitted CS models are very similar to those from a PA model (Hu et al., 1998). Conversely, it is impossible to derive estimates of CS coefficients from a PA model. Still, the choice between CS and PA is usually not crucial to inferential conclusions (Agresti, 2007). That is, if one effect is judged more important than another in a CS model, the same is usually true also in a PA model. An advantage of PA models is that they require less restrictive assumptions for making consistent inferences about coefficients (Heagerty and Kurland, 2001; Hu et al., 1998; Zeger, 1988). For example, in CS models the random effects distribution needs to be correctly specified for consistent inferences (Zeger, 1988). In the PA model, there is no need to specify this distribution, and Liang and Zeger’s (1986) GEE approach yields consistent estimators of coefficients and standard errors even if the assumed working correlation is misspecified.

If the design does not contain any “relevant” information on the response which is not accounted for by the covariates, then an unweighted (SL, PA, or CS) model is to prefer. This is because an unweighted model is easier to understand, explicitly describes the relationship between the probability of response value one and a set of covariates, and, as illustrated in our Monte Carlo simulations, has estimated coefficients with smaller standard errors (cf. Lohr and Liu, 1994). It may be difficult to verify this sufficient condition, but comparisons of unweighted and weighted estimates of coefficients can be a useful tool for improving (if needed) the specification of the model (Korn and Graubard, 1995; Lohr and Liu, 1994; Reiter et al., 2005).

When the sampling design is ignorable, the unweighted and weighted estimates should be similar, since both estimators are consistent (Pfeffermann, 1993). If, however, the unweighted and weighted coefficient estimates are notably different (and especially if the difference is of practical importance), then this indicates a need for further statistical investigation (Lohr and Liu, 1994). In this case, a plausible explanation is that the model may be missing important covariates correlated with the inclusion probabilities or interaction terms involving such covariates.

5. Conclusions

Solid statistical models based on large-scale surveys have a great potential to increase our understanding of factors influencing the occurrence of species and specific habitats. The kind of models investigated in our study can, for instance, be used to explore the effects of climate change or modified forest management methods on the distribution of target species, such as hair lichens. However, our study emphasizes the need to at least explore the effects of complex sampling when applying logistic regression models in large-scale surveys. We also stress the importance of taking spatial correlations into account in the modeling. Ignoring these correlations in the data and/or the sampling design can lead to erroneous conclusions. In (SL, PA, or CS) logistic regression for spatially correlated data collected using complex sampling designs, we recommend performing both unweighted and weighted analyses, and the more the sampling mechanism deviates from simple random sampling the more important this is. If there is a notable difference between the two sets of estimated coefficients, we recommend resolving the discrepancy by adding covariates related to the inclusion probabilities (cf. Nordberg, 1989; Lohr and Liu, 1994; Reiter et al., 2005). If i) the discrepancy cannot be resolved by modifying the model, ii) it is not possible to add all important covariates due to parsimony, and/or iii) the modified model is not “ecologically relevant”, then the inclusion probabilities may contain information on the response which is not accounted for by the covariates in the proposed unmodified model. In this case, when a probability-weighted analysis is presented, the analyst should state that the inclusion probabilities may contain such information. A probability-weighted analysis may also be preferred if it entails only a small loss of efficiency (cf. Korn and Graubard, 1991).

Acknowledgments

This work was supported by the Swedish Research Council (grant 340-2013-5076).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2017.10.006.

References
