Grid-spacing and the quality of abundance maps for species that show spatial autocorrelation and zero-inflation

Olga Lyashevska[a,*], Dick J. Brus[b], Jaap van der Meer[a]

[a]*Department of Marine Ecology*
NIOZ Royal Netherlands Institute for Sea Research
P.O. Box 59 1790 AB Den Burg
Texel, The Netherlands
[b]*Alterra, Wageningen University and Research Centre*
P.O. Box 47, 6700AA
Wageningen, The Netherlands

Abstract

The effect of grid-spacing on the quality of species abundance maps is explored for species that show zero-inflation and spatial autocorrelation. Using a zero-inflated Poisson mixture model multiple fields of the prevalence parameter π and the intensity parameter μ were simulated. A selected field was sampled by grid-sampling with 200, 400, 800, 1600, and 3200 m grid-spacing and used to predict at a fixed set of validation locations by simple kriging with an external drift. The external drift variables were silt, silt squared and altitude. The estimated sampling distribution of MSE against grid-spacing shows that beyond a spacing of 1600 m the mean of MSE increases at a much faster rate. Based on these findings the 1600 m grid which consists of 446 locations for our study area of 2400 km^2 gives a compromise between sampling costs and prediction accuracy.

Keywords: count data; generalized linear geostatistical modeling; autocorrelation; zero-inflation; grid-spacing

1. Introduction

The relationship between species and their environment is generally described by species distribution models~~, also known as~~ ; in particular, by habitat suitability or environmental niche models (Guisan and Thuiller, 2005). Such models are constructed using survey data available at a limited set of sampling locations and allow one to create predictive species distribution maps on the basis of environmental data which are usually available ~~at~~ for a much larger set of locations (Guisan and Zimmermann, 2000). The number of sampling locations is known to affect the accuracy of the species distribution models and maps (Stockwell and Peterson, 2002; Wisz et al., 2008). Knowledge of the trade-off function between the number of sampling locations and the accuracy of the predictions is usually not obtained a priori~~with the risk that the survey either does not~~ . Neither does the survey provide the required accuracy when the number of sampling locations is too low

[*] Corresponding author
Email addresses: olga@herenstraat.nl (Olga Lyashevska), Dick.Brus@wur.nl (Dick J. Brus), Jaap.van.der.Meer@nioz.nl (Jaap van der Meer)

Preprint submitted to Spatial Statistics *June 9, 2016*

or is unnecessarily costly when the number of sampling locations is too high (Caughlan and Oakley, 2001; Reynolds et al., 2011).

~~Yet, the effect~~ Several studies evaluated effects of sample size on the accuracy of species distribution models ~~was evaluated by few (see e.g. Stockwell and Peterson, 2002; Pearson et al., 2007; Wisz et al., 2008; Hanberry et al., 2012)~~ (S
For example, Stockwell and Peterson (2002) assessed sample size requirements for modelling bird species in Mexico by random sampling between 1 ~~to~~ and 100 locations. Wisz et al. (2008) considered three sample sizes (10, 30, and 100 locations) to evaluate the quality of model predictions using data for 46 species obtained from natural history collections. Finally, Hanberry et al. (2012) used sample sizes ranging from 30 to 2500 locations to model tree species in northeasten Minnesota. All these studies consider presence–absence maps, but often, predictive species abundance maps in the form of numerical or biomass density are to be preferred, because they are more informative than presence–absence maps (Vieira et al., 2012; Cozzi et al., 2013). Fortin et al. (1989) constructed such maps using sugar-maple tree density data gathered in southwestern Québec. The authors evaluated the ability to predict spatial patterns using different sample sizes and designs. They considered two sample sizes of 50 and 64 points, both derived from a 200-point dataset.

Using real datasets only, as Fortin et al. (1989) did, ~~the possibilities of studying the effect of the type of sampling design and sample size are limited. This limitation was~~ limits comparison between the effects of different sample designs and sample number sizes as well as uncertainties in the model's parameter values. These limitations were recognised in recent studies, such as those by Perner and Schueler (2004); Rachowicz et al. (2006); Bijleveld et al. (2012) and Foster et al. (2014). ~~These studies follow an approach whereby first~~ An updated approach is to first simulate a spatial field resembling reality as much as possible, a peseudo-reality, ~~is simulated, which is~~ which is then subsequently sampled using different sampling designs. The performance of sampling designs is then compared by confronting predictions with simulated values that serve as ground-truth. Zurell et al. (2010) call this the virtual ecologist approach. Following this approach, Bijleveld et al. (2012) ~~, for example,~~ used the results of an existing intertidal benthic monitoring programme to construct various spatial models with an exponential spatial autocorrelation function. With these models they simulated virtual populations with a Normal distribution and sampled these populations using different sampling designs. They provided a trade-off function between sampling distance and prediction error which was rather flat for those virtual species that hardly showed spatial autocorrelation, but much steeper for species with strong spatial autocorrelation. The ~~authors~~ assumed normality of the data ~~which~~ was clearly violated by the empirical data ~~if only~~ because of the many zero observations.

The assumption of normality is a common practice, ~~but when we are~~ because when dealing with species abundance (count) data ~~, even~~ the more obvious Poisson distribution is rarely applicable. Ecological count data have two properties that ask for a specific treatment, other than relying on the classical assumption of independent and normally distributed data. These properties are zero-inflation (Martin et al., 2005; Clarke and Green, 1988; Lewis et al., 2011) and spatial autocorrelation, i.e. nearby observations are more similar than observations far apart, even when environmental conditions do not differ. Hitherto most studies have dealt with these two properties, but only one at a time (Tyre et al., 2003; Bijleveld et al., 2012).

For example, the first property was accounted by Tyre et al. (2003) who considered the zero-inflated negative binomial model, but ignored autocorrelation. Ignoring spatial autocorrelation in simulation studies on how sampling designs affect the accuracy of estimates of population- or model parameters or the accuracy of spatial predictions, may lead to biased estimates of this accuracy (Legendre et al., 2002).

Contrary to Tyre et al. (2003), the Bijleveld et al. study took account of the autocorrelation by using a stochastic model that included spatial autocorrelation of the error. But, as ~~we~~ mentioned earlier, they simulated normally distributed data. Clearly, there is a need to integrate both properties in a single study and to examine how zero-inflation and autocorrelation may affect recommendations ~~concerning~~ for the optimal sampling design, sample size, and distance between samples.

~~Only few studies~~ Studies that simultaneously address zero-inflation and autocorrelation for species abundances (see e.g. Recta et al., 2012; Boyd et al., 2015) ~~, but none of them address~~ do not treat the question of optimal sampling design. We attempt to fill this gap by following a paper by Lyashevska et al. (2016) in simulating fields with zero-inflated, spatially autocorrelated count data, and sampling the fields repeatedly with different sampling designs. More specifically, we will sample the fields by grid-sampling with a varying spacing. The ~~objective~~ aim of this paper is to quantify the trade-off between ~~the~~ grid-spacing and ~~the~~ accuracy of predictions of species-abundance model parameters on a fine grid for mapping, for species that show zero-inflation and spatial autocorrelation. Most species will show these two properties (see Martin et al., 2005, and references therein) .

2. Materials and methods

2.1. Data

Data used in this paper were zero-inflated (66% are zeros) and autocorrelated counts of a benthic species *Macoma balthica* (Fig. 1a) that were collected in the yearly Synoptic Intertidal Benthic Surveys (SIBES) monitoring programme conducted in the Dutch Wadden Sea (Bijleveld et al., 2012; Compton et al., 2013). The study area, bordered by the barrier islands on the north and by the mainland coast on the south, is formed by intertidal and subtidal mudflats and gullies. The monitoring network consists of 3451 permanent locations on intertidal mudflats at the nodes of a 500 m grid. The square grid is supplemented by 578 locations. These locations were selected by first selecting 578 out of the 3451 grid-points by simple random sampling without replacement. Then at each selected grid-point one point was selected at a uniformly distributed distance between 0 and 250 m distance from the grid-point, in a direction randomly chosen from the four directions defined by the grid-lines (Bijleveld et al., 2012). The total sample size was 4029 locations.

The most important determinants of habitat structure used for mapping the abundance were sediment texture characteristic, more specifically the mass fraction of silt, and altitude (Amsterdam Ordnance Datum, Rijkswaterstaat [1]). To be used as a predictor in mapping, the covariate must be known everywhere in the study area. Therefore the mass fraction of silt was interpolated by inverse distance weighting in ArcGIS 10.0.

2.2. Overview of evaluation method

The starting point of the our procedure for evaluating the sampling designs is a model for the spatial distribution of the zero-inflated and autocorrelated count data. This spatial distribution is modelled through a spatial zero-inflated Poisson mixture model(ZIP)(Lambert, 1992; Agarwal et al., 2002):

$$P(Y_i = y|\eta_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i) & y = 0 \\ (1-\pi_i)\frac{\exp(-\mu_i)\mu_i^y}{y!} & y = 1, 2, 3, \ldots \end{cases} \tag{1}$$

where Y_i is the count at location i, η_i is spatially dependent random effect, π_i the probability of a Bernoulli zero at location i, and $1 - \pi_i$ is the probability of a Poisson count, either zero or non-zero. The intensity (mean number of individuals) of the Poisson process at location i is μ_i. The first part of the model is the overall probability of zero (Hilbe and Greene, 2007).

The parameters π_i and μ_i at location i are random variables modelled by the following submodels:

$$\begin{aligned} \text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) &= \mathbf{x}_{\mathrm{B},i}^T\boldsymbol{\beta}_\mathrm{B} + \eta_{\mathrm{B},i} \\ \log(\mu_i) &= \mathbf{x}_{\mathrm{P},i}^T\boldsymbol{\beta}_\mathrm{P} + \eta_{\mathrm{P},i} \end{aligned} \tag{2}$$

with $\mathbf{x}_{\mathrm{B},i}$ and $\mathbf{x}_{\mathrm{P},i}$ vectors with covariates at location i, $\boldsymbol{\beta}_\mathrm{B}$ and $\boldsymbol{\beta}_\mathrm{P}$ vectors with regression coefficients, and $\eta_{\mathrm{B},i}$, $\eta_{\mathrm{P},i}$ residuals of the spatial trend. Note that the model parameters can be modelled by different sets of covariates.

The residuals $\eta_{\mathrm{B},i}$, $\eta_{\mathrm{P},i}$ at any location i are random variables. The probability distribution of the residuals at all locations in the study area was modelled as

$$\begin{bmatrix} \boldsymbol{\eta}_\mathrm{B} \\ \boldsymbol{\eta}_\mathrm{P} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_\mathrm{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_\mathrm{P} \end{bmatrix} \right) \tag{3}$$

with $\mathbf{C}_\mathrm{B}$ and $\mathbf{C}_\mathrm{P}$ covariance matrices. So note that we assumed that the Bernoulli and Poisson residuals were independent. Testing for this assumption we revealed a weak correlation of 0.3. For both random residuals we further assumed isotropy, so that the covariance of the residuals at any two locations was modelled as a function of the distance h between the two locations. For instance, for the Bernoulli residuals, the covariance was modelled as

$$C_\mathrm{B}(h) = \sigma_\mathrm{B}^2\rho_\mathrm{B}(h;\phi_\mathrm{B}) + \tau_\mathrm{B}^2 \tag{4}$$

[1] www.rijkswaterstaat.nl

with σ^2_{B} the partial sill, ϕ_{B} the range (distance parameter), τ^2_{B} the nugget, and ρ_{B} the correlation function, for instance exponential or spherical (Webster and Oliver, 2007).

The two submodels in 2 are generalised linear mixed models, as they are the sum of a linear combination of covariates describing a spatial trend (fixed effect) and a spatially autocorrelated residual (random effect). Diggle (2007) names this type of models as generalised linear *geostatistical* models.

Following Diggle (2007), ~~hereafter~~ the sum of the trend and residual, representing the transformed model parameter, is referred to as the signal S, for instance $S_{\mathrm{B},i} = \mathbf{x}^T_{\mathrm{B},i}\boldsymbol{\beta}_{\mathrm{B}}+\eta_{\mathrm{B},i}$. For convenience, all the parameters in one model, including the type of correlation function, are collected in a vector: $\boldsymbol{\theta}_{\mathrm{B}} = (\boldsymbol{\beta}_{\mathrm{B}}, \phi_{\mathrm{B}}, \tau^2_{\mathrm{B}}, \sigma^2_{\mathrm{B}}, \rho_{\mathrm{B}})$ and $\boldsymbol{\theta}_{\mathrm{P}} = (\boldsymbol{\beta}_{\mathrm{P}}, \phi_{\mathrm{P}}, \tau^2_{\mathrm{P}}, \sigma^2_{\mathrm{P}}, \rho_{\mathrm{P}})$. To avoid confusion the model parameters $\boldsymbol{\theta}_{\mathrm{B}}$ and $\boldsymbol{\theta}_{\mathrm{P}}$ are referred to as hyperparameters; with model parameters we mean the parameters π and μ.

The aim of evaluating the sampling strategy ~~to be evaluated~~ is to map the prevalence parameter π of the Bernoulli distribution and the intensity parameter μ of the Poisson distribution. ~~So~~ Please note that the ~~aim~~ objective is not to predict the species abundance counts, but to use the observed counts in the sample to estimate, at the desired sites i, the *expected* counts conditional on the values $\mathbf{x_i}$ of the covariates and the ~~*expected* counts , conditional on covariates that define a spatial trend and the observed counts at the sample to be evaluated~~random effects η_i that express spatial dependence. We believe that predicting the counts themselves is not feasible in our situation, and ~~besides~~ not of practical relevance.

~~Broadly speaking, our~~ Our evaluation procedure is as follows. The SIBES data are used to ~~calibrate~~ estimate the parameters of a ZIP model. Several steps are involved in ~~calibration~~estimation. First, a ZIP model is fitted by maximum likelihood assuming that both residuals $\eta_{\mathrm{B},i}$ and $\eta_{\mathrm{P},i}$ are spatially independent. The fitted model parameters are then used to classify a zero count either as a Bernoulli or a Poisson zero, and to construct two datasets: the Bernoulli dataset with zeros (absent) and ones (present), and the Poisson subset, containing the SIBES locations with a one in the Bernoulli dataset, with counts. In the next step these two data sets are used to fit the two submodels for the parameters π_i and μ_i, but now accounting for spatial autocorrelation. This is done by simulating a large sample of signals S_{B} and S_{P} at the SIBES locations by Markov chain Monte Carlo (MCMC) using initial estimates of the hyperparameters, followed by Monte Carlo Maximum Likelihood (MCML) estimation of the hyperparameters. The final MCML estimates of the hyperparameters are used to simulate signals S_{B} and S_{P} at the SIBES locations conditional on the observations at these locations.

The fitted hyperparameters are then used to ~~predict~~ simulate the Bernoulli signal (S_{B}) and Poisson signal (S_{P}) at the nodes of a very fine square grid with a spacing of 100 m covering the study area. This ~~prediction~~ grid is extended with 1000 randomly selected validation locations in between the grid-points. ~~The prediction method is simple kriging with an external drift, using silt, the squared values of silt, and altitude as external drift variables (predictors).~~

In the next step the ~~predicted~~ simulated signals at the grid-nodes and validation points are used to simulate fields with count data. One field with statistics closest to those of the SIBES data is selected, and underlying

S_B and S_P fields are repeatedly sampled by grid-sampling. A range of grid-spacings is applied. Each selected grid is used to predict the model parameters π_i and μ_i at the validation locations. By comparing these predictions with the true (original) model parameters at the validation points the quality of the predictions is assessed. More details on all steps but the first ~~two (calibration~~ (estimation of ZIP model ~~and prediction of signals~~parameters) are given below. For details on the first ~~two steps~~ step we refer to Lyashevska et al. (2016).

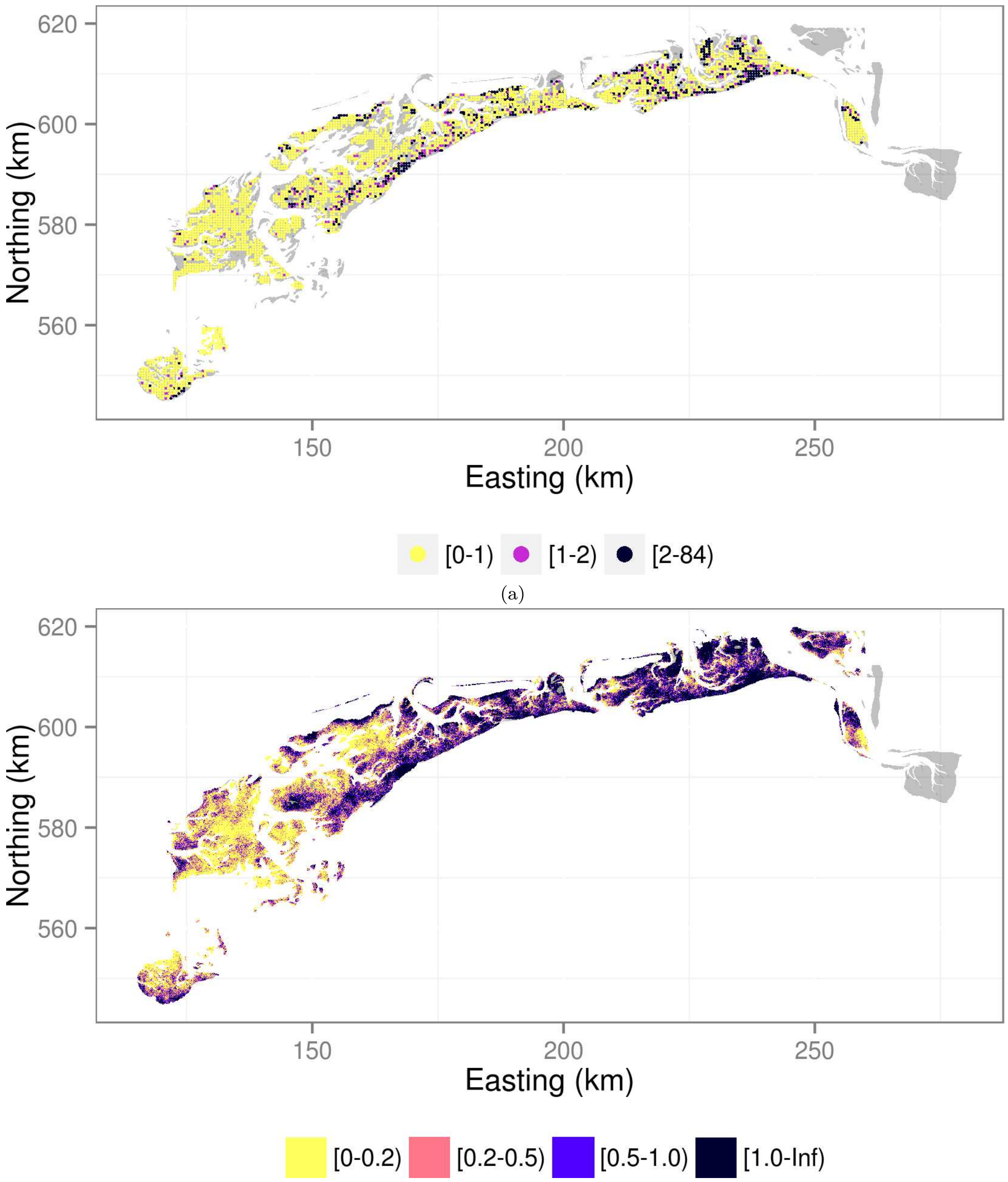
1. ~~After prediction of the signal~~

   The simulated signals S_B at the nodes of ~~the~~ 100 m grid ~~, 100 fields (square grids with spacing of 100 m) with spatially unstructured noise (white noise) are simulated from a Normal distribution with a variance equal to $\hat{\tau}^2_\mathrm{B}$. The predicted signal (S_B) is added to the simulated fields with white noise.The resulting fields are backtransformed by second-order Taylor expansion~~ extended with 1000 validation locations are backtransformed using the inverse of the link function in Eq. 2 to give 100 ~~simulated~~ fields of the prevalence parameter π of the Bernoulli distribution. The same procedure is followed for the S_P to give 100 ~~simulated~~ fields of the intensity parameter μ of the Poisson distribution.

2. ~~Each simulated field of π is used to simulate one field with presence/absence indicators, giving 100 fields with indicators. Besides, each simulated field of μ is used to simulate one field with Poisson counts, giving 100 fields with counts. These two sets of 100 fields are multiplied pair-by-pair, to give 100 fields with zero-inflated counts.~~

   ~~From the 100 fields with zero-inflated counts, one field is selected with summary statistics closest to these summary statistics of the SIBES data. We used as summary statistics the fraction of zeroes, the mean of the Poisson counts, and the variance of the Poisson counts. This leads to three difference values per field. These differences values were standardized by dividing through their standard deviation. The weighted sum of the standardized differences was used as a selection criterion, with weights equal to 0.5 (for fraction of zeroes), 0.25 and 0.25.~~ Apart from the field with zero-inflated counts, the two underlying fields with simulated prevalence parameter values π and simulated intensity parameter values μ are selected, as these are needed in the validation~~(remind that the aim was to predict the model parameters π and μ)~~. Fig. 1 shows a map of the product of π and μ, representing the unconditional intensity (unconditional expected count), and a map of the SIBES count data. There is a clear resemblance between the two maps.

3. The ~~selected field with zero-inflated counts is~~ two underlying fields are sampled on a grid with a spacing of 200, 400, 800, 1600, and 3200 m. For each grid-spacing 100 samples are randomly selected. The corresponding number of grid-points was on average 28505, 7130, 1783, 446, and 110, respectively. An overlay is made of the selected grids ~~and the selected field with simulated zero-inflated counts to obtain the 'pseudo-observations' of counts that will be~~ which are then used in prediction of the prevalence parameter π and the intensity parameter μ at the validation locations.



(a)

(b)

Figure 1: Empirical species abundance map of *Macoma balthica* (a) and unconditional intensity map (b) ~~predicted~~ conditionally simulated to the nodes of 100 m grid~~by simple kriging with an external drift. From Lyashevska et al. (2016)~~.

4. The model parameters π and μ at the 1000 validation locations are predicted by the same method as used in simulating our pseudo-reality, being simple kriging with an external drift, using silt, silt-squared and altitude as external drift variables. Ideally, for each grid-sample the hyperparameters are estimated from the 'pseudo-observations' of zero-inflated counts at the grid-points with Markov chain Monte Carlo maximum likelihood (MCML). Using these hyperparameters signals S_{B} and S_{P} should be simulated again conditional on the simulated counts (pseudo-observations). However, this is not feasible due to the computing time involved. Therefore, in predicting from the selected grid-points to the validation points we used the hyperparameters that were also used to simulate all 100 fields. These hyperparameters (referred to hereafter as the parent-hyperparameters) were estimated by MCML from the SIBES data. In practice these hyperparameters are unknown, so that by using the unknown parent-hyperparameters we ignore the contribution of uncertainty about the hyperparameters to the uncertainty about the predictions.

   To obtain a rough idea about this contribution, per grid-spacing four grid-samples are selected that are used to estimate the hyperparameters. As a consequence the hyperparameters are not fixed but vary between the four samples of a given grid-spacing. The hyperparameters are not estimated from the pseudo-observations of zero-inflated counts at the selected grid-points, but from the Bernoulli and Poisson signals at these points. In doing so we avoid the time-demanding MCML estimation. By using the simulated signals as observations the hyperparameters can be estimated by Residual Maximum Likelihood (REML). We are aware that this estimation procedure does not reflect practice either, and that the contribution of uncertainty about the hyperparameters will be underestimated, but we see it as a first attempt within reasonable computing time.

2.3. Quality measures

The quality of the predicted prevalence parameters was quantified by the Mean Squared Error (MSE); for instance for the prevalence parameter π this MSE equals :

$$\mathrm{MSE}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\pi}_i - \pi_i\}^2 \tag{5}$$

with n the number of validation points ($n = 1000$), $\hat{\pi}_i$ the predicted prevalence parameter at location i and π_i the 'pseudo-observed' intensity parameter. For intensity parameter μ MSE is computed from the subset of validation points with a simulated value of 1 for the presence/absence indicator (species present). This subset contains 211 points. MSE was also calculated for the product of π and μ, representing unconditional intensity (intensity not conditioned on presence). For this product again all 1000 validation points are used.

For each grid-spacing we have 100 grid-samples. All 100 grid-samples are used in prediction with the fixed parent-hyperparameters $\boldsymbol{\theta}_{\mathrm{B}}$ and $\boldsymbol{\theta}_{\mathrm{P}}$ (no sample-specific estimation), leading to 100 estimates of MSE(π), MSE(μ) and MSE($\pi \cdot \mu$). The distribution of these 100 estimates is an estimate of the sampling distribution of the estimated mean quality of model-based predictions of the model parameters π and μ. Only ~~four~~ five grid-samples are used in prediction with variable sample-specific estimates of the hyperparameters $\boldsymbol{\theta}_{\mathrm{B}}$ and $\boldsymbol{\theta}_{\mathrm{P}}$, so these ~~four~~ five estimates give a very rough estimate of the sampling distribution only.

3. Results

3.1. Prevalence

The mean of ~~MSE~~ the 100 MSEs (1 MSE per grid of a given spacing) of the predicted species prevalence parameter π increased with increasing grid-spacing (Fig. 2a). Using the fixed parent-hyperparameters the increase was from ~~a value close to 0~~ 0.006 at 200 m to ~~0.001~~ 0.008 at 3200 m. ~~Except for the largest spacing of 3200 m, the~~ The variance of MSE between the 100 grid samples was small for all grid-spacings.

The mean of the ~~four~~ five MSE values using hyperparameters estimated from the grid-samples, was for all spacings larger than the mean of MSE using the fixed parent-hyperparameters, especially for the largest spacing of 3200 m. This shows that the contribution of uncertainty about the hyperparameters to the uncertainty about the predictions was substantial. Remarkable is the strong increase of the mean MSE beyond a spacing of 1600 m.

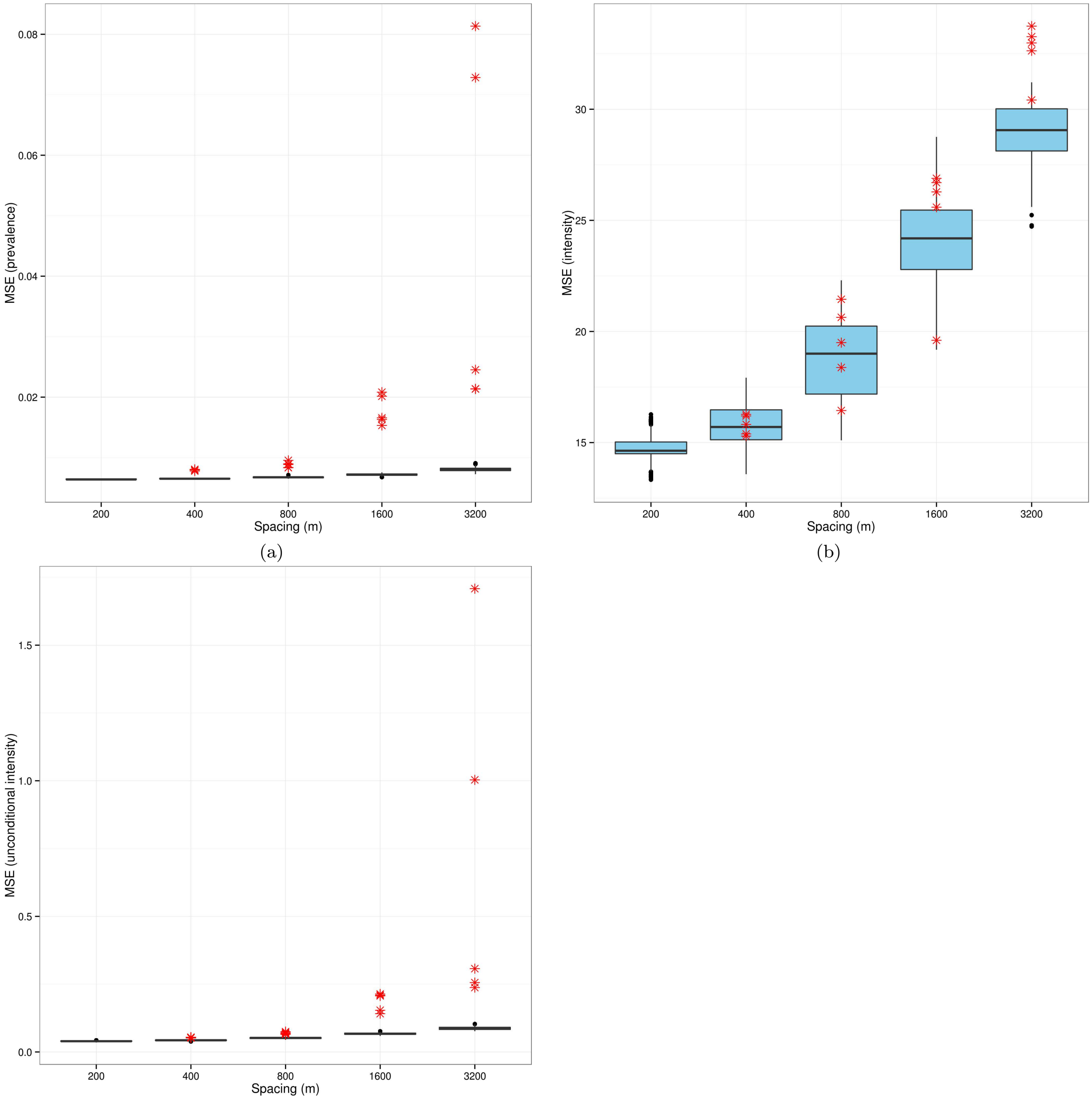
3.2. Intensity

Similar to the prevalence parameter π, the mean of MSE of the predicted intensity parameter μ increased with increasing grid-spacing (Fig. 2b). Using fixed parent-hyperparameters the increase was from ~~0.26~~ 14.73 at 200 m to ~~12.25~~ 28.84 at 3200 m (Fig. 2b). Using hyperparameters estimated from the grid-samples, the mean of MSE increased even more, from ~~2.41~~ 15.89 for 400 m (compared with ~~0.7~~ 15.73 at 400 m for fixed hyperparameters) to ~~13.82~~ 32.32 at 3200 m. ~~In contrast to the prevalence parameter π the mean MSE already started to increase strongly at spacings beyond 400 m.~~

The graph of the MSE for the product of π and μ, the unconditional intensity, is very similar to the graph for the prevalence parameter π. For the first four spacings the increase of the mean MSE was very modest, but beyond a spacing of 1600 m, the increase was much stronger (Fig. 2c).

4. Discussion and conclusions

The ~~objective~~ aim of this paper was to quantify the effect of the grid-spacing on the quality of spatial predictions of abundance of species that show zero-inflation and spatial autocorrelation. We proposed an approach in which multiple fields of the prevalance parameter π and the intensity parameter μ of a zero-inflated Poisson mixture model are simulated by generalized linear geostatistical models. These fields were used to simulate pseudo-realities of zero-inflated autocorrelated counts. One pseudo-reality was then selected with summary statistics that were close to summary-statistics of the available data. This pseudo-reality was then sampled by the sampling design under study, in this study grid-sampling at various grid-spacings. A selected sample was used to predict at a fixed set of validation locations and to compute the MSE of the



(a)

(b)

(c)

Figure 2: The MSE for predicted prevalence (π) (a), predicted intensity (μ) (b) and predicted unconditional intensity ($\pi \cdot \mu$) (c) as a function of grid-spacing for 200, 400, 800, 1600, and 3200 m. Predictions were obtained by simple kriging with an external drift with fixed parent-hyperparameters (blue) and hyperparameters estimated from a sample (red). The first grid-spacing (200 m) with estimated hyperparameters could not be estimated due to memory constraints.

predictions of π and μ. By repeating the selection of samples and the prediction at the validation points, an estimate of the sampling distribution of MSE is obtained.

To construct the graph with the estimated sampling distribution of MSE against grid-spacing the hyperparameters that were used to simulate the pseudo-reality were also used in spatial prediction at the validation points. As a consequence, this graph shows the effect of the grid-spacing *given the model.* This is common practice in designing spatial samples for mapping by kriging. McBratney and Webster (1981) optimized the spacing of grids using as an evaluation criterion the maximum kriging variance as obtained with ordinary kriging. Uncertainty about the variogram used in ordinary kriging is not accounted for. van Groenigen et al. (1999) optimized the spatial coordinates of a given number of sampling locations for ordinary kriging, also assuming that the variogram is known. Brus and Heuvelink (2007) did the same for kriging with an external drift (KED). This study resembles the study described here, apart from that a linear mixed model is used instead of a generalized linear mixed model. The variance as computed with KED does account for uncertainty about the trend coefficients, but does not account for uncertainty about the variogram parameters, so also in this study the contribution of uncertainty about the variogram parameters is ignored.

Alternatively, in order to account for the uncertainty about the hyperparameters in estimating the MSE, the hyperparameters that are used in prediction are estimated from the samples. By repeatedly selecting samples with a given sampling design, estimating the hyperparameters, and predicting at the validation points, a sampling distribution of the MSE is obtained with a mean that will be larger than that of the sampling distribution of MSE obtained with the fixed parent-hyperparameters. The difference in the sampling distributions of MSE reflects the contribution of the uncertainty about the hyperparameters to uncertainty about the predictions due to sampling errors in the estimated hyperparameters.

This procedure for evaluating sampling designs is relatively simple and versatile. However, the applicability of this approach in our case study was hampered by the computing time involved in estimating the hyperparameters from a sample. In the proposed model for the zero-inflated counts spatial dependency is introduced at the level of the model parameters π and μ, which cannot be directly observed. Besides, both model parameters are non-linearly related to environmental covariates. The parameters of such model can be estimated by Monte Carlo Maximum Likelihood (MCML), which involves repeated simulation of long Markov chains, which is time consuming. As an approximation we estimated the hyperparameters by REML from the unobservable model parameters π and μ at the selected sampling locations. Most likely this approximation underestimates the contribution of the uncertainty about the hyperparameters to the uncertainty about the predictions, but it is a first attempt within reasonable computing time. To improve the quality of the estimated MSEs at various grid-spacings a more efficient procedure for estimating the hyperparameters is needed. We welcome research into estimation of the hyperparameters of a ZIP mixture model by integrated nested Laplacian approximation, as proposed by Rue et al. (2009).

For the time being we must base our decisions on the graphs as presented in Figs 2a, 2b and 2c. The graphs of the prevalence parameter π and of the unconditional intensity $\pi \cdot \mu$ show that beyond a spacing of

1600 m the mean of MSE starts increasing at a much faster rate. This is especially true for the MSEs obtained with hyperparameters estimated from the samples. Based on these findings the 1600 m grid which consists of 446 locations for our study area of 2400 km^2 seems to be a good compromise between sampling costs and prediction accuracy. The increase in accuracy of the 800 m grid does not seem to outweigh the fourfold increase in number of sampling units taken. But of course the optimal spacing can only be determined when sampling costs and accuracy are defined in a common unit.

Acknowledgements

This study was part of the Wadden Long-Term Ecosystem Research (WaLTER) project, that is financially supported by the Waddenfonds, and the provinces of Fryslân and Noord Holland (Grant/Award Number: WF209902). We thank the SIBES core-team, numerous volunteers, and the crew of the RV Navicula. In 2010 the SIBES-monitoring was carried out with financial support from NAM, NWO-ALW (ZKO programme) and Royal NIOZ.

Supporting information

Data and R-code can be found on `https://github.com/lyashevska/GridSpacing-Paper.git`

References

Agarwal, D., Gelfand, A., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics 9, 341–355. URL: http://dx.doi.org/10.1023/A%3A1020910605990, doi:10.1023/A:1020910605990.

Bijleveld, A.I., van Gils, J.A., van der Meer, J., Dekinga, A., Kraan, C., van der Veer, H.W., Piersma, T., 2012. Designing a benthic monitoring programme with multiple conflicting objectives. Methods in Ecology and Evolution 3, 526–536. URL: http://dx.doi.org/10.1111/j.2041-210X.2012.00192.x, doi:10.1111/j.2041-210X.2012.00192.x.

Boyd, C., Woillez, M., Bertrand, S., Castillo, R., Bertrand, A., Punt, A.E., 2015. Bayesian posterior prediction of the patchy spatial distributions of small pelagic fish in regions of suitable habitat. Canadian Journal of Fisheries and Aquatic Sciences 72, 290–303.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86–95.

Caughlan, L., Oakley, K.L., 2001. Cost considerations for long-term ecological monitoring. Ecological Indicators 1, 123 – 134. URL: http://www.sciencedirect.com/science/article/pii/S1470160X01000152, doi:http://dx.doi.org/10.1016/S1470-160X(01)00015-2.

Clarke, K.R., Green, R.H., 1988. Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series 46, 213–226.

Compton, T.J., Holthuijsen, S., Koolhaas, A., Dekinga, A., ten Horn, J., Smith, J., Galama, Y., Brugge, M., van der Wal, D., van der Meer, J., van der Veer, H.W., Piersma, T., 2013. Distinctly variable mudscapes: Distribution gradients of intertidal macrofauna across the Dutch Wadden Sea . Journal of Sea Research 82, 103 – 116. URL: http://www.sciencedirect.com/science/article/pii/S1385110113000300, doi:http://dx.doi.org/10.1016/j.seares.2013.02.002. special issue: Proceedings of the International Symposium on the Ecology of the Wadden Sea.

Cozzi, G., Broekhuis, F., McNutt, J., Schmid, B., 2013. Density and habitat use of lions and spotted hyenas in northern botswana and the influence of survey and ecological variables on call-in survey estimation. Biodiversity and Conservation 22, 2937–2956. URL: http://dx.doi.org/10.1007/s10531-013-0564-7, doi:10.1007/s10531-013-0564-7.

Diggle, Peter J.and Ribeiro Jr., P.J., 2007. Model-based Geostatistics. Springer.

Fortin, M.J., Drapeau, P., Legendre, P., 1989. Spatial autocorrelation and sampling design in plant ecology. Vegetatio 83, pp. 209–222. URL: http://www.jstor.org/stable/20038496.

Foster, S.D., Hosack, G.R., Hill, N.A., Barrett, N.S., Lucieer, V.L., 2014. Choosing between strategies for designing surveys: autonomous underwater vehicles. Methods in Ecology and Evolution 5, 287–297. URL: http://dx.doi.org/10.1111/2041-210X.12156, doi:10.1111/2041-210X.12156.

van Groenigen, J., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8, 993–1009. URL: http://dx.doi.org/10.1111/j.1461-0248.2005.00792.x, doi:10.1111/j.1461-0248.2005.00792.x.

Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Hanberry, B., He, H., Dey, D., 2012. Sample sizes and model comparison metrics for species distribution models. Ecological Modelling 227, 29 – 33. URL: http://www.sciencedirect.com/science/article/pii/S0304380011005837, doi:http://dx.doi.org/10.1016/j.ecolmodel.2011.12.001.

Hilbe, J., Greene, W., 2007. Count response regression models, in: Rao, C., Miller, J., Rao, D. (Eds.), Epidemiology and Medical Statistics. Elsevier. Elsevier Handbook of Statistics Series.

Lambert, D., 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 34, pp. 1–14. URL: http://www.jstor.org/stable/1269547.

Legendre, P., Dale, M.R.T., Fortin, M.J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, pp. 601–615.

Lewis, F., Butler, A., Gilbert, L., 2011. A unified approach to model selection using the likelihood ratio test. Methods in Ecology and Evolution 2, 155–162. URL: http://dx.doi.org/10.1111/j.2041-210X.2010.00063.x, doi:10.1111/j.2041-210X.2010.00063.x.

Lyashevska, O., Brus, D.J., van der Meer, J., 2016. Mapping species abundance by a spatial zero-inflated Poisson model: a case study in the Wadden Sea, the Netherlands. Ecology and Evolution URL: http://dx.doi.org/10.1002/ece3.1880, doi:10.1002/ece3.1880.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8, 1235–1246. URL: http://dx.doi.org/10.1111/j.1461-0248.2005.00826.x, doi:10.1111/j.1461-0248.2005.00826.x.

McBratney, A., Webster, R., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables-ii. program and examples. Computers and Geosciences 7, 335–365. Cited By (since 1996) 53.

Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend Peterson, A., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in madagascar. Journal of Biogeography 34, 102–117. URL: http://dx.doi.org/10.1111/j.1365-2699.2006.01594.x, doi:10.1111/j.1365-2699.2006.01594.x.

Perner, J., Schueler, S., 2004. Estimating the density of ground-dwelling arthropods with pitfall traps using a nested-cross array. Journal of Animal Ecology 73, pp. 469–477. URL: http://www.jstor.org/stable/3505657.

Rachowicz, L.J., Hubbard, A.E., Beissinger, S.R., 2006. Evaluating at-sea sampling designs for marbled murrelets using a spatially explicit model. Ecological Modelling 196, 329 – 344. URL: http://www.sciencedirect.com/science/article/pii/S0304380006000731, doi:http://dx.doi.org/10.1016/j.ecolmodel.2006.02.011.

Recta, V., Haran, M., Rosenberger, J.L., 2012. A two-stage model for incidence and prevalence in point-level spatial count data. Environmetrics 23, 162–174. URL: http://dx.doi.org/10.1002/env.1129.

Reynolds, J.H., Thompson, W.L., Russell, B., 2011. Planning for success: Identifying effective and efficient survey designs for monitoring. Biological Conservation 144, 1278 –

1284. URL: http://www.sciencedirect.com/science/article/pii/S0006320710004970, doi:http://dx.doi.org/10.1016/j.biocon.2010.12.002.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 319–392. URL: http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x.

Stockwell, D.R., Peterson, A., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1 – 13. URL: http://www.sciencedirect.com/science/article/pii/S030438000100388X, doi:http://dx.doi.org/10.1016/S0304-3800(01)00388-X.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K., Possingham, H.P., 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. Ecological Applications 13, 1790–1801.

Vieira, C., Seneca, A., Sergio, C., 2012. Floristic and ecological survey of bryophytes from portuguese watercourses. Cryptogamie, Bryologie 33, 113–134.

Webster, R., Oliver, M.A., 2007. Geostatistics for environmental Scientists. 2 ed., John Wiley.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763–773. URL: http://dx.doi.org/10.1111/j.1472-4642.2008.00482.x, doi:10.1111/j.1472-4642.2008.00482.x.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., Grimm, V., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119, 622–635. URL: http://dx.doi.org/10.1111/j.1600-0706.2009.18284.x, doi:10.1111/j.1600-0706.2009.18284.x.