SCIENTIFIC REPORTS
nature research

OPEN

Mapping species richness using opportunistic samples: a case study on ground-floor bryophyte species richness in the Belgian province of Limburg

Thomas Neyens[1,2,3*], Peter J. Diggle[4], Christel Faes[1], Natalie Beenaerts[2], Tom Artois[2] & Emanuele Giorgi[4]

In species richness studies, citizen-science surveys where participants make individual decisions regarding sampling strategies provide a cost-effective approach to collect a large amount of data. However, it is unclear to what extent the bias inherent to opportunistically collected samples may invalidate our inferences. Here, we compare spatial predictions of forest ground-floor bryophyte species richness in Limburg (Belgium), based on crowd- and expert-sourced data, where the latter are collected by adhering to a rigorous geographical randomisation and data collection protocol. We develop a log-Gaussian Cox process model to analyse the opportunistic sampling process of the crowd-sourced data and assess its sampling bias. We then fit two geostatistical Poisson models to both data-sets and compare the parameter estimates and species richness predictions. We find that the citizens had a higher propensity for locations that were close to their homes and environmentally more valuable. The estimated effects of ecological predictors and spatial species richness predictions differ strongly between the two geostatistical models. Unknown inconsistencies in the sampling process, such as unreported observer's effort, and the lack of a hypothesis-driven study protocol can lead to the occurrence of multiple sources of sampling bias, making it difficult, if not impossible, to provide reliable inferences.

Citizen science is a body of research in which scientific investigations are carried out through the involvement of volunteers from the general public in collaboration with experts. This has received increased attention in different scientific fields over the last decade, including ecology[1–4] where public participation provides a low-cost method for data collection in order to provide timely information on ecological processes. For example, in ornithology, studies on the distribution of British birds have used citizen-science data (CSD) to record breeding locations[5] and ringing data, partly collected by citizens, have been used to investigate autumn migration in the French-Swiss Alps[6]. CSD have also been used to assess the effects of habitat degradation on avian population dynamics[7], invertebrate distributions[8], and habitat use of rare species[9].

Price & Lee[10] define a set of citizen science categories, based on varying degrees of citizens' involvement and adherence to a pre-arranged study protocol. When we refer to CSD in this study, we focus on the *active contributory model*, in which participants actively engage in data collection and/or processing, while making individual decisions regarding sampling strategies. Although this model is often used in wildlife monitoring, its economical benefits are offset by the absence of a consistent study design methodology. As a result, it is inadvisable to use CSD for modelling complex ecological processes without understanding the mechanisms that might yield biased

[1]Centre for Statistics, Data Science Institute, Hasselt University, Agoralaan, Building D, B-3590, Diepenbeek, Belgium. [2]Centre for Environmental Sciences, Faculty of Sciences, Hasselt University, Agoralaan, Building D, B-3590, Diepenbeek, Belgium. [3]Leuven Biostatistics and statistical Bioinformatics Centre, Faculty of Medicine, KU Leuven, Kapucijnenvoer 35, block D, box 7001, B-3000, Leuven, Belgium. [4]Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, LA1 4YW, United Kingdom. *email: thomas.neyens@uhasselt.be

inferences on the phenomenon under investigation[11]. Isaac *et al.*[12] identify four main sources of bias in this type of CSD: (i) temporal and (ii) spatial unbalance in data collection, (iii) uneven species detectability, and (iv) uneven sampling effort per visit. Although (i), (ii), and (iii) are not exclusive to CSD, they are likely to be exacerbated in the absence of predefined sampling guidelines. Uneven sampling effort (iv) is more particular to CSD, and may be the product of different sampling inconsistencies, such as differences in the durations spent on the field and the number of data collectors per visit. Hadj-Hammou *et al.*[13] argue that the development of methods to investigate the drivers behind the behaviour of citizens is essential to address these issues.

Recent studies on the biodiversity of ecosystems have reported mixed results on the usefulness of CSD. For example, variation in observational skills among citizens has been found to be an important source of bias in detecting adelgids[14] and anurans[15]. In contrast, Delaney *et al.*[16] concluded that CSD provided useful insights into the distribution of native and invasive crabs, but also highlighted that age and university training of the collector were associated with identification skills. Boudreau & Yan[17] found that citizens were able accurately to detect invasions of non-native water fleas in Canadian lakes. van Strien *et al.*[18] found a good match between predicted trends from CSD and monitoring data on the distribution of dragonflies and butterflies.

Species richness is defined as the number of species present in an ecosystem and is used as a measure of biodiversity. Studies on species richness frequently rely on CSD but their opportunistic nature is often ignored or is addressed through the use of overly simplistic approaches that cannot account for specific sources of sampling bias. An exception is given by Carota *et al.*[19], who model species richness from historical data using a semi-parametric Poisson model with random effects drawn from a mixture of Dirichlet processes. They show that this data-driven approach better captures the large heterogeneity induced by the opportunistic nature of the underlying sampling mechanism than standard Poisson mixed models. Over the last decade, research based on the use of CSD has been mainly focused on the modelling of geographical distributions of a single animal species. A commonly used approach consists of subsampling the data based on the number of location visits and the amount of species collected per location. For example, in their analysis of European ladybird declines, Roy *et al.*[20] only consider data collected over a $1 \times 1\,km^2$ regular grid, where at least two common species were observed. Similarly, Kuussaari *et al.*[21] discard all data on a $10 \times 10\,km^2$ regular grid with less than 40 records of focal farmland butterfly species reported before 1960 to assess changes in their spatial distribution, resulting in the loss of 85% of data points from cells visited before 1960. A generally more robust approach, which we also follow in this paper, is to account for potential sources of bias through the inclusion of key explanatory variables to the model. For example, Szabo *et al.*[22] include the total number of species to model the abundance of several avian species as a proxy for the observer's effort. Kelling *et al.*[23] and Johnston *et al.*[24] use model-based predictions of the number of recorded species to correct for observer-specific species' detectability skills. Occupancy-detection models, originally proposed to model imperfect detection of species[25,26], have also been used to account for observation and reporting bias[18]. However, these models require presence/absence information from repeated site visits and often result in an excessive aggregation of the data over time and space, which is questionable when locations are revisited at highly irregular time intervals.

In this paper, we focus on the spatial prediction of species richness of ground-floor bryophytes in Limburg, Belgium, using data from an expert source and CSD. More specifically, the first source consists of randomised survey data (RSD) collected by a biologist who adheres to a predefined and randomised sampling design protocol. The CSD, in contrast, include non-randomised opportunistic samples obtained by a team of two individuals who are occasionally joined by other collectors. An important aspect of our approach is the explicit modeling of the spatial correlation between observations. This issue has been extensively addressed in modelling species distributions[27–29] with some attempts to account for sampling and detection bias (see, for example, Pacifici *et al.*[30]). Conn *et al.*[31] use geostatistical methods to model ecological data in the presence of *preferential sampling*. This term refers to a special case of opportunistic sampling in which there is stochastic dependence between the sampling design and the reported species counts[32]. In the context of species richness modelling, Chakraborty *et al.*[33] use spatial point processes to estimate the distribution of six different species separately. From each individual point process model, they then draw posterior samples for the estimated intensity functions to predict the overall species richness.

The objective of our study is to assess the reliability of the CSD in predicting bryophyte richness. Hence, we first use a log-Gaussian Cox process model to understand what variables might have affected the opportunistic nature of the sampling mechanism in the CSD. We then fit geostatistical Poisson models to the two data-sets and compare the resulting spatial estimates for bryophyte richness. Model-based geostatistics[34] provides a principled, likelihood-based approach to inference. It also exploits the correlation between recordings of species richness by accounting for unmeasured environmental factors through the inclusion of a latent spatial Gaussian process in the linear predictor of the response variable. To the best of our knowledge, this is the first study that, unlike others where groups of citizens are large and heterogeneous, (i) uses a model-based approach to describe the sampling pattern of individual collectors in a CSD context and (ii) validates the use of the CSD for mapping species richness using geostatistical methods.

Methods

Data. Our study area is the province of Limburg, covering $2{,}414\,km^2$ in the Eastern part of Belgium; we do not include Voeren, a smaller exclave of the Limburg province. The data, which are publicly available[35], are obtained from two sources: the RSD correspond to observations recorded by a biologist working for the Belgian Nature and Forest Agency (ANB); the CSD consist of opportunistic samples collected by two citizens for the Umbrella for Nature Research in Limburg (LIKONA). The citizens lived in the western part of Limburg and were occasionally joined by other citizens. The outcome of interest, available from both sources, is the number of distinct forest ground-floor bryophyte species determined in a lab through microscopy detection or visual inspection of a moss sample. Species richness is often used as an index of biodiversity when the total counts per species are difficult or

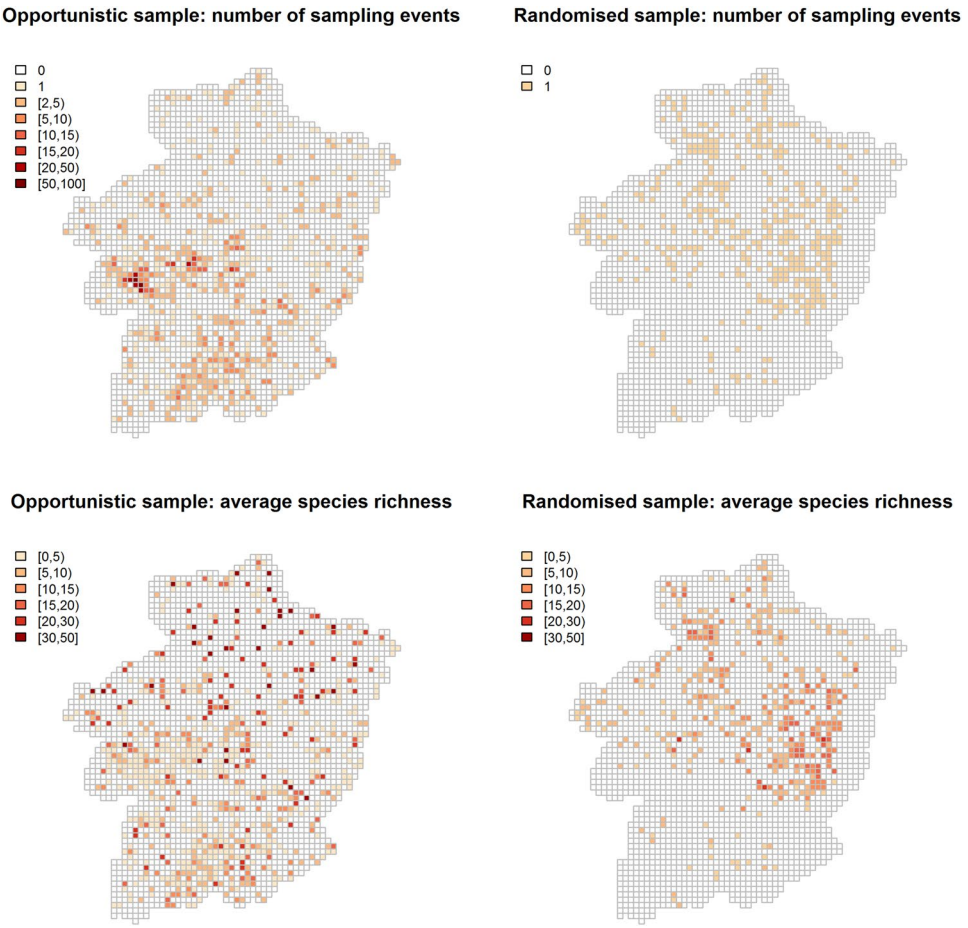

Figure 1. Maps of the number of sampling events and the average observed forest ground-floor species richness over a 1 by 1 km regular grid, from the two data sources.

impossible to assess[36]. Both data-sets include information on the date of collection, the name of the data collectors and the GPS locations of the sampled cell from a 1 by 1 km regular grid created by the Institute for Floristics in Belgium and Luxembourg (IFBL). Each data collector reported all bryophyte species observed during the surveys.

Geographical summaries of sampling events and the observed species richness for both the RSD and CSD are given in Fig. 1. The RSD were collected between 1997 and 1999 and include a total of 420 locations in Limburg that were selected using a randomised systematic lattice-based sampling design, which was restricted to the forested areas of Limburg; a 1 km $\times$ 1 km raster was randomly superimposed on the map of Limburg. All intersections between horizontal and vertical raster lines that landed on an area that was officially defined as a forest, were designated as sampling locations. During the sampling, the data collector walked no more than 50 m in one direction, starting from the sampling location, and repeated this in directions at 90°, 180°, and 270° angles from the initial direction. Sampling was restricted to the prescribed trajectory. The collector assembled samples of all ground-floor bryophytes that, by visual inspection, were species not yet encountered during the sampling event. Those samples were later investigated under laboratory circumstances. There was no fixed or maximal duration of a sampling event, which is a sensible choice, since data gathering is likely to take a longer time on more biodiverse trajectories. More details can be found in Afdeling Bos & Groen[37]. Figure 1 (top right panel) shows the resulting set of sampled locations. The CSD contain 2,088 recordings of bryophyte richness from opportunistically selected locations between 1985 and 2009 (Fig. 1, top left panel). From interviews, we learnt that the selection of locations was often based on their proximity to the home of the two main collectors, with preference for those locations that were more biodiverse according to the collectors' knowledge. Except for a number of sites relatively close to their homes, locations were rarely revisited. The goal of the data collection carried out by the two citizens was to cover the largest possible area in the Limburg province. As a result, unlike the RSD, sampling was not restricted to forested areas and the citizens reported observations of all possible bryophyte species, regardless of habitat and substrate. The two main CSD collectors were volunteer bryophyte experts, with species identification skills that can be assumed to be similar to the RSD's expert knowledge. The observers' efforts, such as the time spent in the field, varied greatly between events and were not documented. However, the interviews also revealed that both collectors were mostly in close proximity and followed the same trajectory when collecting data. There is no available documentation of all Belgian forest ground-floor bryophytes. The species labelled here as forest ground-floor bryophytes, occur in forests, but are not necessarily restricted to forested areas and ground floors, which means that the citizens could have found these species at locations other than forests and/or ground floors. Furthermore,

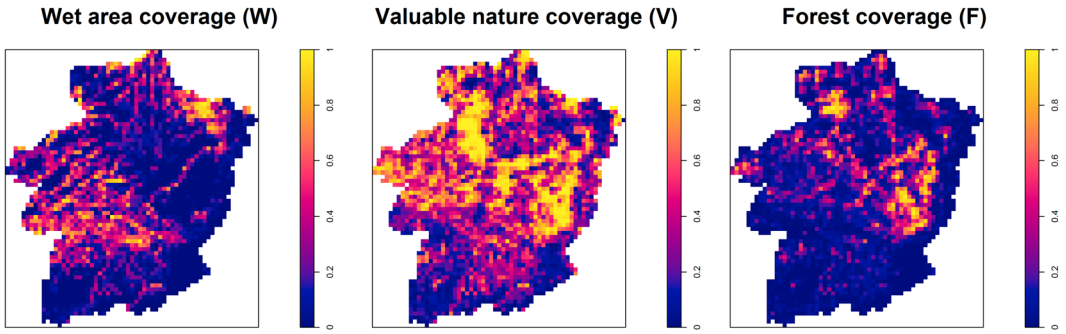

Figure 2. Relative coverages of wet areas, valuable nature, and forests, per IFBL cell.

since spatial information is only available on the grid-level, which mostly consists of multiple habitat types, we do not know which of the CSD's samples were taken from ground floors of forests. As a pragmatic solution, we extract from the CSD database only the observations of species that occur in the RSD data.

Environmental variables. We use information on wetness, environmental evaluation and forest cover to generate ecological variables that might help to explain the spatial variation in bryophyte species richness. The Flemish geological map[38], last updated in 2017, provides wetness categories for each landscape zone within the study area. Since multiple landscape zones may fall within each IFBL cell, we compute the proportion of wet landscape zones (henceforth, W) as shown in Fig. 2. Similarly, we compute the relative coverage of valuable nature (henceforth V; Fig. 2), defined as the proportion of landscapes within an IFBL cell that are wholly or partly labelled as "very valuable" or "valuable" according to the Belgian Biological Evaluation Map anno 2016[39]. Nature evaluation is based on several correlated environmental characteristics, including habitat type, presence of ecologically important fauna and flora and the level of human disturbance. Hence, using the variable V allows us to avoid multicollinearity issues while capturing the main environmental features of each IFBL cell. We do not use temporally varying values for W and V, but it is reasonable to assume that these have only slightly changed between 1985 and 2009, since environmental planning in Limburg has undergone only minor changes throughout the last 40 years. Finally, the relative forest coverage within each IFBL cell (henceforth, F) was calculated, based on the Flemish Forest Map[40].

Statistical analysis. *Modelling of opportunistic sampling as a preferential sampling scheme.* Preferential sampling is a term coined by Diggle *et al.*[32] to denote processes that generate sampling locations X that are stochastically dependent on the underlying spatial process of interest, which in our case corresponds to the species richness at any given location in Limburg. The opportunistic samples from the CSD study might in fact have been generated from a similar mechanism whereby the stochastic dependence between species richness and the selection of a location for data collection is induced by several unmeasured factors, which might increase the likelihood of finding more species. The RSD's sampling locations were chosen non-preferentially, as they were an independent sample from a (non-uniform) distribution on the region of Limburg, in which locations outside forests had a sampling probability of zero, while locations inside forests had an equal sampling probability.

Ignoring stochastic dependence caused by preferential sampling can invalidate our predictive inferences on the outcome of interest, as shown in Diggle *et al.*[32]. Our modelling strategy is inspired by Pati *et al.*[41] who extended the geostatistical method of Diggle *et al.*[32] using a bivariate spatial process. They proposed a joint modelling approach where a geostatistical model of the outcome of interest shares a spatial random effect with a log-Gaussian Cox process (LGCP) that models the choice of sampling locations. Note that by definition, preferential sampling refers to stochastic dependence between species richness and location choice; in other words, it reflects unexplained variability that is shared between both models. Preferential sampling does not occur as a result of variability that is explained by shared predictors.

We proceed as follows. To understand how distance from home and evaluation of nature affected the choice of a sampling location, x, in the CSD data, we model the latter as the realisation of a LGCP with intensity

$$\Lambda(x) = exp\{\alpha_0 + \alpha_1 D(x) + \alpha_2 V(x) + S(x)\}, \quad (1)$$

where: $D(x)$ is the distance between x and the collectors' residential municipality; $V(x)$ is the evaluation of nature at location x; and $S(x)$ is a stationary and isotropic Gaussian process with zero mean, variance σ^2 and Matérn[42] correlation function given by

$$\rho(u; \varphi, \kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}\left(\frac{u}{\varphi}\right)^{\kappa} K_{\kappa}\left(\frac{u}{\varphi}\right),$$

where: $u > 0$ is the Euclidean distance between any two locations; ϕ is a scale parameter regulating how quickly the spatial correlation decays to zero for increasing distance; and $K_k(.)$ denotes the modified Bessel function of the second kind, of order $\kappa > 0$. Zhang[43] warns that under fixed-domain asymptotics σ, ϕ, and κ, cannot be estimated consistently; as a pragmatic approach, we therefore set $\kappa = 1$. Fitting the LGCP model is computationally intensive; we therefore approximate $S(x)$ using a stochastic partial differential equations (SPDE) approach. SPDE

	Parameter	estimate	95% C.I.
intercept	α_0	2.970	[2.822;3.117]
distance to house	α_1	−0.031	[−0.035;−0.027]
value of nature	α_2	2.778	[2.601;2.955]
spatial variance	$\log(\sigma^2)$	2.487	[2.302;2.673]
scale	$\log(\phi)$	2.678	[2.176;3.180]

Table 1. Results for the LGCP model for the opportunistic sampling process of the CSD.

uses a triangulation method based on Gaussian weights with Markov dependencies that approximate the Matérn covariance structure. More details can be found in Lindgren *et al.*[44]. The triangulation mesh was constructed using recommendations outlined in Krainski *et al.*[45].

The objectives of the geostatistical analysis are to model species richness from the two data sources and to quantify the differences between the resulting predictive inferences. In the case of the CSD, we also account for bias that might be induced by preferential sampling.

Let $Y_j(x_i, t_i)$ denote the total number of forest ground-floor bryophyte species collected at location x_i and time $t_i \in \{1985, \ldots, 2009\}$ from source $j \in \{CSD, RSD\}$. Conditionally on a spatial Gaussian process $U_j(x)$ and Gaussian noise $Z_j(x_i)$, we assume that $Y_j(x_i, t_i)$ are mutually independent Poisson random variables with means $\mu_j(x_i, t_i)$. More specifically, for $j = CSD$, we write

$$\begin{aligned}\mu_{CSD}(x_i, t_i) \;=\; & exp\{\beta_{0,CSD} + \beta_{1,CSD}W(x_i) \\ & +\beta_{2,CSD}V(x_i) + \beta_{3,CSD}F(x_i) + \beta_4 log(C_i) \\ & +f(t_i) + \gamma\hat{S}(x_i) + U_{CSD}(x_i) + Z_{CSD}(x_i)\}\end{aligned} \tag{2}$$

where: $\hat{S}(x_i)$ is the predictive mean of the spatial process in Eq. (1) at location x_i used to account for preferential sampling. Based on the results from our exploratory analysis (Appendix; Fig. A.1), we assume a log-linear relationship between $Y_j(x_i, t_i)$ and $\hat{S}(x_i)$; $W(x_i)$, $V(x_i)$ and $F(x_i)$ are the three spatial variables, shown in Fig. 2, at location (x_i); C_i is the number of collectors; and $f(t_i)$ is a cubic spline with knots at 1990, 1995, 2000, 2002, 2004, and 2006. Note that, if all factors that explain the spatial variation in the choice of the sampled locations in the CSD study were available, this would lead to $S(x_i) = 0$ for all i, and hence $\hat{S}(x_i) = 0$ meaning that bias arising from preferential sampling would be completely removed from the model for species richness.

Finally, for $j = RSD$, we write

$$\mu_{RSD}(x_i, t_i) = exp\{\beta_{0,RSD} + \beta_{1,RSD}W(x_i) + \beta_{2,RSD}V(x_i) + \beta_{3,RSD}F(x_i) + \beta_5 t_i^{st} + U_{RSD}(x_i) + Z_{RSD}(x_i)\}, \tag{3}$$

where: $t_i^{st} = t_i - 1997$. We use ν_j^2 and ψ_j to denote the variance and scale of the exponential spatial covariance functions used for $U_j(x)$.

Parameter estimation and predictive comparison. We estimate the parameters of the LGCP model and the Poisson geostatistical models in Section 2.3.1 using the Monte Carlo maximum likelihood (MCML) method[46], implemented in the PrevMap package[47]. More technical details on how the SPDE approach, used for approximation of the Gaussian process $S(x)$, and the MCML method are implemented in order to obtain the parameter estimates can be found in Chapter 7 of[48]. We wish to compare the resulting predictions for $\mu_{RSD}(x,t)$ and $\mu_{CSD}(x,t)$ at locations and times for which both sources provide enough information, noting that $\mu_{CSD}(x,t)$ corresponds to the average species richness adjusted for preferential sampling bias in CSD. We then set $t = 1998$ and consider locations x such that the estimated spatial correlation between $U_j(x)$ and $U_j(x_c)$ is no less than 0.75, with x_c denoting the sampled location from either of two sources that is closest to x, for $j = RSD, CSD$. We then summarise the discrepancy between $\mu_{RSD}(x,t)$ and $\mu_{CSD}(x,t)$ through their relative difference

$$RD(x) = \{\hat{\mu}_{CSD}(x, t) - \hat{\mu}_{RSD}(x, t)\}/max\{\hat{\mu}_{CSD}(x, t), \hat{\mu}_{RSD}(x, t)\}.$$

In computing $\hat{\mu}_{CSD}(x, t)$, the number of collectors is set to $C_i = 2$. As noted in Section 2.1, we assessed that the two collectors from the CSD acted as one in view of their strong interaction during the data collection.

Results

The triangulation mesh used for the SPDE approximation in the LGCP model is provided in the Appendix (Fig. A.2). The parameter estimates in Table 1 indicate that the data collectors, on average, visited locations closer to home and of higher natural value more intensely. In particular, we find that evaluation of nature has the strongest effect, estimating that a location within an environmentally 100% valuable environment has a sampling intensity about ($exp\{2.778\} \approx$) 16 times larger, with the other variables kept equal. However, the predicted surface from the LGCP model (Fig. 3) suggests the presence of additional factors affecting the two collectors' choice of locations. The map shows that a large swathe in the south of Limburg, both far from the collectors' homes and of low natural value, was also sampled with moderate intensity.

For the CSD, we did not find evidence of an association between the predictive mean of the residual spatial process $S(x)$ and species richness, as indicated by the non-significant estimate for the parameter γ at the conventional 5% level ($\hat{\gamma} = -0.031$, 95% confidence interval $[-0.106; 0.044]$). As expected, the mean level of species

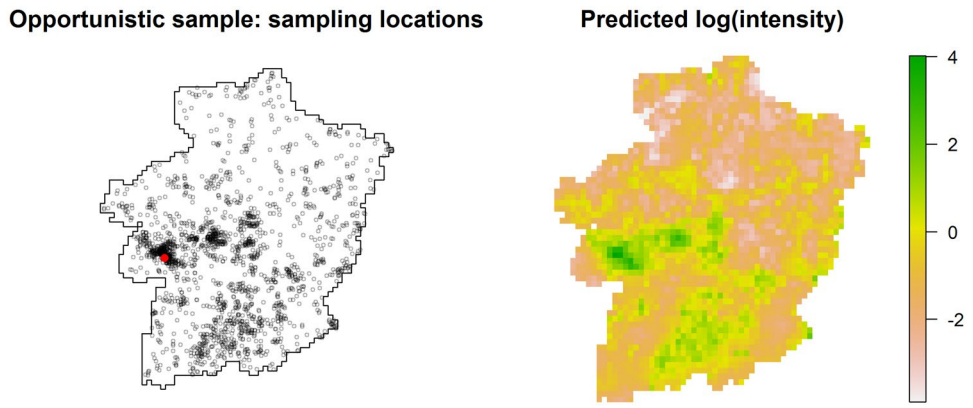

Figure 3. Opportunistic sample. Left: observed sampling locations, with random jitter and a red dot indicating the central point of the collectors' residential municipality. Right: log-transformed intensities, predicted by the log-Gaussian Cox model.

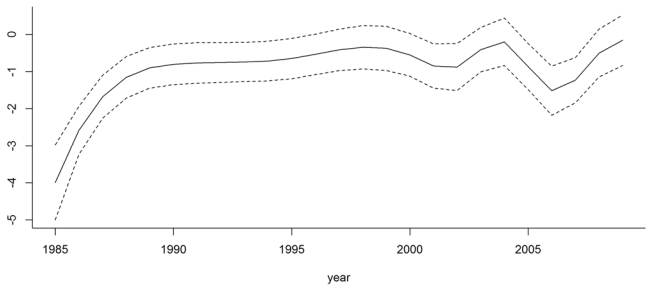

Figure 4. Fitted spline effect (solid line) in the CSD geostatistical model, along with 95% confidence intervals (dashed lines).

	parameter	estimate	95% C.I.	estimate	95% C.I.
		CSD		RSD	
intercept	β_0	−3.990	[−5.003;−2.978]	1.541	[1.189;1.894]
wetness	β_1	0.627	[0.166;1.089]	−0.295	[−0.566;−0.025]
value of nature	β_2	−0.365	[−0.948;0.217]	0.366	[0.078;0.653]
forest cover	β_3	0.984	[0.161;1.627]	0.193	[−0.074;0.461]
spatial variance	$\log(\sigma^2)$	−1.060	[−1.602;−0.518]	−2.379	[−3.354;−1.404]
scale	$\log(\phi)$	1.686	[0.747;2.625]	2.596	[1.245;3.948]
nugget	$\log(\tau^2)$	0.845	[−0.256;1.946]	−3.146	[−5.318;−0.975]

Table 2. Estimates and 95% confidence intervals of parameters common to the CSD and RSD geostatistical models for species richness. Comments on the temporal trends, and the effects of group size in the CSD analysis, not reported in this table, are provided in the main text.

richness significantly increases as the number of collectors also increases ($\widehat{\beta_4} = 1.905$, 95% confidence interval [1.513; 2.396]). The spline function fitted in the CSD model is shown in Fig. 4. The time effect is almost flat during the years in which the RSD were collected (1997 to 1999). For the RSD, the linear time effect was negative ($\widehat{\beta_5} = -0.083$, 95% confidence interval [−0.138; −0.028]). Table 2 and Fig. 5 provide comparisons for the estimates of parameters used in the geostatistical models for both CSD and RSD. We found contrasting results. In the CSD, the estimated regression coefficients for wetness and forest cover were both positive, while the effect of evaluation of nature was not significant. For RSD, these three effects were significantly negative, non-significant and significantly positive, respectively.

Figures 6 and 7 show the predicted spatial surfaces for $\hat{\mu}_{CSD}(x, t)$ and $\hat{\mu}_{RSD}(x, t)$ for $t = 1998$. Predictions from the CSD show higher values in species richness in the northern part of Limburg, while the RSD predictions point to increased values in the north-western and central-eastern parts of Limburg. In addition, Figures 6 and 7 show increased estimation error in the CSD-based predictions. Fig. 8 shows the relative difference at locations where both CSD and RSD could be assumed to be informative (see Section 2.3.2 for more details). We observe that

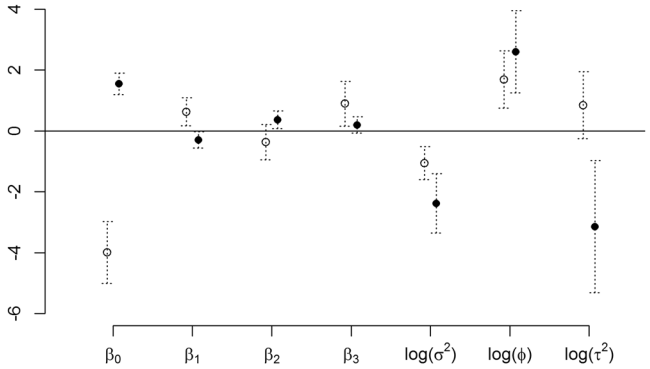

Figure 5. Visual comparison of the point estimates (open circles for the citizen science data (CSD); solid circles for the random study data (RSD)) and 95% confidence intervals (dotted lines) for the parameters common to two geostatistical models that were fitted to the CSD and RSD.

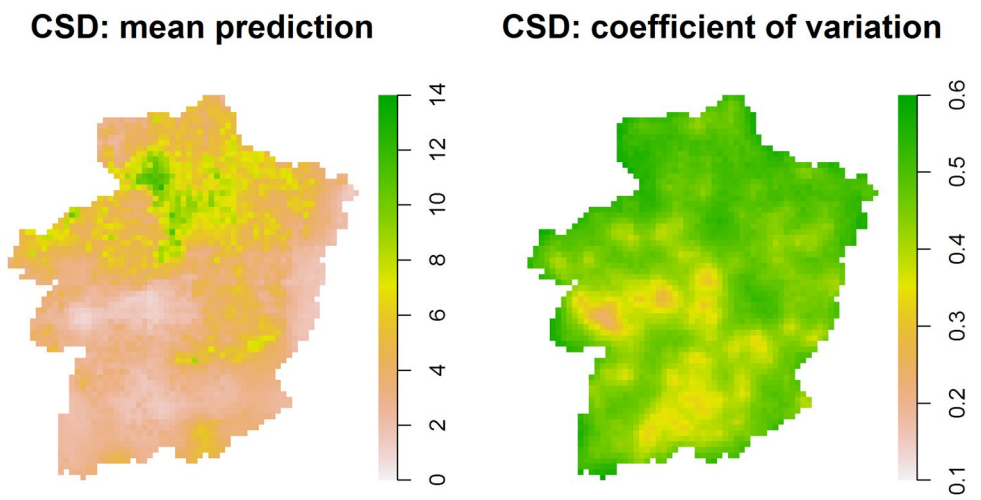

Figure 6. Mean forest ground-floor bryophyte species richness prediction and coefficient of variation (standard error/mean prediction) for 1998 and 2 data collectors, based on the CSD.

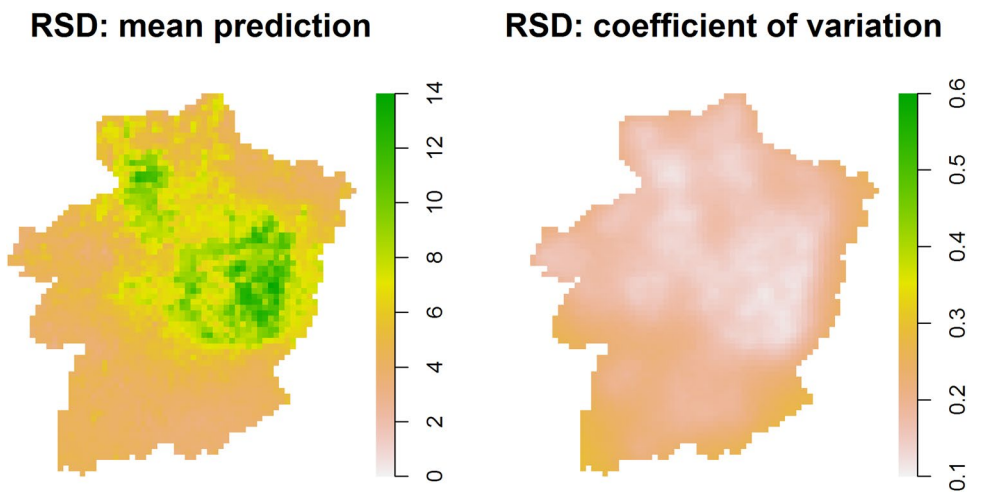

Figure 7. Mean forest ground-floor bryophyte species richness prediction and coefficient of variation (standard error/mean prediction) for 1998, based on the RSD.

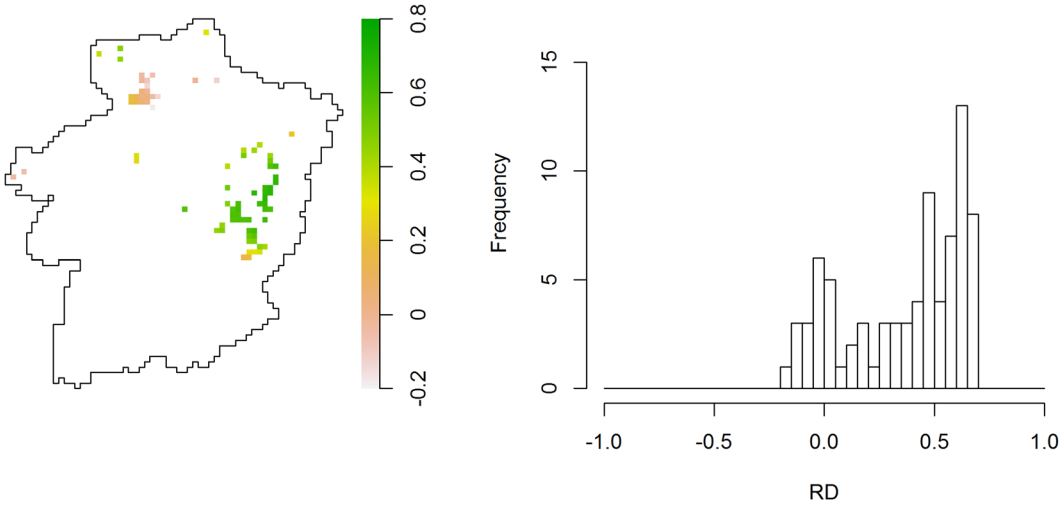

Figure 8. A map and histogram of relative differences ($RD(x)$) between mean predictions for 1998, based on CSD ad RSD, for locations that were visited in both data collections.

overall the CSD provide lower predictions of species richness than the RSD, with relative differences ranging between −0.15 and 0.75.

Discussion

We have assessed the reliability of opportunistic samples on forest ground-floor bryophyte species richness collected by two citizens in Limburg, Belgium, using data from an expert source who adheres to a randomised sampling regime as the gold-standard. In the first part of our analysis, we have fitted a log-Gaussian Cox process (LGCP) model to investigate spatial factors that might have affected the citizens' choice of sampling locations. Our results indicate that locations with a higher value of nature and those located closer to the residential location of the collectors were sampled more intensely than others. However, the fitted LGCP model also suggests that there may be additional factors driving the sampling behaviour of the two citizens. These might be related to personal interests in a specific species or taxonomic group, as documented by Boakes *et al.*[49]. Other authors have highlighted that environmental features that facilitate monitoring often play an important role in citizens' sampling choices. Hadj-Hammou *et al.*[13] observed that citizens visited sampling locations with habitat types that are easily accessible and abundant throughout the study region more frequently than other locations. Ease-of-access has also been shown to be important in the opportunistic placement of camera traps[50].

In the second part of the analysis, we have carried out spatial prediction of bryophyte species richness by fitting geostatistical Poisson models. For the citizen-science data, or CSD, we have accounted for potential bias induced by preferential sampling[32] through the inclusion of the predictive mean of the spatial Gaussian process from the LGCP model into the linear predictor for species richness. This allows us to account for spatial factors that affect both the likelihood of a location being sampled and the variation in species richness. However, since this did not show any significant effect, we conclude that, in this case, preferential sampling may not capture all the main features of the opportunistic sampling process in the CSD. Other sources of bias might not be spatially structured and can only be accounted for when they are accurately reported by the data collector, which is not the case in our study. Here, group size was the only available variable that could be used as proxy for the heterogeneity in the collectors' effort and it was found to have a significant effect on species richness in the CSD. In the CSD analysis, we found positive effects of wetness and forest cover. In contrast, the results from the randomised study (RSD) indicated a negative effect of the wetness index and a positive effect of environmental value. Finally, the spatial estimates for species richness differed greatly at many locations between the CSD and RSD, especially in the eastern parts of Limburg. Note that the estimates were generally smaller in the CSD than in the RSD.

Following the classification by Isaac *et al.*[12] of the sources of bias in CSD, we draw the following conclusions. The spatial unbalance of the data was not an issue in our analysis, since our approach, unlike, for example, occupancy-detection models, does not require data to be aggregated over space. We account for uneven sampling effort by including group size as a covariate in the geostatistical model. However, we were not able to account for additional likely sources of uneven sampling effort as these were not reported. Examples of these are the time spent in the field and the trajectory covered by the citizens, two sampling parameters that are instead controlled for in the RSD's study design. Another limitation of our approach is that it does not account for species- and observer-specific detectability. To account for species-specific detectability, we would require accurate prior information on the probability of detection and how this varies in space for each species so as to incorporate it into a model for the presence/absence of each species. Dorazio and Royle[51] and Dorazio *et al.*[52] predict species richness based on joint species-specific site-occupancy models, in which a species' site-occupancy is modelled as a mixture of a Bernoulli process that determines the likelihood of the species' presence or absence at a given location and a second Bernoulli process regulating how likely that species is to be observed at that location given its presence. The method requires a well-structured study protocol with site re-visits, both of which were unavailable for the CSD. Since only two collectors, occasionally joined by other citizens, carried out the collection

of the CSD, variation in the observer-specific detectability skills is less problematic in our case. This is known to cause inferential issues when data are collected by a highly heterogeneous group of citizens[24]. Furthermore, in our study, collectors in both the CSD and RSD had extensive knowledge of bryophyte systematics and we therefore assume that they all have similar detectability skills. This is arguably a strong assumption, since observer-specific detectability has been shown to vary considerably, even among the most skilled data collectors[53].

Most of the challenges in the analysis of CSD, especially in the active contributory model[10] arise from the lack of a well-defined scientific hypothesis that might provide guidance for the sampling design. The underestimation in bryophyte species richness reported in our analysis of the CSD may be due to the fact that the two investigators were less likely to engage in the close inspection of species on the forest floors, which was the main focus of the RSD study. However, this can as well be an artefact of using the set of species that were collected in the RSD study to extract forest ground-floor bryophytes from the CSD data, which implied that forest ground-floor bryophyte species that were identified by the citizens, but not in the expert study, were not considered for the analysis.

A secondary issue in our study is that the resolution in which the spatial and/or temporal variation is recorded does not align with the resolution at which the true ecological process is at play; e.g., spatial trends in moss species richness will probably vary considerably within each 1km^2 grid cell, but we are unable to investigate this. We face this problem in both CSD and RSD analyses, but it generally poses a difficulty in historical CSD surveys, which have become popular sources to investigate long-term ecological trends. This spatial misalignment can be one of the reasons why an ecological covariate such as value of nature was found to contribute less to the variability in the CSD outcome than expected.

We conclude that in general, we cannot trust inferences that are drawn purely from CSD that were collected without adherence to a strict sampling protocol, due to multiple unreported sources of sampling bias. These may be difficult, if not impossible, to account for in absence of detailed information on the sampling procedures adopted by the citizens. This does not imply that CSD within the active contributory model cannot provide useful information for estimating biodiversity, but rather that standard modelling approaches will be prone to failure. A more promising approach would be to combine the imperfect information from CSD with gold-standard data that can deliver unbiased spatial estimates. This has been achieved in the context of disease mapping[54,55] where joint geostatistical models have been developed in order to remove bias from opportunistically collected samples by analyzing these jointly with data from randomised prevalence surveys. However, we could not apply this modelling framework in our study due to the relatively small temporal overlap between CSD and RSD, which does not allow to reliably estimate the spatially varying bias of the CSD.

Data availability

The data are available at https://doi.org/10.5061/dryad.brv15dv5r.

Received: 7 March 2019; Accepted: 26 November 2019;
Published online: 13 December 2019

References

1. Schmeller, D. *et al*. Advantages of volunteer-based biodiversity monitoring in Europe. *Conserv. Biol.* **23**, 307–316 (2009).
2. Dickinson, J. L. *et al*. The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.* **10**, 291–297 (2012).
3. Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J. & Martin, T. G. Realising the full potential of citizen science monitoring programs. *Biol. Cons.* **165**, 128–138 (2013).
4. Theobald, E. J. *et al*. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biol. Cons.* **181**, 236–244 (2015).
5. Thomas, C. D. & Lennon, J. J. Birds extend their ranges northward. *Nature* **399**, 213–213 (1999).
6. Jenni, L. & Kerry, M. Timing of autumn bird migration under climate change: advances in long-distance migrants, delays in short-distance migrants. *Proc. Biol. Sci.* **270**, 1467–1471 (2003).
7. Zuckerberg, B. & Porter, W. F. Thresholds in the long-term responses of breeding birds to forest cover and fragmentation. *Biol. Cons.* **143**, 952–962 (2010).
8. Warren, M. S. *et al*. Rapid responses of British butterflies to opposing forces of climate and habitat change. *Nature* **414**, 65–69 (2001).
9. Shumba, T., Montgomery, R. A., Rasmussen, G. S. A. & Macdonald, D. W. African wild dog habitat use modelling using telemetry data and citizen scientist sightings: are results comparable? *Afr. J. Wildl. Res.* **48**, 1–13 (2018).
10. Price, C. A. & Lee, H.-S. Changes in participants' scientific attitudes and epistemological beliefs during an astronomical citizen science project. *J. Res. Sci. Teach.* **50**, 773–801 (2013).
11. Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.* **41**, 149–172 (2010).
12. Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* **5**, 1052–1060 (2014).
13. Hadj-Hammou, J., Loiselle, S., Ophof, D. & Thornhill, I. Getting the full picture: Assessing the complementarity of citizen science agency monitoring data. *PLoS ONE* **12**, e0188507, https://doi.org/10.1371/journal.pone.0188507 (2017).
14. Fitzpatrick, M., Preisser, E., Ellison, A. & Elkinton, J. Observer bias and the detection of low-density populations. *Ecol. Appl.* **19**, 1673–1679 (2009).
15. de Solla, S. R. *et al*. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biol. Cons.* **121**, 585–594 (2005).
16. Delaney, D. G., Sperling, C. D., Adams, C. S. & Leung, B. Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biol. Invasions* **10**, 117–128 (2008).
17. Boudreau, S. & Yan, N. Auditing the accuracy of a volunteer-based surveillance program for an aquatic invader *Bythotrephes*. *Environ. Monit. Assess.* **91**, 17–26 (2004).
18. van Strien, A. J., van Swaay, C. A. M. & Termaat, T. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *J. Appl. Ecol.* **50**, 1450–1458 (2013).
19. Carota, C., Nava, C. R., Chiglione, C. & Schiaparelli, S. A Bayesian semiparametric GLMM for historical and newly collected presence-only data: An application to species richness of Ross Sea Mollusca. *Environmetrics* **28**, e2462, https://doi.org/10.1002/env.2462 (2017).

20. Roy, H. E. *et al*. Invasive alien predator causes rapid declines of native European ladybirds. *Divers. Distrib.* **18**, 717–725 (2012).
21. Kuussari, M., Heliölä, J., Pöyry, J. & Saarinen, K. Contrasting trends of butterfly species preferring semi-natural grasslands, field margins and forest edges in northern Europe. *J. Insect Cons.* **11**, 351–366 (2007).
22. Szabo, J. K., Vesk, P. A., Baxter, P. W. J. & Possingham, H. P. Regional avian species declines estimated from volunteer-collected long-term data using List Length analysis. *Ecol. Appl.* **20**, 2157–2169 (2010).
23. Kelling, S. *et al*. Can observation skills of citizen scientists be estimated with species accumulation curves? *PLoS ONE* **10**, e0139600, https://doi.org/10.1371/journal.pone.0139600 (2015).
24. Johnston, A., Fink, D., Hochachka, W. M. & Kelling, S. Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* **9**, 88–97 (2018).
25. MacKenzie, D. I. *et al. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence* (Elsevier, 2006).
26. Royle, J. A. & Kéry, M. A Bayesian state-space formulation of dynamic occupancy models. *Ecology* **88**, 1813–1823 (2007).
27. Chandler, R. B. *et al*. Spatial occupancy models for predicting metapopulation dynamics and viability following reintroduction. *J. Appl. Ecol.* **52**, 1325–1333 (2015).
28. Pacifici, K., Reich, B. J., Dorazio, R. M. & Conroy, M. J. Occupancy estimation for rare species using a spatially adaptive sampling design. *Methods Ecol. Evol.* **7**, 285–293 (2017).
29. Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M. & Fortin, M.-J. Spatial autoregressive models for statistical inference from ecological data. *Ecol. Monogr.* **88**, 36–59 (2018).
30. Pacifici, K. *et al*. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* **98**, 840–850 (2017).
31. Conn, P. B., Thorson, J. T. & Johnson, D. S. Confronting preferential sampling when analyzing population distributions: diagnosis and model-based triage. *Methods Ecol. Evol.* **8**, 1535–1546 (2017).
32. Diggle, P. J., Menezes, R. & Su, T. Geostatistical inference under preferential sampling. *J. Royal Stat. Soc. C* **59**, 191–232 (2010).
33. Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M. & Silander, J. A. Point pattern modelling for degraded presence-only data over large regions. *J. Royal Stat. Soc. C* **60**, 757–776 (2011).
34. Diggle, P. J., Tawn, J. A. & Moyeed, R. A. Model-based geostatistics. *J. Royal Stat. Soc. C* **42**, 3 (1998).
35. Netwerk Natuuronderzoek. *Limburgse Biodiversiteitsatlas* http://biodiversiteitlimburg.be/atlas/ (2018).
36. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
37. Afdeling Bos & Groen *De Bosinventarisatie Van Het Vlaamse Gewest. Resultaten Van De Eerste Vlaamse Bosinventarisatie 1997–1999*. (Ministerie van de Vlaamse Gemeenschap, https://www.natuurenbos.be/beleid-wetgeving/natuurbeheer/bosinventarisatie) (2001).
38. Van Ranst, E. & Sys, C. *Eenduidige Legende Voor De Digitale Bodemkaart Van Vlaanderen (Schaal 1:20 000)*. (Laboratorium Voor Bodemkunde, 2000).
39. De Saeger, S. *et al. Biologische Waarderingskaart En Natura 2000 Habitatkaart: Uitgave 2016*. (Instituut Natuur- en Bosonderzoek, 2016).
40. Informatie Vlaanderen. *Geopunt* http://www.geopunt.be (2000).
41. Pati, D., Reich, B. J. & Dunson, D. B. Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **98**, 35–48 (2011).
42. Matérn, B. *Spatial Variation* (2nd ed.) (Springer, 1986).
43. Zhang, H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Stat. Assoc.* **99**, 250–261 (2004).
44. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Royal Stat. Soc. B* **73**, 423–498 (2011).
45. Krainski., E. T. *et al*. Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA (CRC Press, 2018).
46. Geyer, C. J. & Thompson, E. A. Constrained Monte Carlo Maximum likelihood for dependent data (with discussion). *J. Royal Stat. Soc. B* **54**, 657–699 (1992).
47. Giorgi, E. & Diggle, P. J. PrevMap: an R package for prevalence mapping. *J. Stat. Softw.* **78**, 1–29 (2017).
48. Diggle, P. J. & Giorgi, E. Model-based Geostatistics for Global Public Health: Methods and Applications (CRC Press, 2019).
49. Boakes, E. H. *et al*. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Sci. Rep.* **6**, 33051, https://doi.org/10.1038/srep33051 (2016).
50. Burton, A. C. *et al*. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *J. Appl. Ecol.* **52**, 675–685 (2015).
51. Dorazio, R. M. & Royle, J. A. Estimating size and composition of biological communities by modeling the occurrence of species. *J. Am. Stat. Assoc.* **100**, 389–398 (2005).
52. Dorazio, R. M., Royle, J. A., Söderström, B. & Glimskär, A. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* **87**, 842–854 (2006).
53. Britton, A., Mitchell, R., Potts, J. & Genney, D. Developing monitoring protocols for cost effective surveillance of lichens. *Lichenologist* **46**, 471–482 (2014).
54. Giorgi, E., Sesay, S. S., Terlouw, D. J. & Diggle, P. J. Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *J. Royal Stat. Soc. A.* **178**, 445–464 (2015).
55. Diggle, P. J. & Giorgi, E. Model-based geostatistics for prevalence mapping in low-resource settings (with discussion). *J. Am. Stat. Assoc.* **111**, 1096–1120 (2016).

Acknowledgements

The Belgian Nature and Forest Agency (ANB), the Institute for Nature and Forest Research (INBO), and the Umbrella for Nature Research in Limburg (LIKONA) are gratefully acknowledged for providing the data used in this study and commenting on our results. In particular, we thank the following persons for their support and insights: Cécile Nagels (LIKONA), Luc Crevecoeur (LIKONA), Wouter Van Landuyt (INBO), Dirk De Beer (INBO), and Martine Waterinckx (ANB). The largest part of this study has been conducted when Thomas Neyens was funded as a postdoctoral researcher by the Flemish Research Foundation (12S7217N).

Author contributions

T.N. carried out the analysis and wrote the first draft of the manuscript. E.G. implemented the models in the R software environment. T.N., P.J.D., C.F., and E.G. developed the modelling framework of the statistical analysis. N.B. and T.A. provided the necessary ecological interpretations. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55593-x.

Correspondence and requests for materials should be addressed to T.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.