A Probabilistic Generating Process of Citizen Science Data: Modeling and Estimation of Parameters

Kwaku Peprah Adjei[1] and Jorge Sicacha-Parada[2] and Bob O'Hara[3] and Ingelin Steinsland[4]

[1] *Norwegian University of Science and Technology (NTNU), Norway, kwaku.p.adjei@ntnu.no*
[2]*Norwegian University of Science and Technology (NTNU), Norway, jorge.sicacha@ntnu.no*
[3]*Norwegian University of Science and Technology (NTNU), Norway, bob.ohara@ntnu.no*
[4]*Norwegian University of Science and Technology (NTNU), Norway, ingelin.steinsland@ntnu.no*

Citizen Science refers to the open engagement of the public in scientific activities. For example, several biodiversity projects encourage regular citizens to report the species they observe. Enabling citizens to feed databases enlarges the spatial coverage and the temporal resolution of biodiversity data. However, it comes at the risk of having biased unstructured sampling designs, information focused on more easily detectable species and misspecification of the species observed, among others.

Given the inherent biases to Citizen Science data, making inference with them needs to account for these biases. In order to do it, we propose a probabilistic generating process of Citizen Science data. It starts by assuming that true occurrence of species is a Spatial Point Pattern. Then, three typical sources of bias in Citizen Science are regarded as sources of thinning for the true point pattern. These are: the sampling process of Citizen Scientists; the detection probability which is characteristic of each species, and the misclassification of the observed species.

Recently, Machine Learning has become a popular approach for Citizen Science in biodiversity. Another interesting feature of our research involves incorporating species classification probabilities from Machine Learning algorithms into our statistical model framework. Using a Bayesian approach, we aim to make inference of key parameters that explain the ecological process of each species.

In this work, we will simulate a three-stage thinning procedure for the occurrences of a given species and we will make use of the model we propose to estimate the true ecological parameters.

Central Limit Theorems for Ornstein–Uhlenbeck processes on Yule trees

Krzysztof Bartoszek[1] **and Torkel Erhardsson**[2]

[1] *Linköping University, Sweden, krzysztof.bartoszek@liu.se; krzbar@protonmail.ch*
[2] *Linköping University, Sweden, torkel.erhardsson@liu.se*

The field of phylogenetic comparative methods, the study of traits on the between species level, that takes into account shared evolutionary history, can be viewed as a biological application of branching Markov processes. The Ornstein–Uhlenbeck process is the current workhorse of the continuous–time–continuous–trait modelling setup. A key question is then its asymptotic behaviour. Central Limit Theorems (CLTs) for the sample average have been found in general situation [1, 5].

Alternatively, if one specializes on the pure birth tree model for the phylogeny one can find such CLTs through careful consideration of the coalescent times [2, 4]. The key property used when finding these CLTs is that conditional on knowing the phylogeny, the observations are jointly Gaussian. Hence, an application of Stein's method to this setting would provide the CLTs alongside rates of convergence. We find, based on Stein's method, general bounds in the Kolmogorov and Wasserstein distances between mixtures of normal distributions and a limiting normal approximations that are expressed as functions of the first two conditional moments [3]. We apply these to Ornstein–Uhlenbeck processes evolving on pure–birth trees. Not only do these bounds confirm previous weak convergence results, but they also provide rates of convergence to the limiting distribution.

KB's research is supported by the Swedish Research Council (Vetenskapsrådet) grant no. 2017–04951.

References

[1] Adamczak, R. and Miłoś, P. (2015). CLT for Ornstein–Uhlenbeck branching particle system. *Electronic Journal of Probability*, **20**, 1 – 35.

[2] Bartoszek, K. (2020). A Central Limit Theorem for punctuated equilibrium. *Stochastic Models*, **36**, 473 – 517.

[3] Bartoszek, K. and Erhardsson, T. (in press). Normal approximation for mixtures of normal distributions and the evolution of phenotypic traits. *Advances in Applied Probability*.

[4] Bartoszek, K. and Sagitov, S. (2015). Phylogenetic confidence intervals for the optimal trait value. *Journal of Applied Probability*, **52**, 1115 – 1132.

[5] Ren, Y.–X., Song, R. and Zhang, R. (2014). Central limit theorems for supercritical branching Markov processes. *Journal of Functional Analysis*, **266**, 1716 – 1765.

Model-based inference for abundance estimation using presence/absence data from large-area forest inventories together with covariate data from remote sensing

Benoît Gozé[1], Magnus Ekström[2], Göran Ståhl[3], Bengt-Gunnar Jonsson[4], Jörgen Wallerman[5], Saskia Sandring[6], Jonas Dahlgren[6]

[1] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, benoit.goze@slu.se*

[2] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, magnus.ekstrom@slu.se*

[3] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, goran.stahl@slu.se*

[4] *Department of Natural Sciences, Mid Sweden University, Bengt-Gunnar.Jonsson@miun.se*

[5] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, jorgen.wallerman@slu.se*

[6] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, saskia.sandring@slu.se*

[7] *Department of Forest Resource Management, Swedish University of Agricultural Sciences,jonas.dahlgren@slu.se*

In this paper, we investigate methods to estimate plant population size and intensity with help from presence/absence data. Presence/absence sampling is a useful and relatively simple method for monitoring state and change of plant species communities. Moreover, it has advantages compared to traditional plant cover assessment, the latter being more prone to surveyor judgement error. We use inhomogeneous Poisson point process models concerning plant locations, and generalised linear models (GLM) with a complementary log-log link function for linking presence/absence data to plant intensity. In these models, auxiliary covariate information coming from remote sensing (i.e. wall-to-wall data) are used. We estimate plant intensity for a selection of forest plants, as well as plot probabilities of presence of these species for parts of Sweden. The estimations of plant population sizes and intensities are made not only locally but also for larger regions. An estimator of the variance of the estimator of the expected number of plants in an area of interest is also proposed. For evaluating the estimators, we use both Monte-Carlo simulations, where we create artificial plant populations, and empirical data from the Swedish National Forest Inventory (NFI). We also develop a test for our models, to check the underlying Poisson point process model assumption and protect inference against model misspecification. The suggested hypothesis test is evaluated through Monte-Carlo simulations and shows good power properties.

Joint modelling of the peatland vegetation cover using non-stationary multivariate Gaussian processes

Juho Kettunen[1], **Lauri Mehtätalo**[2], **Eeva-Stiina Tuittila**[3], **Aino Korrensalo**[4] **and Jarno Vanhatalo**[5]

[1] *University of Eastern Finland, Finland, juho.kettunen@uef.fi*
[2] *University of Eastern Finland, Finland, lauri.mehtatalo@uef.fi*
[3] *University of Eastern Finland, Finland, eeva-stiina.tuittila@uef.fi*
[4] *University of Eastern Finland, Finland, aino.korrensalo@uef.fi*
[5] *University of Helsinki, Finland, jarno.vanhatalo@helsinki.fi*

Knowledge of the amount and structure of the vegetation are often needed inputs for modeling e.g. the greenhouse gas exchange and for ecosystem models in general. Collecting vegetation cover data through ground inventory surveys is time consuming and identifying species requires specific expertize from the field crew. Therefore, the sample plots are typically small and sampling fraction is low. However, based on such data, one should be able to produce maps on distributions of species, estimate species total abundance as well as provide uncertainty estimates. We propose a novel Gaussian process-based Bayesian hierarchical joint species distribution model to estimate areal vegetation percentage cover of sphagnum mosses and vascular plants on peatland. The proposed model improves the state-of-the-art join species distribution models [2, 1] in three ways. 1) We use the Dirichlet-Multinomial distributions to model interspecific exclusion between competing species as well as the uncertainty in observation process; 2) we use non-stationary multivariate Gaussian process to describe species specific niche preference over the study area; and 3) we propose novel approaches to validate and compare the predictive performance of joint species distribution models. We demonstrate the model by analysing the vegetation cover of an extensively studied boreal minerotrophic fen, which is part of Siikaneva peatland study site in Southern Finland. The area is part of Integrated Carbon Observation System (ICOS) network. Our results improve the total vegetation cover estimates compared to the existing species distribution models and provide quantitative uncertainty estimates for them for the first time. Our novel methods allow us to infer also the species interactions. We present maps of the species covarage and show that our new method performs better than the existing species distribution models that are commonly used in similar tasks.

References

[1] Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling With Applications in R*, Cambridge University Press, Cambridge

[2] Vanhatalo, J., Hartmann, M. and Veneranta, L. (2020). Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data. *Bayesian analysis*, **15**, 415 – 447.