Integrating hierarchical statistical models and machine learning algorithms for ground-truthing drone images of the vegetation: taxonomy, abundance and population ecological models

Christian Damgaard, Bioscience, Aarhus University, Vejlsøvej 25, 8600 Silkeborg

1 Abstract

In order to fit population ecological models, e.g. plant competition models, to the new drone-aided image data, we need to develop statistical models that may take the new type of measurement uncertainty into account and quantify its importance for statistical inferences and ecological predictions. Here, it is proposed to quantify the uncertainty and bias of image predicted plant taxonomy and abundance in a hierarchical statistical model that is linked to ground-truth data obtained by the pinpoint method. It is critical that the error rate in the species identification process is minimized when the image data are fitted to the population ecological models, and several avenues for reaching this objective are discussed.

2 Introduction

Using drones that record multi-spectral photography and LIDAR, it has now become possible to obtain spatio-temporal ecological data at a fine-scaled resolution. These new data collection possibilities provide a quantum leap compared to earlier methodologies for monitoring ecological processes, e.g. competitive plant growth (e.g. Tay et al. 2018). However, in order to use the new drone-aided image data types for modelling plant ecological processes, there is a need to develop statistical models that are especially tailored towards these new image data types.

Plant competition is a population ecological proces, where plant growth is reduced by the presence of neighbouring plants. When investigating interspecific interactions in light-open vegetation, where it often is difficult to distinguish individual plants, the population growth of a species is modeled as a function of the local abundance of other species (Adler et al. 2009, Damgaard 2011). Previously, plant competitive interactions have been modelled using non-destructive measurements of plant abundance, e.g. using pin-point data, where the vertical density (number of times a plant species is touched by a thin pin) is recorded several times during the growing season in permanent plots. Vertical density is correlated to plant biomass (Jonasson 1983), and plant growth and interspecific interactions may consequently be estimated from repeated pin-point measurements of vertical density (Damgaard et al. 2009, Damgaard et al. 2011, 2014, Merlin et al. 2015, Ransijn et al. 2015). However, it is now possible to radically upscale the non-destructive measurements of plant abundance by repeated drone-aided recordings of multi-spectral and LIDAR image data of the vegetation. The new image data encompass vast possibilities, but also a new challenge. Compared to pin-point data, which is assembled by persons trained in plant taxonomy, the new image data come without plant taxonomic information or abundance measures.

Currently, image data from drones are being collected in several plant ecological laboratories, and valuable experience on how to recognize plant species is being collected. It seems to be a natural choice to use machine learning algorithms for fiting the information from the new image data to observed ground truth of species taxonomy or abundance and, currently, research is focussed on how best to use such machine-learning algorithms for predicting purposes in plant ecology (e.g. Sun et al. 2017, Tay et al. 2018). The aim of this study is to discuss the use of such predictions obtained by machine learning algorithms for fitting empirical population ecological models, e.g. competition models. More specifically, the aim is to to specify statistical models that will allow us to quantify the possible bias and uncertainties of species identification and abundance predictions obtained by machine learning algorithms, so that image data may be used to fit population ecological models with a known degree of uncertainty.

Here, it is proposed to use the confusion matrix of the chosen machine learning algorithm for quantifying the uncertainty when identifying species taxonomy and integrate a Bayesian hierarchical modelling approach with machine learning algoritms for quantifying the uncertainty when estimating species abundance. In this study, the proposed general statistical model will be outlined and tentatively specified with suggested relevant statistical distributions. The developed statistical models are needed for fitting population ecological models of plant communities and make quantitative ecological predictions of plant community and ecosystem dynamics, including quantitative assessments of the proces or structural uncertainty.

3 Methods and models

3.1 Pinpoint data – vertical density

In a number of ground-truthing plots at a natural or semi-natural habitat site with light-open vegetation, plant species taxonomic identity and abundance is determined by the pinpoint method. A pinpoint frame with *n* grid points is placed in the vegetation and the position of the frame is recorded using high-accuracy GPS. At each grid point, a thin pin is inserted into the vegetation and the sequence in which different plant species touch the pin is recorded. Such sequence pinpoint data allow the determination of several derived plant abundance measures, e.g. cover, top cover, and vertical density at the spatial resolution of a single pin, a number of neighboring pins, and the plot. Furthermore, it is possible to aggregate the species data to higher taxonomic levels or species groups.

Depending on the vegetation and the studied ecological question, various measures of plant abundance may be relevant, but here we will focus on the vertical density at the spatial level of the plot. Importantly, it is assumed that the pinpoint measure of vertical density is an unbiased sample of the true, but unknown, vertical density.

3.2 Machine learning algorithms of image data

A drone is used to record multi-spectral images and LIDAR data of the site with the ground-truthing plots at a resolution that is sufficient to compare the image data with the pinpoint data. Using standard image software, e.g. *Agisoft*, a 3D model of the site is constructed and the information of the different bands is summarized at the approximate position of each pin in the pinpoint frame. Using supervised machine-learning algorithms, the taxonomic identity and vertical density of each species is predicted from the image data at the spatial resolution of the plot.

The species taxonomic identity is predicted from the information in the multi-spectral image bands as well as information on texture etc. In species-rich plant communities, it is to be expected that not all species can be distinguished with sufficient accuracy, and species that cannot be reliably distinguished are aggregated into a common species group. Since the overall objective of the proposed statistical method is to fit plant population ecological models, it is more important that all plants are accounted for than that each species is identified precisely. Furthermore, when constructing plant population ecological models of species-rich plant communities it is typically necessary to aggregate plant species into plant species groups or functional groups anyway (e.g. Damgaard 2015). In the following, the term species may either mean a single plant species or a group of plant species.

The vertical density of each species is predicted using the 3D modelling of the vegetation and LIDAR data. It is assumed that the vertical density predicted from the image data may be a biased sample of the true, but unknown, vertical density, and that the direction and magnitude of the bias is species-specific.

3.3 Statistical models

By aggregating species with similar image information, using different auxiliary information, e.g. time series image data, and different supervised machine-learning algorithms, it is possible to maximize the probability of correct species identification. However, there will always be a non-zero probability of false identification. The probabilities of falsely identifying an entity of vertical density to a wrong species is called a confusion matrix, which is a right stochastic matrix, or transition matrix, of real numbers, where each row sums to one. If all species are correctly identified, then the confusion matrix is the identity matrix. The confusion matrix is fitted using the data from the ground-truth plots and is, consequently, susceptible to sampling errors, and it is here assumed that each row in the confusion matrix is distributed according to a Dirichlet distribution ($M1$).

$$M1: \boldsymbol{p_i} \sim Dir(\boldsymbol{\alpha_i}) \qquad (1),$$

where $\boldsymbol{p_i}$ is a row vector of p_{ik}, which are the probabilities of classifying species *i* as species *k*, and $\boldsymbol{\alpha_i}$ is a row vector of α_{ik}, which are the number of times species *i* is categorized as species *k* by the supervised machine-learning algorithm (Frigyik et al. 2010).

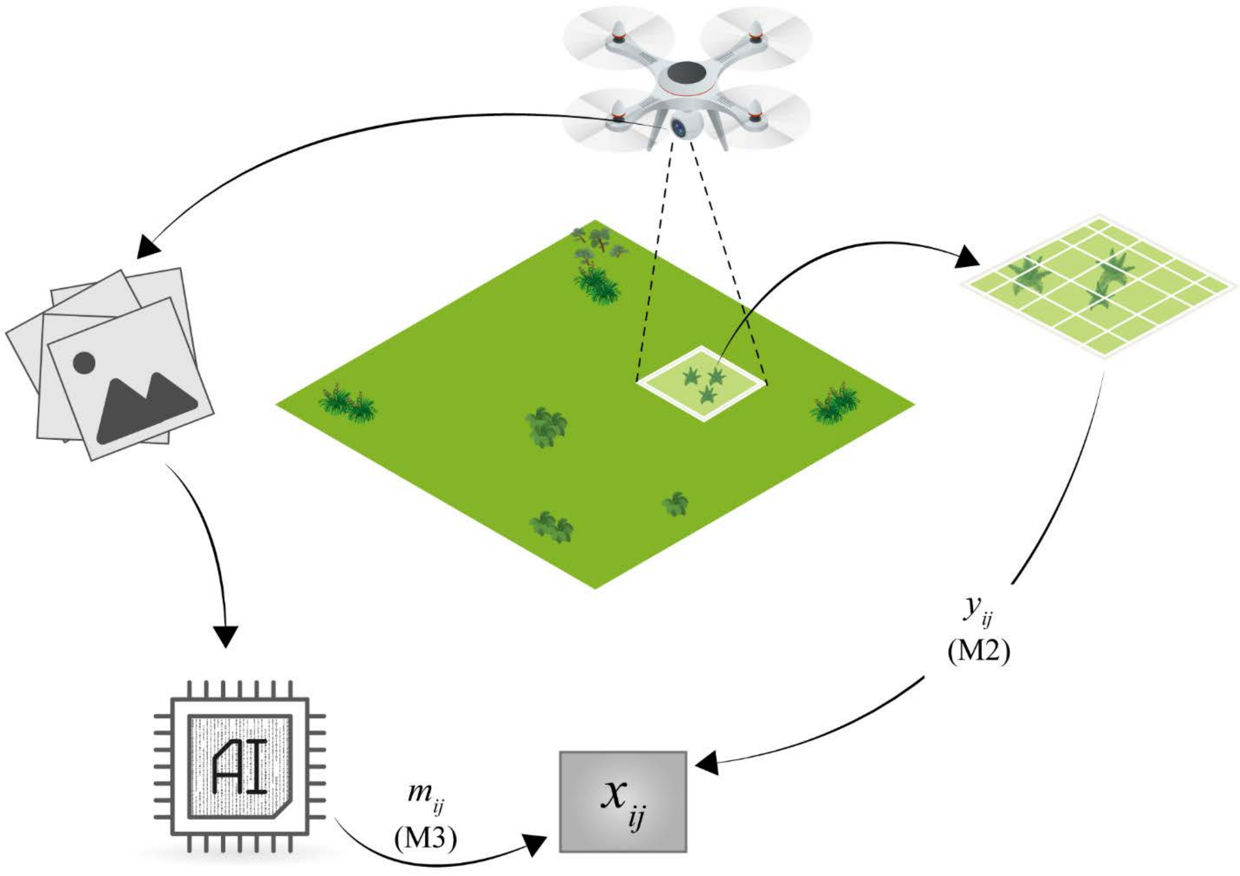
The hierarchical model for determining the uncertainty of the vertical density measured by the drone images is outlined in figure 1. The true, but unknown, vertical density of species i at plot j is denoted x_{ij}. The pinpoint vertical density of species i at plot j observed by the pinpoint method is denoted y_{ij}, and assumed to be distributed according to a generalized Poisson distribution ($M2$) with mean parameter x_{ij} and a species-specific scale parameter ρ_i (Damgaard 2014, Damgaard et al. 2014). The predicted vertical densities from the image data at the level of the plot are denoted m_{ij} and assumed to be distributed according to a reparametrized gamma distribution ($M3$) with mean $x_{ij} + \tau_i \, x_{ij}$, where τ_i is a species-specific bias parameter, and ν_i is a species-specific scale parameter.

$$M2\colon y_{ij} \sim GP\left(x_{ij}, \rho_i\right) \qquad (2),$$

$$M3\colon m_{ij} \sim Gamma(x_{ij} + \tau_i\, x_{ij}, \nu_i) \qquad (3).$$

In order not to confound possible false species identification with the uncertainty in predicting the vertical density, the species identity of the predicted vertical density data in the ground-truth plots is corrected manually before fitting $M3$.

Fig. 1. Outline of the hierarchical model for determining the uncertainty of the vertical density measured by the drone images. The true, but unknown, vertical density of species i at plot j in ground-truthing plot j is modelled by the latent variable x_{ij}. The posterior distribution of the latent variable is calculated using both i) the vertical density predicted from the information from the drone images using machine-learning algorithms (m_{ij}) that are modelled using $M3$, and ii) the vertical density measured by the pinpoint method (y_{ij}) that is modelled using $M2$.



The idea is now to fit the measurement equations $M1$ and $M3$ to the information in the ground-truthing plots and keep these fitted measurement equations *fixed* when fitting plant population ecological models to the image data of the other plots at the site.

3.4 Population ecological modelling using image data

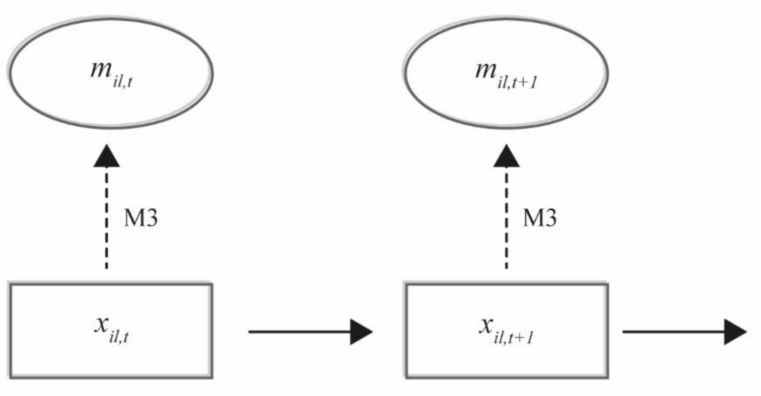
The ultimate aim of the statistical models is to be able to fit plant population ecological models, e.g. competition modes, to time-series image data with a known degree of uncertainty. Following a discrete Lotka-Volterra competition model and earlier population ecological modelling studies, where interspecific interactions are modelled using pin-point abundance data (Damgaard et al. 2009, Damgaard et al. 2011, 2014, Merlin et al. 2015, Ransijn et al. 2015), the following general species interaction modelling framework may be followed:

$$x_{i,t+1} = f_i(x_{i,t}) \sum_j Exp(-c_{ij}\, x_{j,t}) \qquad (4),$$

where f_i is a species-specific growth function in the absence of interspecific interactions and c_{ij} measure the competitive effect of species *j* on the growth of species *i*.

The population ecological model (eqn. 4) may now be applied on a selected "vegetation plot" *l* that has the same size as the ground-truthing plots, but where only image data are available. The model (eqn. 4) is the process equation in a hierarchical model, where the measurement equation of the true, but unknown, vertical density of species *i* in a selected "plot" *l* at time *t*, $x_{il,t}$ is specified by the predicted vertical density $m_{il,t}$ and the fitted $M3$ (Fig.2).

Fig. 2. Hierarchical population ecological model fitted to image data from a selected "vegetation plot" *l* that has the same size as the ground-truthing plots, but where only image data are available. The true, but unknown, vertical density of species i at time *t* is modelled by the latent variable $x_{il,t}$ and the solid arrows are the process equation (eqn. 4). The dashed arrows are the fitted measurement equations ($M3$) that link the vertical density predicted from the information from the drone images (m_{ij}) to the latent variables.



The output of running the selected machine-learning algorithms on the image data of plot *l* is a vector, $\boldsymbol{m}_l$, where each element in the vector contains the predicted species identity and the corresponding predicted vertical density of that species in the plot. However, the species identity is determined with some uncertainty from the image data, and this uncertainty needs to also be included in the uncertainty of the population ecological modeling. This uncertainty is proposed to be included when fitting the model using a numerical MCMC procedure by drawing $\boldsymbol{m}_l^d$ during the model fitting procedure according to $\boldsymbol{m}_l$ and the fitted $M1$. More specifically, for each entity of vertical density in $\boldsymbol{m}_l$, a new species identity is randomly drawn using the fitted $M1$, and the resulting vertical densities are collected by their drawn species identity into the matrix $\boldsymbol{m}_l^d$. The frequency of drawing a new $\boldsymbol{m}_l^d$ may be set to every 100th MCMC iteration, but the sensitivity of this frequency setting to the overall convergence properties of the MCMC must be checked by visual inspection of the sampling chains.

4 Discussion

The chosen statistical distributions $(M1, M2 \text{ and } M3)$ are, in my opinion, natural choices for modelling the statistical uncertainty of the different stochastic processes and, except for $M3$, they have been applied in a number of empirical studies. However, the outlined modelling concept is general, and alternative specifications of the suggested statistical distributions may be relevant in other cases. For example, the bias correction in $M3$ is suggested to be proportional to the vertical density, but if more detailed information on the bias is available, then this information should, of course, be used to specify $M3$.

The reason for choosing vertical density obtained by the pinpoint method as the measure of plant abundance in the ground-truthing plots is threefold. i) the vertical density is a non-destructive method for measuring plant abundance that has been shown to be correlated with plant biomass, ii) the vertical density measure has previously been shown to be usefull for fitting plant population ecological models, and iii) it is possible to aggregate the abundance of single species into the abundance of species groups. However, other measures of plant abundance with similar charactersitica can be used instead, and then the statistical distribution used in M2 should be modified accordingly.

Generally, I find it important that abundance measures allow for the aggregation of abundances across species groups, e.g. counts of individuals, biomass, or vertical density. In species-rich communities, it will not be practically possible, or even desirable, to construct dynamic population ecological models where all species are accounted for individually. Instead, it is important to construct taxonomic or ecologically meaningful species groups that allow the results of population ecological models to be generalized acoss sites. This necessity to group species may be compared to the plant trait-based approach of summarizing the ecological functions of local plant communities by the mean and variances of selected plant traits (e.g. Garnier et al. 2016).

It is critical that the error rate in the species identification process is minimized when the image data are fitted to the population ecological models. In order to meet this requirement, a number of actions can be applied: i) use time-series image data to identify species-specific changes in the image data ii) aggregate species with similar characteristics in the image data into a species group, iii) only select plots with species groups that are clearly distinct in the image data for population ecological modelling. Regarding the later suggestion, note that in the population ecological modelling of plant competitive growth it is not neccesarry to include all plots or a random selection of plots in the fitting process. Instead, it is a valid approch to select plots and model competitive interactions where species of particular interest are locally coexisting (e.g. Damgaard et al. 2014).

The future use of drone-aided image data is predicted to be of immense importance in fitting plant population ecological models to vegetation data at an unprecedented large geographical scale as well as a fine-scaled resolution in time. This will allow us to make quantitative ecological predictions of plant community and ecosystem dynamics, including quantitative assessments of the process or structural uncertainty, which are urgently needed for targeting and prioritizing societal effort in conserving natural habitats.

5 References

Adler, P. B., J. HilleRisLambers, and J. M. Levine. 2009. Weak effect of climate variability on coexistince in a sagebrush steppe community. Ecology **90**:3303-3312.

Damgaard, C. 2011. Measuring competition in plant communities where it is difficult to distinguish individual plants. Computational Ecology and Software **1**:125-137.

Damgaard, C. 2014. Quantitative plant ecology: statistical and ecological modelling of plant abundance. Aarhus University, E-book.

Damgaard, C. 2015. Modelling pin-point cover data of complementary vegetation classes. Ecological Informatics **30**:179-184.

Damgaard, C., T. Riis-Nielsen, and I. K. Schmidt. 2009. Estimating plant competition coefficients and predicting community dynamics from non-destructive pin-point data: a case study with *Calluna vulgaris* and *Deschampsia flexuosa*. Plant Ecology **201**:687–697.

Damgaard, C., B. Strandberg, S. K. Mathiassen, and P. Kudsk. 2011. The combined effect of nitrogen and glyphosate on the competitive growth, survival and establishment of *Festuca ovina* and *Agrostis capillaris*. Agriculture Ecosystems & Environment **142**:374– 381.

Damgaard, C., B. Strandberg, S. K. Mathiassen, and P. Kudsk. 2014. The effect of glyphosate on the growth and competitive effect of perennial grass species in semi-natural grasslands. J Environ Sci Health B **49**:897-908.

Frigyik, B. A., A. Kapila, and M. R. Gupta. 2010. Introduction to the Dirichlet Distribution and Related Processes. University of Washington.

Garnier, E., M. L. Navas, and K. Grigulis. 2016. Plant functional diversity. Organism traits, community structure, and ecosystem properties. Oxford University Press, Oxford, UK.

Jonasson, S. 1983. The point intercept method for non-destructive estimation of biomass. Phytocoenologia **11**:385-388.

Merlin, A., A. Bonis, C. F. Damgaard, and F. Mesléard. 2015. Competition is a strong driving factor in wetlands, peaking during drying out periods. PLOS ONE **10**:e0130152.

Ransijn, J., C. Damgaard, and I. Schmidt. 2015. Do competitive interactions in dry heathlands explain plant abundance patterns and species coexistence? Plant Ecology **216**:199-211.

Sun, Y., Y. Liu, G. Wang, and H. Zhang. 2017. Deep Learning for Plant Identification in Natural Environment. Computational Intelligence and Neuroscience **2017**:6.

Tay, J. Y. L., A. Erfmeier, and J. M. Kalwij. 2018. Reaching new heights: can drones replace current methods to study plant population dynamics? Plant Ecology **219**:1139-1150.