Observed Cases and Hypothetical Controls to Study the Influence of Trees on Understory Vegetation

Sharon Kühlmann-Berenzon(1)
Urban Hjorth(1)(2)
(1) Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Eklandagatan 86, S-412 96 Gothenburg, Sweden
(2) Department of Business Administration,
Computer Science, Economics and Statistics
Örebro University
S-701 82 Örebro, Sweden
sharon@math.chalmers.se

February 18, 2003

Abstract: The relation between the presence of an understory vegetation species and the influence of tree species was studied using measurements gathered in plots distributed all over Finland. The presence of the understory was observed in several quadrats within each plot. The collective influence of the trees was quantified by the influence potential index, based on the size of the individual trees and their spatial distribution. Given the size of the study area, large scale factors, such as climate and latitude, were expected to be significant. To avoid such factors, a conditional logistic model was derived, where the probability of presence was conditioned on another event also subject to large-scale factors. The resulting model may be interpreted as a matched case-control study: the case is the observed set of quadrats in the plot in terms of absence and presence of the vegetation in the individual quadrats; and the controls are permutations of this set. The results from the odds ratio were valid for both conditional and unconditional logistic models. The analysis of *Vaccinium vitis idaea* indicated that the odds of presence decreased with the increasing influence of Scots pine and Norway spruce; for *Calamagrostis arundinacea*, the decrease was due to the influence of Norway spruce.

Keywords: Boreal forest, Conditional logistic regression, Ecological field theory, Influence potential, Matched case-control, Understory-overstory relationship.

1 Introduction

The abundance of many understory vegetation species has decreased in the Finnish forest since the 1950s, as reported by Reinikainen, Mäkipää, Vanha-Majamaa, and Hotanen (2000). As reasons for the decline, they mention forest management and agricultural practices, which have modified site conditions, age class distribution, and tree species composition (see also e.g. Mäkipää and Heikkinen 2002). The understory is defined as the layer of vegetation beneath the canopy of the trees and includes grasses, herbs, dwarf bushes, and mosses. It plays an important role in the forest by giving shelter to animals, protecting and enriching the soils, and providing fodder. Efforts to model the understory vegetation have been limited (McKenzie and Halpern 1999), although such models could help to evaluate and predict the effect of alterations like the ones mentioned before, as well as those caused by natural disturbances and climate change.

It is known that canopy trees affect the understory, but further details and quantification of the effect have not been studied extensively. The influence potential (IP) index was introduced by Kuuluvainen and Pukkala (1989) and Kuuluvainen, Hokkanen, and Järvinen, and Pukkala (1993) to quantify the collective effect of single trees on other plants. The index has also been modified and applied by Saetre (1999) and Økland, Rydgren, and Økland (1999). Throughout these studies, IP was used to model the effect of the trees on the understory vegetation in small and relatively homogeneous boreal stands of Scots pine (*Pinus sylvestris*) (Kuuluvainen and Pukkala 1989; Kuuluvainen et al. 1993), mixed Norway spruce (*Picea abies*) and hairy birch (*Betula pubescens*) (Saetre 1999), and Norway spruce (Økland et al. 1999). Moreover, the understory was grouped into categories (Kuuluvainen and Pukkala 1989, Økland et al. 1999) or described by principal components (Saetre 1999).

Our specific objective was to model the presence of individual understory vegetation species as a function of IP from different tree species. For this study, the observations were gathered in the permanent sample plots of the 1985-86 National Inventory of Finland. The database consists of more than 3000 plots distributed over the entire country, and thus includes information from different tree stand compositions, ecological conditions, and management practices.

As the data was collected in an extended study area, an additional challenge emerged: large-scale factors, such as climate and latitude, interfered with the signal from IP. The large-scale effects were not of interest, but the local-scale process, i.e. the effect of the nearby trees as measured by IP, was of importance. This condition alone made the logistic model unsuitable for the purposes of the study. The search for a model that avoided the undesired large-scale factors led to a conditional logistic regression, which can be transformed into a matched case-control model. In this application, the cases consisted of observed plots, while the controls were hypothetical plots based on permutations of the observed ones. The fitted models provided estimates of the effect that a tree species had on the presence of an understory species, thereby obtaining results regarding the relationship between tree and understory species.

Sections 2 and 3 of the paper describe the data and the IP function applied. We then present the unconditional logistic model and show why it was not appropriate for this study. The conditional logistic model is described in Section 5, and an example is provided. An application to the presence of *Vaccinium vitis idaea* and *Calamagrostis*

arundinacea using IP of Scots pine, Norway spruce, and hairy birch and silver birch (*Betula pendula*) serves as an illustration in Section 6. The last section provides mathematical details on how the conditional logistic model was derived for this problem, and the Annex includes additional the mathematical results.

2 Data

The Finnish Forest Research Institute (METLA) gathered data on trees and understory vegetation from the permanent sample plots (PSP) during the 1985-86 National Forest Inventory. The PSP were established for monitoring purposes and consisted of 3009 circular plots located on forestry land. The plots had a radius of 9.77 m ($A = 300$ m^2) and were distributed in clusters on a grid over Finland ($A = 337\ 000$ km^2). The clusters in Southern Finland consisted of four plots on a north-south transect, with 400 m between plots and 16 km between clusters; in Northern Finland, the clusters were formed by three plots, with 600 m between plots, and 24 km north-south and 32 km east-west between clusters.

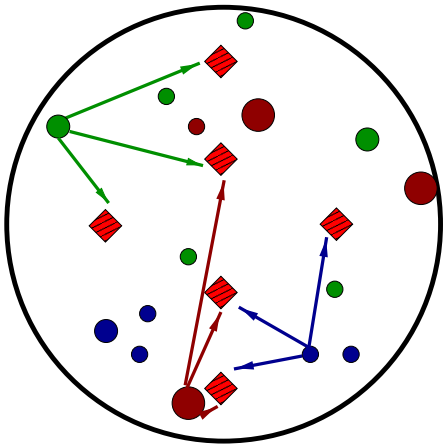

Figure 1: *Permanent sample plot of the 1985-86 National Forest Inventory of Finland: The diamonds represent the quadrats, and the circles, the trees, where the different shades of grey represent species. The quadrats were 2 m^2, and the radius of the plot was 9.77 m. The arrows shows in a schematic way, how every tree influences each quadrat, and each quadrat is influenced by every tree.*

The understory vegetation was assessed visually in six quadrats of 2 m^2, which were located at 3 m and 8 m north and south of the plot center, and at 6 m east and west; see Fig. 1. Not all six quadrats, however, were consistently measured, and therefore the number of observed quadrats in a plot varied between one and six. Information on trees with diameter at breast height (DBH) larger than 10.5 cm was additionally recorded, particular for our study were species, location, and DBH. Other information on the plot included the identification of tree stands, and soil type of the quadrat and of the tree stand.

We believe that the results from this data will be relevant for Finland, as well as for other similar boreal forests in the same latitudes, such as those in Sweden, Norway, and Russia.

3 Influence potential of trees

Ecological field theory (EFT) was originally introduced by Wu, Sharpe, Walker, and Penridge (1985) as a theoretical approach to the study of interactions among individual plants. EFT assumes that a field or domain exists around every plant, where the plant influences the availability of resources according to its own characteristics and other environmental factors. As the plant adds or subtracts resources, it facilitates or suppresses the growth of other plants situated inside the influence field. This general framework allows different individuals from the same community to be compared by: 1. considering specific characteristics of each plant, and 2. incorporating the spatial configuration through the physical domain of the plant. Furthermore the way the plant influences its domain can be described mathematically (Wu et al. 1985; Walker, Sharpe, Penridge, and Wu 1989).

Kuuluvainen and Pukkala (1989) and Kuuluvainen et al. (1993) applied the idea of EFT to a single-tree index that quantified the influence of the trees on other vegetation. The calculation of the influence potential (IP) consists of two steps: the effect of an individual tree on the location is described as a kernel function based on the size of the tree and distance between the tree and the vegetation; then the effects of the trees are aggregated into the total influence potential exerted on the vegetation. Saetre (1999) and Økland et al. (1999) also applied IP using slightly different mathematical formulations from the original.

In this study we used an influence potential function similar to that of Kuuluvainen et al. (1993) and Saetre (1999): influence potential on a quadrat q (IPQ) from trees of species T was defined as

$$\mathrm{IPQ}(q;T) = \sum_{t \in T} D_t \cdot \exp\left(-\frac{\| \mathbf{t} - \mathbf{q} \|^2}{c_T}\right),$$

where D_t is the DBH of tree t; $\mathbf{t}$ is the spatial coordinates of the tree; $\mathbf{q}$ is the spatial coordinates of the quadrat; $\| \mathbf{t} - \mathbf{q} \|$ is the Euclidean distance between the tree and the quadrat; and c_T is a parameter specific for the tree species T and reflects the range of influence of the tree species. IPQ(q,T) sums the effects of all trees of species T, where the effect of an individual tree decreases exponentially with the squared distance between the tree and the quadrat (Fig. 2), and is adjusted according to the tree's DBH. The range of influence of a tree species is defined as $\sqrt{-\log(0.01)\ c_T}$, i.e. the distance at which the unadjusted effect of a tree reaches 0.01 (Fig. 1); a tree beyond that distance is not expected to have a significant influence on the quadrat. Furthermore, the absence of a tree species results in IPQ equal to zero.

Other possible influence functions can be defined (cf. Kuuluvainen et al. 1993; Økland et al. 1999; Saetre 1999). One generalization is to let the parameter c depend on DBH. Although this is biologically reasonable, it also introduces additional complexity in the estimates, and we have therefore opted for the current simpler form.

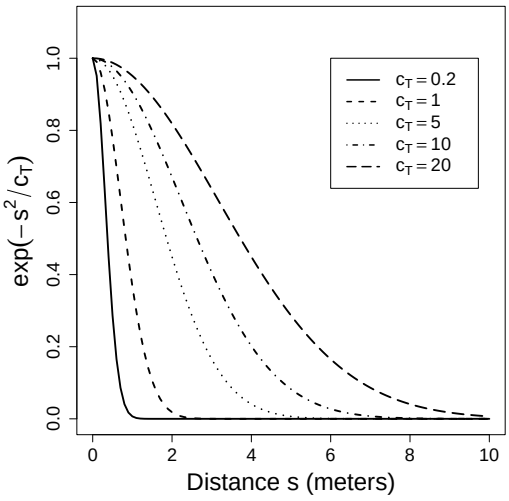

Figure 2: *Influence range: Effect function* $\exp(-s^2/c_T)$ *plotted against the distance* s *for different values of* c_T*, where* s *is the Euclidean distance between the tree and the quadrat. The larger* c_T *is, the larger the influence range of the species, i.e. the distance at which the effect drops to* 0.01*. In IPQ, the effect of every tree is scaled by its DBH.*

4 Logistic regression

4.1 In general

Binary data are observations of an event Y with two possible outcomes, for example an event where success ($Y = 1$) or failure ($Y = 0$) may occur. The presence and absence of an understory species is also an example of data of this type. A model for these situations estimates the probability of the outcomes for each k-th observation, i.e. $P(Y_k = 1) = p_k$ and $P(Y_k = 0) = 1 - p_k$, using information from independent variables or covariates $\mathbf{X}$, where $\mathbf{X}$ represents the design matrix (see e.g. Hosmer and Lemeshow 1989; McCullagh and Nelder 1989; and Collett 1991). In this particular problem, we are interested in modeling the probability of the presence of an understory species, using as covariates the influence potential of Norway spruce, Scots pine, and birch.

One possible model for this type of data is the logistic model (McCullagh and Nelder, 1989)

$$\eta_k = \log\left(\frac{p_k}{1-p_k}\right) = \mathbf{x}_k'\boldsymbol{\beta}; \tag{1}$$

where $\mathbf{x}_k' = [x_{k1}, \ldots, x_{kI}]$ is the k-th row vector of covariates. The inverse transformation gives the success probabilities as

$$p_k = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})}.$$

The assumption behind the logistic model is that the conditional distribution of $Y_k \mid \mathbf{x}_k$ is Bernoulli with expected value p_k.

4.2 For the PSP data

Before setting up a logistic model as Eq. 1, we need to consider additional issues related to the problem. The data collected in the PSP consists of measurements carried out throughout a very extensive study area, where large-scale factors are responsible for much of the variation observed. For example, certain species prefer warmer conditions and are present more often in plots situated in the South of Finland (see e.g. Reinikainen et al. 2000). Therefore large-scale factors must be specified in the model together with the influence potential of the trees. Since the plots are relatively small, however, the large-scale factors affect all the quadrats in the plot in similar way, and thus the large-scale factors are plot-level covariates. Furthermore, some large-scale factors that have been measured, e.g. latitude, can be specified explicitly in the model, but other factors, unknown or unmeasurable, can be included through a plot-level intercept. IPQ, however, is different for each quadrat, since it depends on the distance from the tree to the quadrat.

A logistic model for the presence of an understory vegetation in quadrat q in plot k that includes both the large-scale factors and IPQ could be

$$P(W_{kq} = 1) = \frac{\exp(a_k + \mathbf{g}_k'\boldsymbol{\alpha} + \mathbf{x}_{kq}'\boldsymbol{\beta})}{1 + \exp(a_k + \mathbf{g}_k'\boldsymbol{\alpha} + \mathbf{x}_{kq}'\boldsymbol{\beta})}, \tag{2}$$

where W_{kq} is the presence or absence of the understory plant in quadrat q of plot k; a_k, the plot-level intercept; $\mathbf{g}_k'$, the row vector of explicit large-scale factors for plot k; $\mathbf{x}_{kq}'$, the row vector of IPQ measurements from the different tree species for the quadrat q in plot k; and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the corresponding coefficients.

We are, however, more interested in modeling the relationship between understory and the influence of the trees at the local-scale than on the large-scale factors. As large-scale factors vary more clearly throughout the study area than the local-scale factors, the large-scale factors are expected to dominate the model.

5 Conditional logistic regression

5.1 In general

One way to avoid the large-scale factors is to condition on them. In other words, we calculate probabilities conditioned on another event that is also subject to the same large-scale factors. In this way, the large-scale factors are canceled if the model is appropriate, leaving the local-scale characteristics for further study. Such a model is called conditional logistic regression. Its mathematical structure coinsides with the matched case-control model used in medical applications (Collett 1991; Woodward 1999). Matched case-control is a better description of the way the data is collected in medical studies, while conditional logistic regression describes the technicality of the model and the solution to our problem at hand.

In matched case-control models, strata are formed consisting of cases and control. All subjects in one strata are matched according to characteristics that are not relevant for the study, but that affect the probability of success or failure. The case is the success, and the control is the failure. The probabilities of success are then obtained as functions of covariates of interest, given the matching characteristics.

A matched case-control study in a medical context would compare patients that have a particular disease, i.e. the cases, with patients that do not, i.e. the controls. Both cases and controls are matched in each stratum according to non-relevant characteristics, for example age and sex. The researcher conducting this study will want to find out which covariates affect the presence of the disease and to what degree, while avoiding the effects of age and sex.

In analogous way, we wish to estimate the probabilities of finding an understory vegetation in the forest as a function of IPQ, but conditioning on the large-scale factors. The large-scale factors are known to affect the probabilities of presence and absence, but are not of interest for the study, and for this reason can be considered as matching characteristics.

A conditional logistic model, in general, compares the values of the covariates $\mathbf{x}$ of the cases with those of the controls in the same strata k as in

$$P(\text{case}_\text{k} \mid \text{match}_\text{k}) = \left\{ 1 + \sum_{\text{control}} \exp\left[(\mathbf{x}_\text{k}^{\text{control}} - \mathbf{x}_\text{k}^{\text{case}})' \boldsymbol{\beta} \right] \right\}^{-1}.$$

5.2 For the PSP data

The strata for the PSP data are constructed based on an observed case and several hypothetical controls. Instead of modeling the presence of an understory species in an individual quadrat, we consider the set of all quadrats in the plot,

$$\mathbf{w}_k' = (w_1, \ldots, w_q, \ldots, w_{n_k}), \tag{3}$$

i.e. the set of quadrats $q = 1, \ldots, n_k$ in plot k, where $n_k = 1, \ldots, 6$ is the number of quadrats measured in the plot. Each w_q indicates whether the understory vegetation is present in the quadrat ($w_q = 1$) or absent ($w_q = 0$).

The matching is carried out according to the number of measured quadrats n_k, and the number of quadrats z_k where the understory species was present, i.e. the number of times $w_q = 1$. Given n_k and z_k, we can construct different patterns of $\mathbf{w}_k$. For example, if three quadrats were measured in a plot ($n_k = 3$), and the understory species was found in two of them ($z_k = 2$), then any of the following $\mathbf{w}_k$s are possible: (0, 1, 1), (1, 0, 1), or (1, 1, 0). In other words, the understory species might be present in quadrats 2 and 3, in 1 and 3, or in 1 and 2.

In this way, the stratum consists of all the possible patterns $\mathbf{w}_k$ matched according to n_k and z_k. Only one of them, however, was in fact observed in the field, and this one we designate as the case; the other patterns are hypothetical, and therefore considered to be controls. The number of patterns possible in a plot, given n_k and z_k, is

$$M_k + 1 = \binom{n_k}{z_k}.$$

Furthermore we designate a specific pattern as $\mathbf{w}_k^{(j)}$, with $j = 0$ for the case, and $j = 1, \ldots, M_k$, for the controls.

As mentioned before, the covariate IPQ is different for each quadrat. If the column vector $\mathbf{x}_{ki}' = [x_{k1,i}, \ldots, x_{kn_k,i}]$ represents the i-th covariate for the n_k quadrats of

plot k, a value for that covariate at pattern-level is obtained by evaluating the original measurements according to where the understory species was present:

$$\mathbf{x}'_{ki}\mathbf{w}_k^{(j)} = x_{k1,i}\, w_1^{(j)} + x_{k2,i}\, w_2^{(j)} + \cdots + x_{kn_k,i}\, w_{n_k}^{(j)}. \tag{4}$$

The values of $x_{kq,i}$ from the quadrats where the understory species was present are added, and those where the species was absent are ignored, and so (4) represents the total value of covariate i for the plot k weighted by the j-th pattern. We call this pattern-level IP, influence potential of the pattern (IPP):

$$\mathrm{IPP}_{ki}^{(j)} = \mathbf{x}'_{ki}\,\mathbf{w}_k^{(j)}.$$

With the definitions of what are cases, what are controls, what are the matching variables, and what covariates to use, we can apply the conditional logistic model to the presence of an understory species as

$$P(\mathbf{w}_k^{(0)} \mid z_k, n_k) = \left\{ 1 + \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\,\mathbf{w}_k^{(j)} - \mathbf{X}'_k\,\mathbf{w}_k^{(0)} \right)' \boldsymbol{\beta} \right] \right\}^{-1}. \tag{5}$$

The derivation is explained in detail in section 7. The model compares IPP from the pattern observed in the field ($\mathbf{X}'_k\,\mathbf{w}_k^{(0)}$) to IPP from the hypothetical patterns ($\mathbf{X}'_k\,\mathbf{w}_k^{(j)}$). The complete design matrix of IPP comparisons from all plots, ($\mathbf{X}'\,\mathbf{w}^{(j)} - \mathbf{X}'\,\mathbf{w}^{(0)}$), is of size $(M \times I)$: the number of rows is equal to the total number M of hypothetical patterns in all plots ($M = \sum_k M_k$), and the number of columns corresponds to the number of IPQ variables I. The likelihood (Eq. 9 in Appendix) of the conditional probabilities in (5) is used to obtain the maximum likelihood estimates of $\boldsymbol{\beta}$; in this particular model, the log likelihood of the null model, i.e. when $\boldsymbol{\beta} = \mathbf{0}$, simplifies to a constant equal to $\sum_k \log(M_k + 1)$.

5.3 Example

As an example of how to adjust the original IPQ measurements according to the patterns, we present a plot where quadrats 1, 2, 3, and 4 were recorded; the understory species was observed in quadrats 1 and 3, and the following values of IPQ of Scots pine and Norway spruce were obtained:

Quadrat	Presence	IPQ(pine)	IPQ(spruce)
1	1	2	0
2	0	5	0
3	1	1	0
4	0	8	0

The influence potential of Norway spruce is zero since no trees of that species were found in the plot. Here $n_k = 4$ and $z_k = 2$, and so six patterns are possible. The matrix of original IPQ measurements is the same for all patterns, but after it is multiplied by the patterns $\mathbf{w}_k^{(j)}$, a different IPP is obtained for each as in Table 1. The matrix with the differences between the controls and the case of IPP(pine) and IPP(spruce) becomes the new design matrix for the conditional logistic regression.

Table 1: *Example: IPQ measurements are weighted by the pattern* $\mathbf{w}_k^{(j)}$ *to obtain IPP. Y indicates which pattern is a case (*$Y = 1$*)and which are controls (*$Y = 0$*).*

j	$\mathbf{w}_k^{(j)}$	IPQ(pine) $= \mathbf{x}_{k1}$	IPQ(spruce) $= \mathbf{x}_{k2}$	IPP(pine) $= \mathbf{x}'_{k1}\mathbf{w}_k^{(j)}$	IPP(spruce) $= \mathbf{x}'_{k2}\mathbf{w}_k^{(j)}$	Y
0	(1, 0, 1, 0)			3	0	1
1	(1, 1, 0, 0)	2	0	7	0	0
2	(1, 0, 0, 1)	5	0	10	0	0
3	(0, 1, 1, 0)	1	0	6	0	0
4	(0, 1, 0, 1)	8	0	13	0	0
5	(0, 0, 1, 1)			9	0	0

5.4 Odds ratio

In logistic regression, an important measure is the odds $p/(1-p)$, i.e. the relation between the probability of success and the probability of failure. The odds ratio

$$\Psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

compares the odds of two events with probabilities of success p_0 and p_1 (Woodward 1999).

In a conditional logistic model with all covariates fixed except IPP_i, the odds ratio is calculated as

$$\Psi_i = \exp\left[\Big((\text{IPP}_i + 1) - \text{IPP}_i\Big)\beta_i\right].$$

From this follows that $\exp(\beta_i)$ can be interpreted as the odds ratio when IPP_i increases by one unit and all other covariates are kept constant. In other words, it indicates how much more likely a pattern with $\text{IPP}_i + 1$ is relative to a pattern that has IPP_i. This is often more interesting and useful than estimating the actual probability $P(\mathbf{w}_k^{(j)} \mid z_k, n_k)$, because it gives information on the effect of the covariate on the odds of the studied event.

Furthermore, the odds ratio $\exp(\beta_i)$ also applies to the unconditional probabilities, since the coefficients in the conditional model in (5) are the same as those in the unconditional model in (2) (see section 7 on derivation of the model). This means that the interpretation of the effect of a covariate is valid both in terms of IPP and of IPQ.

6 Application

The conditional logistic regression was applied to the study of the presence of *Vaccinium vitis-idaea L.*, and of *Calamagrostis arundinacea L.*. For the analysis, a homogeneous subset of plots was selected, consisting of those plots with only one tree stand and with all quadrats situated on mineral soils. As covariates, the influence potential of Scots pine, Norway spruce, and birch (hairy and silver birch) were analyzed; these

tree species represented the dominating species in the study area. The IPQ measurements were corrected for edge-effects; see Kühlmann-Berenzon (2002) for details on the method.

For each understory species, only those plots where $0 < z_k < n_k$ were used, since otherwise there is only one possible pattern. The analysis was carried out with the statistical package R v. 1.4.1 (Ihaka and Gentelman 1996), using the package survival v. 2.8-2, which contains a function for fitting conditional logistic model. When using this software, it is not necessary to calculate the covariate matrix explicitly; instead it suffices to provide the IPP and Y columns as in the Table 1, where Y indicates whether the pattern is a case ($Y = 1$) or a control ($Y = 0$). Alternatively, the Newton-Raphson algorithm may be used to find the estimates; the likelihood and the first and second derivatives are provided in the Appendix for this purpose.

Before carrying out the analysis, the likelihood for the model including IPQ of Norway spruce, Scots pine, and birch was numerically maximized in order to obtain the optimal value of the c_T for each species. With these given c_T, further analysis were carried out. In general, the distributions of IPQ were highly negatively skewed, and logarithms did not normalize the observations.

The goodness of fit of the models was tested with the likelihood ratio test at the 5% level of significance. Main effects were first included, and subsequently interactions. Residual plots were also used to check for outliers, in particular Pearson residuals (Hosmer and Lemeshow 1989; Collett 1991), and the delta-beta graphs of Pregibon (1984); the latter show the change in the estimate of a coefficient when a complete stratum is ignored. No particular outliers were observed in any of the cases.

6.1 *Vaccinium vitis idaea*

According to Reinikainen et al. (2000), *V. vitis idaea*, known as cowberry, is one of the most frequent field layer species, and it is most abundant on relatively dry and poor sites with open canopy. Older Scots pine stand are known to be preferred habitat type for this berry.

From the 3009 plots contained in the database, only 211 plots fulfilled the criteria for the analysis, i.e. one tree stand, all quadrats located in mineral soils, and the berry present in at least one quadrat, but not in all the quadrats of the plot. In these plots, 98% of the recorded trees were Norway spruce, Scots pine or birch; some relevant additional statistics are given in Table 2.

Table 2: *Vaccinium vitis-idaea: Mean (st.dev.) number of trees per plot, DBH per tree, and IPP with optimal c_T in observed and hypothetical patterns; and influence range based on optimal c_T.*

	Tree Species					
	Pine		Spruce		Birch	
Number of trees per plot	8.53	(6.26)	6.18	(5.24)	3.08	(2.48)
DBH of tree (cm)	17.33	(5.97)	16.83	(5.59)	14.38	(4.25)
Influence range (m)	1.11		6.28		3.45	
IPP(Observed)	0.448	(1.75)	5.315	(10.37)	1.057	(3.49)
IPP(Hypothetical)	0.743	(2.63)	6.071	(12.38)	1.042	(3.82)

The optimal c_T for Scots pine, Norway spruce, and birch were 0.27, 8.56, and 2.58. Additional 768 hypothetical patterns were possible to construct. The mean and standard deviation of IPP for observed and hypothetical patterns according to tree species is shown in Table 2. Although there was a larger number of Scots pine trees and their mean size was also larger, its mean IPP was relatively low. The reason was the small c_T: at 0.27 it corresponds to a range of influence of 1.11 m, i.e. Scots pine trees located further than 1.11 m away from the quadrat had virtually no influence on the cowberry in the quadrat; for Norway spruce, the corresponding range was 6.28 m, and for birch, 3.45 m.

The most parsimonious model, with a log likelihood -309.15 (log likelihood of null model = -314.68), included the influence potential of Scots pine and Norway spruce. The estimated odds ratio indicate that with every increase of one unit in IPP of Scots pine, the odds of finding cowberry decreased 10.5%; in the case of Norway spruce, the decrease in the odds was of 3.7% (Table 3).

Table 3: *Vaccinium vitis-idaea: Estimates, standard errors, odds ratio, and 95% confidence interval of odds ratio for IPP(pine) and IPP(spruce).*

	Coef.	s.e.(coef)	Odds ratio	95% CI Odds ratio
IPP(Pine)	-0.1107	0.0510	0.895	(0.810; 0.989)
IPP(Spruce)	-0.0375	0.0167	0.963	(0.932; 0.995)

6.2 *Calamagrostis arundinacea*

Calamagrostis arundinacea is a type of reed grass abundant only in southern Finland and known to be frequent on relatively humid site types. This grass is most abundant in young stands, and its abundance decreases by stand age.

For the analysis of *C. arundinacea*, 236 plots were used. The dominating tree species in terms of number of trees per plot was Norway spruce, with an average of more than 10 trees per plot (Table 4). The optimal c_T for Scots pine, Norway spruce, and birch were 0.07, 9.10, and 0.15, and the corresponding influence ranges were 0.57 m, 6.48 m, and 0.83 m. Norway spruce, additional to being the dominating species, also had the widest influence range, which was reflected in the large means of IPP for both observed and hypothetical patterns (Table 4). The design matrix included 236 observed cases and 1003 hypothetical controls.

The best possible model included only IPP of Norway spruce as significant covariate ($\hat{\beta} = -0.0233$, s.e.$(\hat{\beta}) = 0.00691$); the log likelihood of the model was -362.29, and of the null model, -368.23. As the estimated coefficient was negative, this meant that an increase of one unit in IPP (or IPQ) of Norway spruce led to a decrease by 0.977 in the odds ratio, with a 95% confidence interval of $(0.964, 0.990)$.

7 Derivation of the conditional logistic model for PSP

Model (2) is not appropriate for studying the local-scale effects of IPQ on the presence of understory species, since the large-scale effects dominate and are not completely

Table 4: *Calamagrostis arundinacea: Mean (st.dev.) number of trees per plot, DBH per tree, optimal c_T, and IPP with optimal c_T in observed and hypothetical patterns; and influence range based on optimal c_T.*

	Tree Species					
	Pine		Spruce		Birch	
Number of trees per plot	8.44	(7.94)	10.28	(6.83)	3.78	(3.19)
DBH (cm)	19.42	(7.48)	18.79	(6.95)	18.40	(6.95)
Influence range (m)	0.57		6.48		0.83	
IPP(Observed)	0.168	(1.38)	25.112	(34.16)	0.041	(0.52)
IPP(Hypothetical)	0.175	(1.22)	33.051	(39.03)	0.066	(0.68)

known. For simplicity, we will assume that a_k includes all large-scale factors, both those that can be specified and those that cannot, and proceed to derive in detail the conditional logistic model for the problem.

7.1 The modeling object

Since the large-scale factors are nuisance variables, and they affect at a plot level, then the strata for the conditional logistic model must consist of plots. One possibility is to model the set of quadrats in the plot $\mathbf{w}_k$ as in (3), instead of the quadrats individually.

7.2 The conditioning event

The conditioning event must also be subject to the same large-scale factors as $\mathbf{w}_k$. Given that a species is prone to grow in warmer areas, we would expect the number of quadrats where the species was present to be greater in the south than in the north; similar arguments can be used for any other condition. This provides us with an event that is also subject to large-scale factors, namely, the number of quadrats where the species was present. This event, denoted by z_k, can be applied plot-wise, and in fact,

$$z_k = \sum_{q=1}^{n_k} w_{kq}. \tag{6}$$

7.3 The matched stratum

With the modeling object $\mathbf{w}_k$ and the conditioning event z_k, it is possible to construct the different patterns of $\mathbf{w}_k$, where all patterns have the same n_k and z_k; n_k must also be considered, since otherwise z_k is expected to increase or decrease depending on the number of measured quadrats. In the matched case-control terminology, the case is the pattern $\mathbf{w}_k$ observed in the field, while the controls are all other patterns, and together they build one stratum. If the species is present in all or absent in all the quadrats of a plot, then only one pattern is possible, i.e. the observed one, and in that case no valuable information is obtained from that plot, thus such cases are excluded from the analysis.

7.4 The probabilities

Using the logistic model (2), the probability of presence or absence of an understory species in a quadrat, can be written as

$$P(W_{kq} = w) = \frac{\exp\Big[w(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})\Big]}{1 + \exp(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})},$$

where w may be 0 or 1 to indicate absence or presence. By assuming independence among the quadrats, the probability of the pattern $\mathbf{w}_k$ is obtained from the probabilities of the individual quadrats as

$$\begin{aligned} P(\mathbf{w}_k) &= \prod_{q=1}^{n_k} P(w_{kq}) \\ &= \prod_{q=1}^{n_k} \frac{\exp\Big[w_{kq}(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})\Big]}{1 + \exp(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})}. \end{aligned} \tag{7}$$

Furthermore, by Eq. 6, the probability of z_k is based on those patterns $\mathbf{w}_k$ that have n_k and z_k. The set containing those patterns is denoted by

$$A_{z_k} = \left\{ j : \sum_{q=1}^{n_k} w_{kq}^{(j)} = z_k, j = 0, \ldots, M_k \right\}.$$

From Eq.(7) for each pattern, the probability of observing the understory in z_k out of n_k quadrats is calculated as

$$\begin{aligned} P(Z_k = z_k) &= \sum_{j \in A_{z_k}} P\left(\mathbf{w}_k^{(j)}\right) \\ &= \sum_{j \in A_{z_k}} \prod_{q=1}^{n_k} \frac{\exp\Big[w_{kq}^{(j)}(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})\Big]}{1 + \exp(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})}. \end{aligned} \tag{8}$$

7.5 The conditional probability

The conditional probability of $\mathbf{w}_k$ given z_k and n_k is obtained by combining equations Eqs. (7) and (8) in

$$P(\mathbf{w}_k^{(0)} \mid z_k, n_k) = \frac{\prod_{q=1}^{n_k} \frac{\exp\left[w_{kq}^{(0)}\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)\right]}{1+\exp\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)}}{\sum_{j \in A_{z_k}} \prod_{q=1}^{n_k} \frac{\exp\left[w_{kq}^{(j)}\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)\right]}{1+\exp\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)}}$$

This represents the probability of the observed pattern in relation to the probability of all possible patterns with the same number n_k of measured quadrats and the same number of occurrences z_k of the understory species.

With further simplifications (see Appendix), the conditional probability may be expressed as

$$
\begin{aligned}
P(\mathbf{w}_k^{(j)} \mid z_k, n_k) &= \frac{\prod_{q=1}^{n_k} \exp\left(w_{kq}^{(0)}\, \mathbf{x}_{kq}'\boldsymbol{\beta}\right)}{\sum_{j=0}^{M_k} \prod_{q=1}^{n_k} \exp\left(w_{kq}^{(j)}\, \mathbf{x}_{kq}'\boldsymbol{\beta}\right)} \\
&= \left\{ 1 \; + \; \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}_k'\mathbf{w}_k^{(j)} - \mathbf{X}_k'\mathbf{w}_k^{(0)} \right)' \boldsymbol{\beta} \right] \right\}^{-1}.
\end{aligned}
$$

The advantage of this approach is that the large-scale factor a_k affecting the plot is canceled out. This is possible because all patterns in a plot, both observed and unobserved, are subject to the same large-scale effects a_k, and because they all have the same number of 0's and 1's as specified in (6).

8 Discussion

Our objective was to study the presence of an understory vegetation species as a function of the surrounding trees. The effect of the trees was measured using the influence potential on a quadrat IPQ, which took into account both the characteristics of the individual trees as well as the spatial pattern of the stand. Since the data was collected in an extensive study area, the measurements on the presence of the understory species reflected large-scale factors as well as local-scale effects. In order to isolate the local-scale effects measured by IPQ, we derived a model that was conditioned on the large-scale factors and which resulted in a conditional logistic model or matched case-control model. Consequently, the model fulfilled the two criteria required by the problem: it modeled the presence of the understory species, and it concentrated on the local-scale effects, avoiding the large-scale factors.

In this application, the event of interest was not the presence of the understory species in an individual quadrat, but rather the pattern observed in the quadrats of the plot in terms of the presence and absence. Hypothetical patterns were subsequently created by considering all other possible patterns with the same number of quadrats and occurrences. The patterns were also used to adjust the IPQ measurements accordingly: the influence potential of the pattern IPP represented the total sum of IPQ weighted by the pattern. The regression model then compared the hypothetical patterns to the observed pattern in terms of the IPP covariates. The advantage of this approach lies on the fact that specification of large-scale factors is not required, since the factors shared by the cases and the controls are canceled.

Shifting the event of interest from the presence of the understory species on a quadrat to the pattern was not important when the results were interpreted in terms of the odds ratio. The relative change in the odds ratio due to the effect of a tree species was the same in terms of IPQ and IPP, since the coefficients in the conditional and the unconditional logistic models are the same.

The results obtained from analyzing *V. vitis-idaea* showed that an increasing influence of Scots pine and Norway spruce decreased the odds of finding cowberry in the forest. In other words, the greater the presence of Scots pine and Norway spruce, the

less cowberry is likely to be found in the forest. In the case of the grass *C. arundinacea*, an increasing influence of Norway spruce also caused a decrease in the odds ratio. These results coincide with previous ecological knowledge regarding the habitat of these species, i.e. that they require light conditions and poor soils most often related to Scots pine and Norway spruce.

The main assumption of the model presented is that the quadrats within a plot are independent, or independent given the covariates if these are random. This assumption, of course, may be questioned for the current problem. An alternative possibility is to use random effects that would account for the correlation among the quadrats, as done by Chowdhury and McGilchrist (2001).

9 Appendix: Formulas for conditional logistic model

9.1 Simplification and matrix notation

$$
\begin{aligned}
&P(\mathbf{w}_k^{(j)} \mid z_k, n_k) = \\
&= \frac{\prod_{q=1}^{n_k} \exp\left[w_{kq}^{(0)}(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})\right]}{\prod_{q=1}^{n_k} 1 + \exp\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)} \cdot \frac{\prod_{q=1}^{n_k} 1 + \exp\left(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_{j=0}^{M_k} \prod_{q=1}^{n_k} \exp\left[w_{kq}^{(j)}(a_k + \mathbf{x}'_{kq}\boldsymbol{\beta})\right]} \\
&= \frac{\prod_q \exp\left(w_{kq}^{(0)}\ a_k\right) \prod_q \exp\left(w_{kq}^{(0)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_j \left[\prod_q \exp\left(w_{kq}^{(j)} a_k\right) \prod_q \exp\left(w_{kq}^{(j)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)\right]} \\
&= \frac{\exp\left(a_k\ \sum_q w_{kq}^{(0)}\right) \prod_q \exp\left(w_{kq}^{(0)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_j \left[\exp\left(a_k \sum_q w_{kq}^{(j)}\right) \prod_q \exp\left(w_{kq}^{(j)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)\right]} \\
&= \frac{\exp(a_k\ z_k) \prod_q \exp\left(w_{kq}^{(0)}\ \mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_j \left[\exp(a_k\ z_k) \prod_q \exp\left(w_{kq}^{(j)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)\right]} \\
&= \frac{\prod_{q=1}^{n_k} \exp\left(w_{kq}^{(0)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_{j=0}^{M_k} \prod_{q=1}^{n_k} \exp\left(w_{kq}^{(j)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)} \\
&= \frac{\exp\left(\sum_{q=1}^{n_k} w_{kq}^{(0)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)}{\sum_{j=0}^{M_k} \exp\left(\sum_{q=1} w_{kq}^{(j)}\mathbf{x}'_{kq}\boldsymbol{\beta}\right)} \\
&= \frac{\exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]}{\sum_{j=0}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)}\right)'\boldsymbol{\beta}\right]} \\
&= \left\{1\ +\ \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\right\}^{-1}.
\end{aligned}
$$

9.2 Likelihood

Assuming that the plots are independent:

$$
L = \prod_k \left\{1 + \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\right\}^{-1}. \tag{9}
$$

9.3 First derivative

($\mathbf{X}_k$ represents the matrix of covariates and patterns of plot k; $\mathbf{x}_{ki}$ the column vector of covariate i for plot k; $l = \log L$).

$$\frac{\partial l}{\partial \beta_i} = -\sum_k \frac{\sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right] \cdot \left(\mathbf{x}'_{ki}\mathbf{w}_k^{(j)} - \mathbf{x}'_{ki}\mathbf{w}_k^{(0)}\right)}{1 + \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]}$$

9.4 Second derivative

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_i \partial \beta_m} &= \\
&= \sum_k \left\{ \frac{1}{\left\{1 + \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\right\}^2} \right. \\
&\cdot \left\{ \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\left(\mathbf{x}'_{ki}\mathbf{w}_k^{(j)} - \mathbf{x}'_{ki}\mathbf{w}_k^{(0)}\right) \cdot \right. \\
&\cdot \left. \sum_{j=1}^{M_k} \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\left(\mathbf{x}'_{km}\mathbf{w}_k^{(j)} - \mathbf{x}'_{km}\mathbf{w}_k^{(0)}\right) \right\} - \\
&- \left. \sum_{j=1}^{M_k} \frac{\exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]\left(\mathbf{x}'_{ki}\mathbf{w}_k^{(j)} - \mathbf{x}'_{ki}\mathbf{w}_k^{(0)}\right)\left(\mathbf{x}'_{km}\mathbf{w}_k^{(j)} - \mathbf{x}'_{km}\mathbf{w}_k^{(0)}\right)}{1 + \exp\left[\left(\mathbf{X}'_k\mathbf{w}_k^{(j)} - \mathbf{X}'_k\mathbf{w}_k^{(0)}\right)'\boldsymbol{\beta}\right]} \right\}
\end{aligned}$$

References

[1] Breslow, N.E., and Day, N.E. (1984), *Statistical methods in cancer research 1: The analysis of case-control studies,* Lyon: IARC.

[2] Chowdhury, S.R., and McGilchrist, C.A. (2001), "Matched Case Control Studies with Random Exposure Effects," *Biometrical Journal,* 40, 431–446.

[3] Collett, D. (1991), *Modelling Binary Data,* London: Chapman & Hall.

[4] Hosmer, D.W., and Lemeshow, S. (1989), *Applied logistic regression,* New York:John Wiley & Sons.

[5] Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics,* 5, 299–314.

[6] Kühlmann-Berenzon, S. (2002), "Tree Influence on Understory Vegetation: an Edge Correction and a Conditional Model," Tech. Rep. No. 2002:12, School of Mathematical Sciences, Chalmers University of Technology.

[7] Kuuluvainen, T., and Pukkala, T. (1989), "Effect of Scots pine seed trees on the density of ground vegetation and tree seedlings," *Silva Fennica,* 23, 159–167.

[8] Kuuluvainen, T., Hokkanen, T.J., Järvinen, E., and Pukkala, T. (1993), "Factors related to seedling growth in a boreal Scots pine stand: a spatial analysis of a vegetation-soil system," *Canadian Journal of Forest Research,* 23, 2101–2109.

[9] Mäkipää, R., and Heikkinen, J. (2003), "Lanscape-scale changes in the relative abundance of terricolous bryophyte and macrolichen species across Finland from 1951 to 1995," *Journal of Vegetation Science,* to appear.

[10] McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models,* 2 ed., London: Chapman & Hall.

[11] McKenzie D., and Halpern C.B. (1999), "Modeling the distributions of shrub species in Pacific northwest forests," *Forest Ecology and Management,* 114, 293–307.

[12] Pregibon, D. (1984), "Data analytic methods for match case-control studies," *Biometrics,* 40, 639–651.

[13] Økland, R., Rydgren, K., and Økland, T. (1999), "Single-tree influence of understorey vegetation in a Norwegian boreal spruce forest," *OIKOS,* 87, 488–498.

[14] Reinikainen, A., Mäkipää, R., Vanha-Majamaa, I., and Hotanen, J.P. (eds.) (2000), "Summary: Changes in the frequency and abundance of forest and mire plants in Finland since 1950," in *Kasvit Muuttuvassa Metsäluonnossa,* Helsinki: Kustannusosakeyhtiö Tammi.

[15] Saetre, P. (1999), "Spatial patterns of ground vegetation, soil microbial biomass and activity in a mixed spruce-birch stand," *Ecography,* 22, 183–192.

[16] Woodward, M. (1999), *Epidemiology: study design and data analysis,* London: Chapman & Hall.