Mechanical-statistical modelling in ecology: from outbreak detections to pest dynamics

S. Soubeyrand*‡, S. Neuvonen† and A. Penttinen‡

January 17, 2008

Abstract. Knowledge about large-scale and long-term dynamics of (natural) populations is required to assess the efficiency of control strategies, the potential for long-term persistence, and the adaptability to global changes such as habitat fragmentation and global warming. For most natural populations, such as pest populations, large-scale and long-term surveys cannot be carried out at a high resolution. For instance, for population dynamics characterised by irregular abundance explosions, i.e. outbreaks, it is common to report detected outbreaks rather than measuring the population density at every location and time event. Here, we propose a mechanical-statistical model for analysing such outbreak occurrence data and making inference about population dynamics. This spatio-temporal model contains the main mechanisms of the dynamics and describes the observation process. This construction enables us to account for the discrepancy between the phenomenon scale and the sampling scale. We propose the Bayesian method to estimate model parameters, pest densities and hidden factors, i.e. variables involved in the dynamics but not observed. The model was specified and used to learn about the dynamics of the European pine sawfly (*Neodiprion sertifer* Geoffr., an insect causing major defoliation of pines in northern Europe) based on Finnish sawfly data covering the years 1961–1990. In this application, a dynamical Beverton-Holt model including a hidden regime variable was incorporated into the model to deal with large variations in the population densities.

Keywords. European pine sawfly; Bayesian inference; Hidden variable; Scale discrepancy; Spatio-temporal model; Two-regime assumption.

*INRA, UR546 Biostatistics and Spatial Processes, 84914 Avignon, France. Corresponding author: samuel.soubeyrand@avignon.inra.fr

†METLA, PL 68 80101 Joensuu, Finland

‡Department of Mathematics and Statistics, PO Box 35, 40014 University of Jyväskylä, Finland

1 Introduction

Knowledge about the large-scale dynamics of pests such as forest insects is especially needed to determine efficient control strategies and to predict changes in population densities caused by environmental variation like global warming. For producing such knowledge, collecting and analysing spatio-temporal data on pest dynamics over a large region and along several decades is useful, but monitoring densities of pest populations at a high resolution and at a large spatio-temporal scale is often difficult (very expensive) or even impossible. In contrast, binary data indicating occurrences of outbreaks are crude but are more readily available for longer time spans. In Finland, for instance, detected occurrences of pine sawfly (*Neodiprion sertifer*) outbreaks were gathered at the municipality level across three decades (1961-1990). Thus, for each year and each municipality, we know whether or not an outbreak of pine sawfly was detected and we know the values of some covariates which are expected to be related to the pest dynamics.

In this communication, we are interested in the information about pest dynamics which can be gleaned from binary data indicating occurrences of local and annual outbreaks. Such binary data can be analysed directly using regression models to learn why outbreaks occur; see for example Virtanen et al. (1996). Such data can also be used to better understand hidden underlying processes by applying an 'inverse approach' which consists of inferring about hidden processes of the dynamics based on observed patterns (Grimm et al., 2005; Wiegand et al., 2003). Following this approach, we developed a mechanical-statistical model incorporating a model of the pest dynamics and a model of the observation process. The mechanical part of the model contains knowledge about the main mechanisms of the dynamics (e.g. growth, density-dependence and migration), as well as unknowns which must be inferred. The statistical part of the model makes the link between the dynamics and the observations. This hierarchical approach combining a process model and a data model, sometimes called 'physical-statistical modelling' or 'state-space modelling', has been formalised and applied in environmental science (Berliner, 2003; Campbell, 2004; Wikle, 2003a) and ecology (Buckland et al., 2004; Rivot et al., 2004; Wikle, 2003b).

There is often a problem of scale when trying to learn about hidden processes of a large-scale dynamics based on occurrence data. This issue is called 'change of support' (Chilès and Delfiner, 1999; Wikle, 2003b) and arises when the phenomenon scale and the sampling scale do not coincide. Indeed, data on pest outbreaks are often collected at the level of administrative units (low resolution), but the dynamics may show variations at a finer resolution. In the case studied below, for instance, an outbreak detected in a municipality does not occur within the whole municipality but only in some restricted parts of it. The mechanical-statistical model that we propose here was developed as an analysis tool making the phenomenon scale and the sampling scale compatible. This compatibility is made possible by the hierarchical structure of the model in which a data model and a process model are explicitly specified (see Wikle, 2003b; Wikle and Berliner, 2005). For discussions about discrepancies between the scales of phenomenon, sampling and analysis, see Dungan et al. (2002) in ecology and Soubeyrand et al. (2007) in epidemiology.

The spatial resolution of our mechanical-statistical model is the same as the one of the data set, i.e. the unit, but the model construction originates from the dynamics at the subunit level. The model can be briefly described as follows:

- *Statistical part.* The administrative (or observation) units are divided into subunits of equal size, and the outbreak occurrences in the observation units are modelled conditionally on the pest abundances in the subunits. This model is stochastic to account for undetected outbreaks.
- *Mechanical part.* Subsequently, the pest abundances in subunits are modelled as a stochastic process conditional on covariates and past abundances in subunits.

Three difficulties arise in the development of the mechanical part: (i) abundances in subunits are not observed and form a hidden process whose dimensionality (i.e. the number of unknowns) can be much larger than the number of observations; (ii) observed covariates are only measured at the unit level; (iii) there may be spatio-temporal heterogeneity which is not explained by the observed covariates. In order to overcome difficulties (i) and (ii), the following approximation is made: pest abundances in subunits are modelled conditionally on pest densities in the past and covariate values at the unit level (and no longer at the subunit level). The pest density in a unit is defined as the mean of pest abundances in the corresponding subunits. In order to overcome difficulty (iii), additional hidden processes representing unobserved covariates are incorporated into the model.

The resulting model is hierarchical. We use a Bayesian procedure based on Markov chain Monte Carlo (MCMC, Robert and Casella, 1999) to make inferences about pest densities, covariate effects, hidden processes reflecting unobserved covariates, and unknown parameters.

The model and the Bayesian procedure were applied to the dynamics of the European pine sawfly in Finland based on outbreak data collected annually for three decades (1961–1990) at the municipality level. The shapes of the model components were specified to account for particular features of the dynamics. In particular, the forward function reflecting mortality and reproduction processes between two successive years was modelled by a modified Beverton-Holt model fluctuating between two regimes, low and high, which may depend on predation pressure and climatic conditions, for example. The regime variables were not observed and were assumed to form a hidden process.

In sum, the contribution of the proposed approach is threefold. The hierarchical structure of the model enables us to handle the discrepancy between the data scale and the dynamics scale. Making an approximation in the mechanical part of the model allows the dimensionality of the unknowns to be reduced and, consequently, permits reasonable inferences. Incorporating hidden processes into the model enables the investigation of unobserved underlying factors influencing the dynamics. Thus, our modelling approach enables us to infer about the main mechanisms governing the large-scale dynamics (but not those governing the micro-scale dynamics, mainly because of the (necessary) approximation which is made).

In the following, we describe the mechanical-statistical model in a generic context (Section 2) and then present the estimation procedure (Section 3). Next, the model is used to analyse the dynamics of the European pine sawfly in Finland (Section 4) where we focus on (i) the influence of observed covariates on the pest dynamics and (ii) the temporal and spatial dependencies not explained by the observed covariates. This study is completed by a discussion (Section 5).

2 Mechanical-statistical model

2.1 Context and notations

Spatial and temporal resolutions. Here we assume that space and time are discrete. The study region is divided into I spatial units labelled by i, and each unit is divided into J_i subunits of equal sizes. The pairs (i, j) are used to denote the subunits ($i = 1, \ldots, I$ and $j = 1, \ldots, J_i$). In the following, the unit corresponds to the resolution at which data are collected whereas the subunit corresponds to the resolution at which the dynamics is modelled. In the sawfly example, units are municipalities and subunits are pine areas of one hectare.

Let $t = 0, \ldots, T$ index the time; the interval between t and $t + 1$ corresponds to the duration of one life cycle, typically one year.

Dynamics variables. The non-negative variable S_{ijt} denotes the pest abundance in subunit (i, j) at time t. This quantity is assumed to correspond to a fixed stage of the life cycle. Let $\mathbf{S}_t = \{S_{ijt} : i = 1, \ldots, I, j = 1, \ldots, J_i\}$ be the set of pest abundances in all the subunits at time t. The non-negative variable $\bar{S}_{it}$ defined by

$$\bar{S}_{it} = \frac{1}{J_i} \sum_{j=1}^{J_i} S_{ijt}$$

is the pest density in unit i at time t. Let $\bar{\mathbf{S}}_t = \{\bar{S}_{it} : i = 1, \ldots, I\}$ be the set (or map) of pest densities in all the units at time t.

The vector Z_{ijt} is assumed to encode environmental characteristics of subunit (i, j) that influenced mortality and reproduction processes during the time interval $(t - 1, t]$. Let $\mathbf{Z}_t = \{Z_{ijt} : i = 1, \ldots, I, j = 1, \ldots, J_i\}$. We also introduce the aggregated variables

$\bar{Z}_{it}$ whose components are scalar functions (e.g. mean or median) of the corresponding components of Z_{ijt} $(j = 1, \ldots, J_i)$. Let $\bar{\mathbf{Z}}_t = \{\bar{Z}_{it} : i = 1, \ldots, I\}$.

Data variables. The binary variable Y_{it} is equal to one if a pest outbreak was detected at time t in unit i and zero otherwise. The spatial extent of an outbreak may be smaller than the unit area. So, formally, an outbreak is detected in unit i if there is a subunit (i, j) such that the event $S_{ijt} > d$ occurs and is observed, where $d > 0$ is a threshold over which the pest abundance is considered as high. A difference is made between "$S_{ijt} > d$ occurs" and "$S_{ijt} > d$ is detected" because some of the subunits and times such that $S_{ijt} > d$ may be unobserved during the survey.

In addition, some components of $\bar{Z}_{it}$ are observed and are denoted by $\bar{Z}_{it}^{(o)}$ (see Section 3 for the distinction between observed and hidden components).

2.2 Model for the observation process

We first introduce an auxiliary variable: for subunit (i, j) and time t, Y_{ijt} indicates if the event $S_{ijt} > d$ occurred and was observed. By assuming that the intensity and efficiency of the survey are uniform in space and time, the detection variables Y_{ijt} are independently drawn from Bernoulli distributions conditional on pest abundances S_{ijt},

$$Y_{ijt} \mid S_{ijt} \underset{\text{indep.}}{\sim} \text{Bernoulli}\left\{\kappa \mathbf{1}(S_{ijt} > d)\right\}, \tag{1}$$

where $\kappa \in [0, 1]$ is the probability of observing $S_{ijt} > d$ if this event occurs, and $\mathbf{1}(\mathrm{E})$ is the indicator function taking value one if event E occurs and zero otherwise.

The distributions of the observed detection variables Y_{it} at the unit level are then obtained by aggregation of the subunit detection variables. Under the assumption made above and conditionally on pest abundances $\mathbf{S}_t$, the binary variables Y_{it} are independent and drawn from Bernoulli distributions

$$Y_{it} \mid \mathbf{S}_t \underset{\text{indep.}}{\sim} \text{Bernoulli}\left\{P(Y_{it} = 1 \mid S_{i1t}, \ldots, S_{iJ_it})\right\}, \tag{2}$$

with success probabilities

$$P(Y_{it} = 1 \mid S_{i1t}, \ldots, S_{iJ_it}) = 1 - \prod_{j=1}^{J_i} \{1 - \kappa \mathbf{1}(S_{ijt} > d)\}. \tag{3}$$

This expression was obtained as follows: the events $Y_{it} = 1$ and $\sum_{j=1}^{J_i} Y_{ijt} \geq 1$ are identical, hence the success probability is equal to $P\left(\sum_{j=1}^{J_i} Y_{ijt} \geq 1 \middle| S_{i1t}, \ldots, S_{iJ_it}\right)$ which is equal to $1 - \prod_{j=1}^{J_i} P(Y_{ijt} = 0 \mid S_{ijt})$ using equation (1).

2.3 Dynamical model for pest abundances in subunits

As the probabilistic behaviour of the observed detection variables are conditional on pest abundances in subunits, we now build a model for the abundances, that is to say, a model for the spatio-temporal pest dynamics.

The conditional expectation $E(S_{ijt} \mid \mathbf{S}_{t-1}, \mathbf{Z}_t)$, denoted $E_c(S_{ijt})$ for short, of pest abundance in (i,j) at time t given past abundances $\mathbf{S}_{t-1}$ and subunit factors $\mathbf{Z}_t$ is assumed to satisfy

$$E_c(S_{ijt}) = f(S_{ij,t-1}, Z_{ijt})\left(1 - w_{ij\to\mathbb{V}(i,j)}\right) + \sum_{(i',j')\in\mathbb{V}(i,j)} f(S_{i'j',t-1}, Z_{i'j't}) w_{i'j'\to ij}. \quad (4)$$

The terms appearing in this equation, thereafter called space-time dynamic equation, have the following meanings:

- The quantity $f(S_{ij,t-1}, Z_{ijt})$ is the potential pest abundance generated at time t by pests in subunit (i,j) at time $t-1$. The function f reflects mortality and reproduction processes between observation times $t-1$ and t. It is called the forward function. The presence of Z_{ijt} as an argument of f is to show that mortality and reproduction may be influenced by local environmental conditions. Here, f is unspecified for sake of generality, but its shape will be specified in the case-study.
- The weights $w_{ij\to\mathbb{V}(i,j)}$ and $w_{i'j'\to ij}$ reflect population transfers between subunits and successive years (e.g. migrations of pests and predators, spread of (un)favourable conditions). More precisely, $w_{i'j'\to ij}$ reflects the transfer from (i',j') to (i,j) between times $t-1$ and t. The weight $w_{ij\to\mathbb{V}(i,j)} = \sum_{(i',j')\in\mathbb{V}(i,j)} w_{ij\to i'j'}$ reflects the transfer from (i,j) to other subunits; $\mathbb{V}(i,j)$ is the set of all the subunits within all the units, except subunit (i,j). The migration probabilities are assumed to be independent of t.

Equation (4) gives the conditional expected value of S_{ijt}. In order to account for additional local variability, stochasticity is introduced in the dynamical model. Conditional on past abundances $\mathbf{S}_{t-1}$ and subunit factors $\mathbf{Z}_t$, S_{ijt} are assumed to be drawn from the independent gamma distributions with shape parameters $E_c(S_{ijt})^{1-\gamma}$ and scale parameters $E_c(S_{ijt})^{\gamma}$ ($\gamma \in \mathbb{R}$)

$$S_{ijt} \mid \mathbf{S}_{t-1}, \mathbf{Z}_t \underset{\text{indep.}}{\sim} \text{Gamma}\left(E_c(S_{ijt})^{1-\gamma}, E_c(S_{ijt})^{\gamma}\right). \quad (5)$$

The expected value of a variable with gamma distribution being equal to the product of the shape and scale parameters, equation (5) is consistent. The conditional variance of S_{ijt} is the product of the shape parameter and the squared scale parameter, $V(S_{ijt} \mid \mathbf{S}_{t-1}, \mathbf{Z}_t) = E_c(S_{ijt})^{1+\gamma}$. So, parameter γ modulates the dispersion of the probabilistic distribution of S_{ijt} and especially the risk of extreme events such as high abundances which correspond to strong outbreaks.

2.4 Approximation of the space-time dynamic equation

Let us first motivate the approximation. If the migration probabilities $w_{i'j'\to ij}$ and the forward function f were specified up to unknown parameters, then the hierarchical model constructed from distributions (2) and (5) could be directly fitted to data in order to infer about the dynamics. Indeed, pest abundances S_{ijt} could be viewed as random effects forming a hidden process, and a Monte Carlo based method (Robert and Casella, 1999; Wei and Tanner, 1990) could be applied to infer about the process and the unknown parameters. However, in such a hierarchical model, the number of random effects can be very large compared to the number of observations (in the case-study more than 120,000 subunits scattered in the 431 municipalities are considered and there are 30 years of data) and, consequently, the estimation procedure can be unfeasible.

Hence, we adopt the following strategy. The space-time dynamic equation (4) is approximated using auxiliary variables, namely the pest densities $\bar{S}_{it} = (1/J_i)\sum_{j=1}^{J_i} S_{ijt}$ in the units. Under this approximation, the distribution of the binary observation process can be given conditionally on the densities in units. Because these densities are unobserved, they are treated as random effects whose number is the same as the number of observations (431 municipalities $\times$ 30 years in the application). Under this new hierarchical model, a Monte Carlo based inference method can reasonably be applied.

To approximate the space-time dynamic equation, it is assumed that the inter-subunit transfer weights $w_{i'j'\to ij}$ are symmetric and that the forward function f is continuously differentiable. Then, a first-order Taylor's expansion yields an approximation of $E_c(S_{ijt})$ given by

$$\bar{E}_c(S_{ijt}) = f(\bar{S}_{i,t-1}, \bar{Z}_{it}) + \sum_{i'=1}^{I}\{f(\bar{S}_{i',t-1}, \bar{Z}_{i't}) - f(\bar{S}_{i,t-1}, \bar{Z}_{it})\}w_{i\to i'}, \qquad (6)$$

where $w_{i\to i'}$ is the transfer weight from unit i to unit i'. Recall that $\bar{S}_{i,t-1}$ is the pest density in unit i at time $t-1$ and $\bar{Z}_{it}$ encodes mean environmental characteristics of unit i influencing the dynamics between times $t-1$ and t. So, $E_c(S_{ijt})$ is approximated by the sum of a mean effect $f(\bar{S}_{i,t-1}, \bar{Z}_{it})$ and relative fluxes of pests between units. This may be a crude approximation, but it enables us to (i) catch the main temporal and spatial dependencies of the studied dynamics and (ii) link the pest abundances with the observed covariates.

2.5 Expression of the hierarchical model

Based on the approximation made above, it is now possible to build a hierarchical model consisting of (i) a model for the dynamics of pest densities $\bar{S}_{it}$ in units, and (ii) a model for the observed detection variables Y_{it} conditional on the pest dynamics at the unit resolution.

Replacing $E_c(S_{ijt})$ by $\bar{E}_c(S_{ijt})$ implies that the conditional distributions of pest abundances S_{ijt} in subunits are now given by

$$S_{ijt} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t \underset{\text{indep.}}{\sim} \text{Gamma}\left(\bar{E}_c(S_{ijt})^{1-\gamma}, \bar{E}_c(S_{ijt})^{\gamma}\right), \tag{7}$$

and that the conditional distribution of pest densities $\bar{S}_{it} = (1/J_i)\sum_{j=1}^{J_i} S_{ijt}$ is

$$\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t \underset{\text{indep.}}{\sim} \text{Gamma}\left(\bar{E}_c(S_{ijt})^{1-\gamma}, \bar{E}_c(S_{ijt})^{\gamma}\right).$$

Hence, the conditional expected value of S_{it}, denoted by $\bar{E}_c(\bar{S}_{it}) = E(S_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t)$, is equal to $\bar{E}_c(S_{ijt})$. Consequently, the dynamical model for the pest densities in units is

$$\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t \underset{\text{indep.}}{\sim} \text{Gamma}\left(\bar{E}_c(\bar{S}_{it})^{1-\gamma}, \bar{E}_c(\bar{S}_{it})^{\gamma}\right) \tag{8}$$

$$\bar{E}_c(\bar{S}_{it}) \;= f(\bar{S}_{i,t-1}, \bar{Z}_{it}) + \sum_{i'=1}^{I}\{f(\bar{S}_{i',t-1}, \bar{Z}_{i't}) - f(\bar{S}_{i,t-1}, \bar{Z}_{it})\}w_{i\to i'}. \tag{9}$$

From equations (3) and (7), it can be shown that the outbreak-detection variables Y_{it} conditional on the past pest densities $\bar{S}_{t-1}$ in units and mean environmental factors $\bar{Z}_{it}$ are independently drawn from Bernoulli distributions

$$Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t \underset{\text{indep.}}{\sim} \text{Bernoulli}\left\{P(Y_{it} = 1 \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t)\right\} \tag{10}$$

with success probabilities

$$\begin{aligned} P(Y_{it} = 1 \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t) =& 1 - \left\{1 - \kappa P(S_{ijt} > d \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t)\right\}^{J_i} \\ =& 1 - \left[1 - \kappa\left\{1 - F_{\bar{E}_c(\bar{S}_{it})^{\gamma}}\left(\frac{d}{\bar{E}_c(\bar{S}_{it})}\right)\right\}\right]^{J_i}, \end{aligned} \tag{11}$$

where F_x is the cumulative distribution function of the gamma distribution with shape parameter x^{-1} and scale parameter x. Whereas in equation (3) the probability of a detected outbreak in a unit is conditional on pest abundances S_{ijt} in the corresponding subunits, in equation (11) the probability of a detected outbreak is conditional on $\bar{\mathbf{S}}_{t-1}$ and $\bar{\mathbf{Z}}_t$ which characterise the distribution of S_{ijt}. The advantage of the approximation is that it is possible to replace the actual values of pest abundances by characteristics of their probability distribution. Consequently, the pest abundances in subunits no longer appear in the model, and are replaced by pest densities in units.

Equations (8) and (9) model the pest dynamics at the unit resolution, whilst equations (10) and (11) model the observed detection variables conditional on this dynamics. This set of equations defines the mechanical-statistical model which can be used to analyse data on outbreak detections and infer the large-scale pest dynamics.

3 Bayesian formulation of the model and inference method

Consider a situation where the outbreak-detection variables Y_{it}, the unit sizes J_i and some of the environmental factors grouped in the vectors Z_{it} are observed, and the aim is to infer the dynamics of pest densities, the effects of observed covariates, the hidden processes reflecting unobserved covariates, and the unknown parameters. We adopt the Bayesian approach and implement it using an MCMC algorithm using the following notations and assumptions. The environmental factors are denoted separately with respect to their observation status. A mark (o) is used to denote observed factors and a mark (h) is used to denote hidden factors. For example, $\bar{Z}_{it}^{(o)}$ and $\bar{Z}_{it}^{(h)}$ stand for the components of $\bar{Z}_{it}$ which are observed and hidden, respectively ($\bar{Z}_{it} = (\bar{Z}_{it}^{(o)}, \bar{Z}_{it}^{(h)})$). A priori, the hidden factors $\bar{Z}_{it}^{(h)}$ are assumed to be independently drawn from a given parametric model. In addition, the forward function f and the migration probabilities $w_{i \to i'}$ are assumed to have parametric forms. The parameters of f, $w_{i \to i'}$ and $\bar{Z}_{it}^{(h)}$, together with parameter γ and κ, introduced in the previous section, are grouped into the vector θ. Besides, $\bar{\mathbf{S}}$ denotes the set of pest densities from time $t = 0$ to time $t = T - 1$ in all the units; $\mathbf{Y}$, $\bar{\mathbf{Z}}^{(o)}$ and $\bar{\mathbf{Z}}^{(h)}$ denote, respectively, the sets of outbreak-detection variables, observed factors and hidden factors from time $t = 1$ to time $t = T$ in all the units. There is a time lag between the sets $\bar{\mathbf{S}}$ on one hand and $\mathbf{Y}$, $\bar{\mathbf{Z}}^{(o)}$ and $\bar{\mathbf{Z}}^{(h)}$ on the other because of the conditioning on the past which can be seen in (9). The vector $\bar{\mathbf{Z}}$ denotes the union of $\bar{\mathbf{Z}}^{(o)}$ and $\bar{\mathbf{Z}}^{(h)}$.

The posterior joint distribution of the hidden processes $\bar{\mathbf{S}}$, $\bar{\mathbf{Z}}^{(h)}$ and the parameter vector θ is proportional to

$$P(\bar{\mathbf{S}}, \bar{\mathbf{Z}}^{(h)}, \theta \mid \mathbf{Y}, \bar{\mathbf{Z}}^{(o)}) \propto P(\bar{\mathbf{Y}} \mid \bar{\mathbf{S}}, \bar{\mathbf{Z}}, \theta) P(\bar{\mathbf{S}} \mid \bar{\mathbf{Z}}, \theta) P(\bar{\mathbf{Z}}^{(h)} \mid \bar{\mathbf{Z}}^{(o)}, \theta) P(\theta \mid \bar{\mathbf{Z}}^{(o)}).$$

The right-hand side terms of this posterior are constructed using the model assumptions. Thus, the conditional distribution of $\mathbf{Y}$ satisfies

$$P(\mathbf{Y} \mid \bar{\mathbf{S}}, \bar{\mathbf{Z}}, \theta) = \prod_{i=1}^{I} \prod_{t=1}^{T} P(Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t, \theta),$$

where $P(Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t, \theta)$ is the Bernoulli probability distribution (10).

Because of the iterative structure of the dynamical model for pest densities (see equations (8) and (9)), the conditional distribution of $\bar{\mathbf{S}}$ is proportional to

$$P(\bar{\mathbf{S}} \mid \bar{\mathbf{Z}}, \theta) \propto P(\bar{\mathbf{S}}_0) \prod_{i=1}^{I} \prod_{t=1}^{T-1} P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_{\mathbf{t}}, \theta),$$

where $P(\bar{\mathbf{S}}_0)$ denotes the prior of the initial conditions and $P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_{\mathbf{t}}, \theta)$ is the gamma density (8).

As the hidden factors are mutually independent and independent of the observed factors, the conditional distribution of $\bar{\mathbf{Z}}^{(h)}$ simplifies into

$$P(\bar{\mathbf{Z}}^{(h)} \mid \bar{\mathbf{Z}}^{(o)}, \theta) = \prod_{i=1}^{I} \prod_{t=1}^{T} P(\bar{Z}_{it}^{(h)} \mid \theta),$$

where $P(\bar{Z}_{it}^{(h)} \mid \theta)$ stands for the model of the hidden factors mentioned at the beginning of this section.

It is also assumed that the prior distribution of the parameters is independent of the observed factors and simplifies to $P(\theta \mid \bar{\mathbf{Z}}^{(o)}) = P(\theta)$.

The posterior distribution given above can be computed via an MCMC method using a Metropolis-Hastings algorithm for updating the hidden processes and the parameters at each iteration (Robert and Casella, 1999). In order to speed up the algorithm, a block-acceptance strategy based on the decomposition of the posterior distribution can be adopted. Appendix A shows how the algorithm was implemented for the case-study carried out below.

4 Application: large-scale dynamics of pine sawfly

4.1 Ecological context and data

The European pine sawfly (*Neodiprion sertifer*) is a defoliating insect which has four life stages: eggs inserted into pine needles (overwintering stage), larvae feeding on pine needles, cocoons located in the upper layers of soil, and flying adults which oviposit in the autumn. The sawfly population is endemic in Finland but its dynamics is characterised by irregular outbreaks (Juutinen, 1967; Hanski, 1987).

During 1961-1990 in Finland, sawfly outbreaks were recorded at the municipality level by forest owners, forest authorities and the Finnish Forest Research Institute (METLA). An outbreak was detected in a municipality if a pine stand with a strong intensity of sawfly-damage was observed. Such a level of damage occurs when the density of larvae feeding on the pines is clearly higher than in the endemic situation. The spatial extent of an outbreak can vary from only few hectares to thousands of hectares of pine forest (Juutinen and Varama, 1986). Generally, the outbreak areas are rather small, but in some cases, over 25% of the pine forests within a municipality can suffer from outbreaks.

Figure 1 provides graphical descriptions of the outbreak data. Twenty one municipalities in the north and south-west (those with shading lines) were removed from the study to improve homogeneity in the data. Indeed, in the northernmost municipalities the life cycle of sawflies can be longer than in the other municipalities, and in the archipelago in the southwest, the lake ratio which is used as a covariate (see below) is not characteristic because of the presence of the sea. Outbreaks were detected in 29 of the 30 study years and in 52% of the municipalities. The total occurrence rate of detected outbreaks is 5.7%. There are strong spatial and temporal heterogeneity: outbreaks occur most frequently in southern Finland (Fig. 1 a) and there seems to be some degree of synchrony in the outbreaks, their frequency being highest around 1960, 1980 and 1990 (Fig. 1 b). Outbreak periods usually have short durations (Fig. 1 c), whereas the distribution of periods without outbreak is more or less uniform if we do not consider the municipalities without outbreak during the study period (Fig. 1 d).

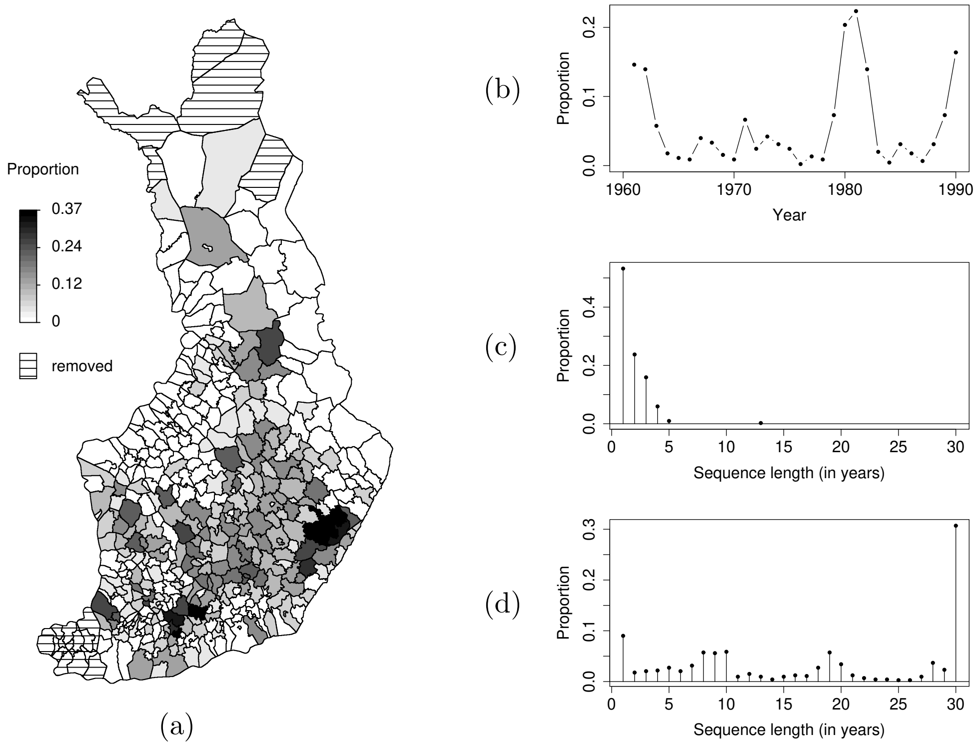

Figure 1: Summary statistics of outbreak data. (a) Spatial variation in the proportion of years with detected outbreaks; municipalities covered by shading lines were removed from the study. (b) Temporal variation in the proportion of municipalities with detected outbreaks. (c) Distribution of durations of outbreak periods at the municipality level (an outbreak period is a set of consecutive years along which outbreaks were detected in a given municipality). (d) Distribution of durations of periods without detected outbreak at the municipality level.

Among the factors which directly or indirectly cause variation in sawfly densities at various scales are winter temperature, predation, virus, parasitoid, soil property, needle quality and acid rain (Hanski, 1987, 1990; Larsson and Tenow, 1984; Larsson et al., 2000; Neuvonen et al., 1990; Saikkonen and Neuvonen, 1993; Saikkonen et al., 1995; Virtanen et al., 1996). Here we analyse the effects of three environmental factors expected to be related to the sawfly dynamics and collected at the municipality resolution.

- *Extreme winter temperature (EWT).* The critical temperature for the death of *N. sertifer* eggs is about $-36°$C (Austarå, 1971; Virtanen et al., 1996). For each municipality and year, we consider as a covariate the occurrence of temperatures below $-36°$C, defined as EWTs. To build this covariate, the minimum winter temperature (MWT) was collected for the study period at 24 meteorolog-

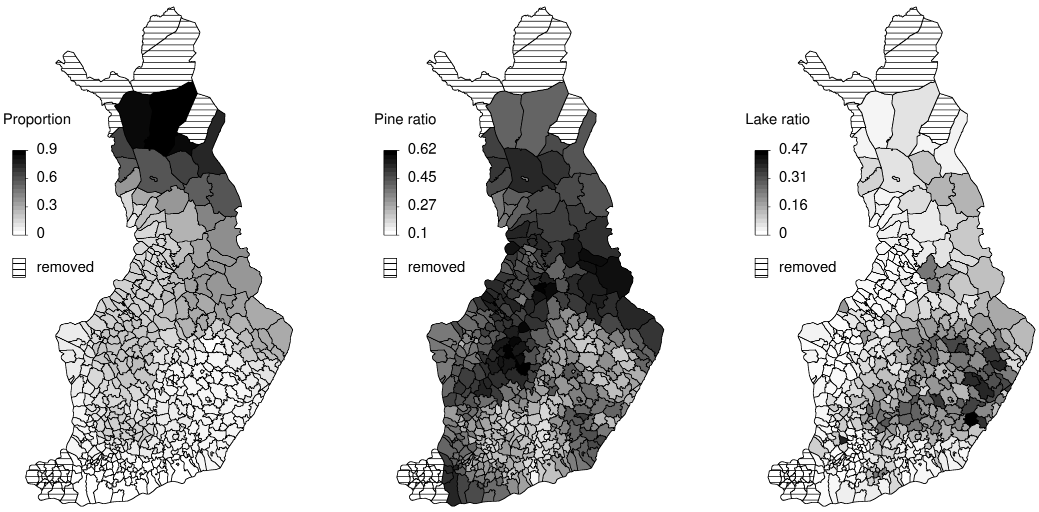

Figure 2: Maps of environmental factors. Left: Proportion of winters with minimum temperatures below $-36°$C. Centre: Ratio of pine area over total area at the municipality level. Right: Ratio of lake area over total area at the municipality level.

ical stations (NORDKLIM, www.smhi.se/hfa_coord/nordklim). Then, the MWT was interpolated at the municipality barycentres using kriging with linear trend (Chilès and Delfiner, 1999). Finally, the occurrence of EWT was simply obtained by thresholding: it is equal to one if the interpolated MWT was below $-36°$C and zero otherwise. Figure 2 (left) shows the spatial variations of the proportion of EWTs.

- *Pine ratio.* The ratio of pine area over total area of a given municipality is assumed to be constant during the study period and is based on the 8th National Forest Inventory which was carried out during 1986–1994 (VM18, wwww.metla.fi/metinfo/vmi; see Figure 2, centre).
- *Lake ratio.* The ratio of lake area over total area of a given municipality is assumed to be constant during the study period and is based on the official statistics of the National Land Survey of Finland (NLSF, www.maanmittauslaitos.fi/default.asp?id=894; see Figure 2, right).

4.2 Model specification and prior distributions

To adapt the model to the sawfly case-study, the mechanical-statistical model was specified as follows:

Space and time. The units are the municipalities and the subunits are areas of one hectare covered by pines. Hence, the number J_i of subunits in municipality i is the number of hectares covered by pine in this municipality. In addition, time $t = 0$ corresponds to year 1960, and times $t = 1, \dots, T$ correspond to the period 1961–1990 for which data are available.

Sawfly densities and outbreak threshold. In this study, the sawfly density S_{it} is assumed to correspond to the density of cocoons which should effectively estimate the density of larvae which feed on pine needles and which cause defoliation. Indeed, most of the feeding occurs at the last larval stage, just before the formation of cocoons (defoliation caused by small larvae which died is neglected but most needle consumption occurs when larvae are large; see Larsson and Tenow, 1979).

As mentioned previously, outbreaks were detected by observing the level of defoliation. Based on available data, defoliation is at the "outbreak level" when the abundance of cocoons in the subunit of one hectare was about 10^6 or more (cf. Hanski, 1987). This number corresponds to the threshold d incorporated in the model of the observation process. So, up to a multiplicative factor of one million, the outbreak threshold is fixed to $d = 1$. Thereafter, the unit of the abundances (resp. densities) is the million (resp. million per hectare).

Transfer weights. The transfer weight $w_{i_1 \to i_2}$ between municipalities i_1 and i_2 is assumed to decrease with distance $\Delta_{i_1 i_2}$ between the municipality centres and to be zero if this distance is greater than a threshold δ. Its parametric shape is given by

$$w_{i_1 \to i_2} = \frac{\exp(-\Delta_{i_1 i_2}/\omega)\mathbf{1}(\Delta_{i_1 i_2} < \delta)}{\sum_{i=1}^{I} \exp(-\Delta_{i_1 i}/\omega)\mathbf{1}(\Delta_{i_1 i} < \delta)},$$

where ω is a positive parameter. In the estimation algorithm, δ was fixed at 20km (using a threshold enables us to keep the computations in the estimation procedure moderate).

Forward function. For the forward function f, we specify a parametric shape which accounts for heterogeneity not explained by the observed factors. With this shape, the system can fluctuate between two regimes: one under which outbreaks are not expected (endemic situation), and the other under which outbreaks are possible (epidemic situation); see Section 5 for a discussion of the two-regimes assumption. Thus, f is assumed to satisfy

$$f(\bar{S}_{i,t-1}, \bar{Z}_{it}) = \frac{\exp(\alpha' \bar{Z}_{it}^{(o)}) \bar{S}_{i,t-1}}{1 + \beta_1^{H_{it}} \beta_0^{1-H_{it}} \exp(\alpha' \bar{Z}_{it}^{(o)}) \bar{S}_{i,t-1}}, \tag{12}$$

where the so-called observed factors $\bar{Z}_{it}^{(o)}$ have four components: the constant value one, the occurrence of EWT, the pine ratio and the lake ratio (see above); α is a vector of unknown real parameters (α' is the transpose of α whose dimension is four); H_{it}

is an unknown binary (0/1) variable modelling the fluctuation between low and high regimes; β_1 and β_0 are positive unknown parameters.

Let us explain equation (12). If migrations are neglected, then the space-time dynamic equation (9) becomes

$$\bar{E}_c(\bar{S}_{it}) \approx f(\bar{S}_{i,t-1}, \bar{Z}_{it}) = \frac{a\bar{S}_{i,t-1}}{1 + b\bar{S}_{i,t-1}},$$

where $a = \exp(\alpha' \bar{Z}_{it}^{(o)})$ and b is equal to $a\beta_1$ if $H_{it} = 1$ and $a\beta_0$ if $H_{it} = 0$. This is a Beverton-Holt model (Geritz and Kisdi, 2004) parametrised by a which represents the annual growth rate if the saturation factors are neglected, and b which accounts for saturation factors. However, our dynamic model is not a simple Beverton-Holt model because parameters a and b can vary in time and space (also because the model accounts for stochasticity and spatial transfers). Parameter a depends on the observed factors $\bar{Z}_{it}^{(o)}$. Parameter b enables the model to fluctuate between two regimes of sawfly density: a low one (outbreaks not expected) and a high one (outbreaks possible). The difference between the two regimes increases with the discrepancy between β_0 and β_1. The value of H_{it} indicates which regime takes place in municipality i between times $t-1$ and t. Hence, we obtain a sort of dynamic Beverton-Holt model whose shape is changed in time and space because of variations in local and annual conditions. The concept of equilibrium which comes with the Beverton-Holt model ($(a-1)/b$ is the equilibrium density) also must be updated: in the classical Beverton-Holt model, the equilibrium is stable; in our model, the stability of the equilibrium $(a-1)/b$ depends on the stability of factors $\bar{Z}_{it}^{(o)}$ and H_{it}. For example, if a is constant and greater than one and if the probability that $H_{it} = 1$ is very low, then b usually equals β_0 and, consequently, $(a-1)/\beta_0$ may be viewed as a quite stable equilibrium, but not $(a-1)/\beta_1$.

Prior distributions. The variables H_{it} are unobserved and correspond to the hidden factors denoted by $\bar{Z}_{it}^{(h)}$ in the Bayesian formulation of the model. We chose independent Bernoulli priors for H_{it} ($i = 1, \ldots, I$, $t = 1, \ldots, T$) with success probability $\eta \in [0, 1]$; we chose an improper uniform prior on $\mathbb{R}$ for logit(η).

Informative priors for β_0 and β_1 were chosen from data on the number of cocoons per hectare at the low and high regimes. From equation (12), the saturation value when $H_{it} = 1$ (high regime) is $1/\beta_1$. Indeed, a/b is the saturation value for a Beverton-Holt model parametrised by a and b. We associate this saturation value with the carrying capacity of the pine forest which is about 10^7 cocoons per hectare, that is to say ten fold the outbreak abundance (10^6 cocoons in one hectare, see above). Thus, $1/\beta_1$ is about tenfold the threshold $d = 1$ (i.e. $\beta_1 \approx 0.1$) and so we chose a Gaussian prior for $\log(\beta_1)$ with mean $\log(0.1)$ and standard deviation 0.1. The logarithm transformation was used because β_1 is positive. When $H_{it} = 0$ (low regime), the saturation value is $1/\beta_0$. No direct information is available on the carrying capacity in the low regime.

However, as the endemic density is about 10^4 cocoons per hectare, we considered that the carrying capacity in low regime is about 5×10^4, that is to say 0.05-fold the outbreak abundance (10^6). It follows that β_0 is approximately equal to 20, and we specified a Gaussian prior for $\log(\beta_0)$ with mean $\log(20)$ and standard deviation 0.1.

Because there was no information on other dynamics parameters and the initial sawfly densities, we chose an improper uniform prior on $\mathbb{R}^{6+I}$ for the four components of α, $\log(\omega)$, γ and $\log(\bar{S}_{i0})$ ($i = 1, \ldots, I$) (we recall that γ modulates the dispersion of the sawfly density distribution (see equation (8)). The logarithm transformation was used for the positive quantities. There was also no information on the probability κ of detecting an outbreak in a subunit if the outbreak actually occurs, except that this probability is small. Specifying a vague prior for this parameter yields identification problems for the other model parameters so we specified an informative prior by assuming that about 20% of the one-hectare outbreaks are detected. We chose a Gaussian prior for $\text{logit}(\kappa)$ with mean $\text{logit}(0.2)$ and standard deviation 0.01 (this choice is discussed in Section 5).

4.3 Results

Output of the MCMC algorithm (see Section 3 and Appendix A) applied to the sawfly data set are presented here.

Example of dynamics within a municipality. Figure 3 illustrates the restoration of the temporal dynamics within a municipality, namely Kauhajoki which contains 649ha of pine forest and is located in the West of Finland. We clearly see two sorts of distributions for the cocoon abundance (Fig. 3 left): some which are concentrated on low values and the others with a greater dispersion and a tail (above the detection threshold one) with a significant mass. Besides, for years with detected outbreaks, the posterior probability of high regime is one; For the other years, this probability fluctuates at lower values (Fig. 3 right).

Effects of observed covariates. We see on Figure 4 (four left panels) that the occurrence of extreme winter temperature (EWT) had a negative effect on the growth rate between successive years, whereas the ratio of lake area had a positive effect. The ratio of pine forest had no significant effect (the value zero is clearly within the posterior distribution).

The unsaturated annual growth rate $\exp(\alpha' Z_{it}^{(o)})$ (see equation (12)), whose posterior distribution is displayed in Figure 4 (right), is a function of the observed factors. The relative influences of the three factors in the unsaturated annual growth rate can be assessed by comparing the following sums of squares. The sum of squares of the linear predictor $\alpha' Z_{it}^{(o)}$ has posterior median and 95% posterior interval 660 [470;990]. The contribution of the occurrence of EWT to this sum has posterior median and 95%

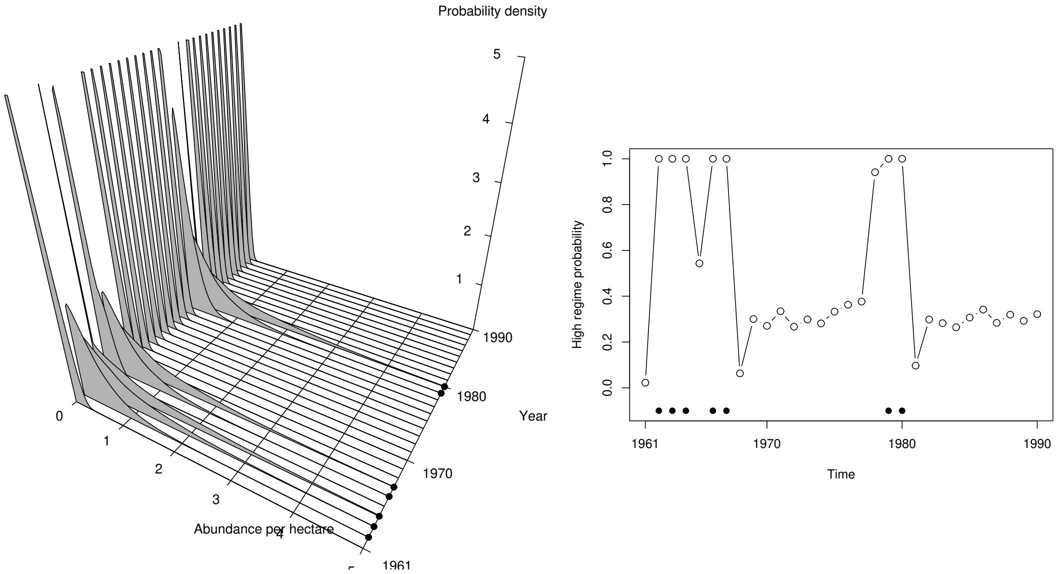

Figure 3: Temporal dynamics within the municipality Kauhajoki. Left: Posterior median of the distribution of cocoon abundances in the subunits of 1ha for each year. Right: Posterior probabilities that the municipality is in high regime ($H_{it} = 1$) for each year. On both plots the black dots indicate the years with detected outbreaks.

posterior interval 140 [40;325]. For the pine ratio and the lake ratio, these quantities are, respectively, 75 [10;185] and 430 [275;685]. Thus, a large part of the variation in the unsaturated annual growth rate is explained by the lake ratio and the occurrence of EWT.

Spatial and temporal dependence. The posterior median of the transfer weight is about half of the weight of no transfer for a municipality at 6km, the minimum distance between municipality barycentres, and is about the tenth of the weight of no transfer for a municipality at 20km (Fig 5, left).

We plotted the 95% posterior envelope of the forward function f (Fig. 5, right) by accounting for the variation in the observed factors, the regime variable and the parameters. As mentioned in Section 4.2, the degree of stability of each equilibrium depends on the stability of the corresponding regime. The posterior median and the 95%-posterior interval of the probability that $H_{it} = H_{i,t+1} = 0$ are 0.45 and [0.40;0.51]. The posterior median and the 95%-posterior interval of the probability that $H_{it} = H_{i,t+1} = 1$ are 0.12 and [0.09;0.15]. Thus, the low regime is quite stable whereas the high regime is rather volatile.

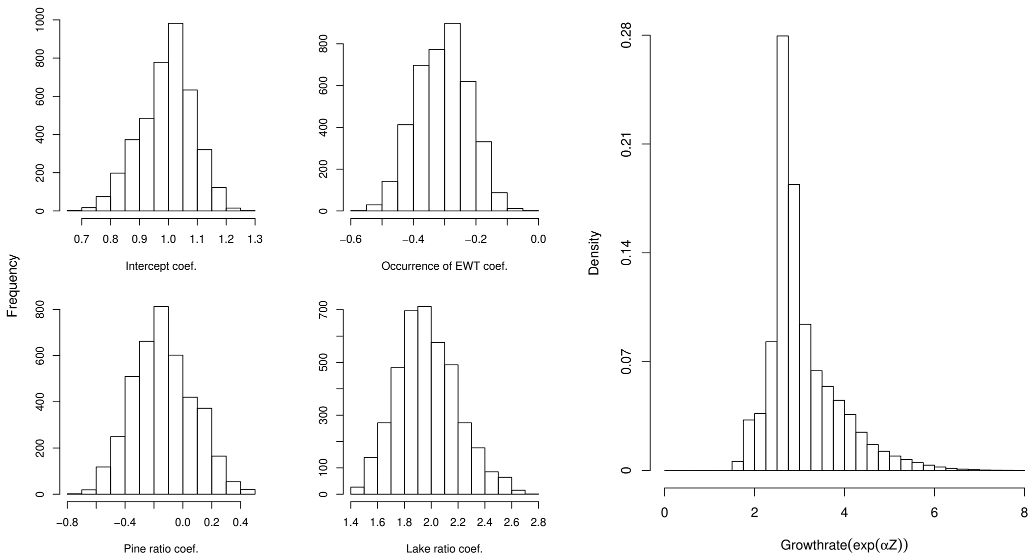

Figure 4: Effects of observed factors. Four left panels: Posterior distributions of the components of α. Right panel: Posterior distribution of the unsaturated annual growth rate $\exp(\alpha' Z_{it}^{(o)})$ (i.e. when the saturation factors are neglected; see equation (12)). This distribution accounts for the variation in the factors $Z_{it}^{(o)}$ and the posterior distributions of the components of α.

5 Discussion

We developed a mechanical-statistical approach to infer large-scale pest dynamics from outbreak occurrence data collected at a crude (administrative) resolution. Our approach can be used to estimate the distributions of pest densities, the effects of observed factors and the role of hidden factors. Furthermore, modelling the observation process enables us to account for missed outbreaks and to handle the discrepancy between the sampling scale and the dynamics scale. An approximation made in the model of the dynamics reduces the dimensionality of unknowns.

Using model output for furthering the study. We illustrated the use of the approach by applying it to data on outbreaks of the European pine sawfly in Finland. Using the approach, we described the spatial and temporal dependence of the dynamics and assessed the effects of covariates. The influence of extreme minimum winter temperatures was revealed earlier by Virtanen et al. (1996) who considered only the spatial variation in temperatures. The negative effect of EWT on outbreaks has a simple mechanistic explanation —the supercooling ability of *N. sertifer* eggs allows them to survive in temperatures as cold as, but not colder than $-36°$C (Austarå, 1971). We used both spatial and temporal variations in our model. Nevertheless, minimum winter temperatures can vary extensively at the landscape scale depending on the local

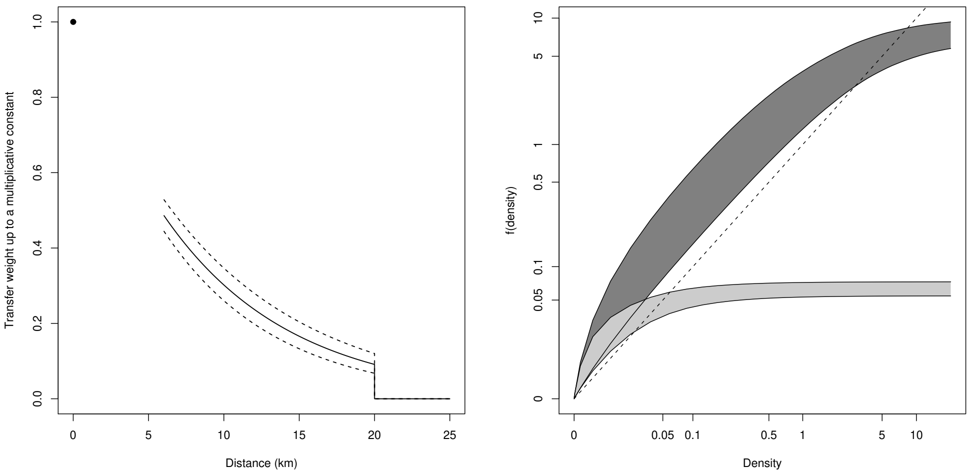

Figure 5: Left: transfer weight (up to a multiplicative constant) against distance; the solid line shows the posterior median and the dashed lines delimit the 95%-posterior envelope. Example for the calculation of the transfer weights: if a given municipality has two neighbour municipalities which are at 7.5km and 14km, respectively, then the transfer weights for these municipalities are approximately 0.4/(1+0.4+0.2)=0.25 and 0.2/(1+0.4+0.2)=0.125. Right: 95%-posterior zone of the forward function f. This zone accounts for the variation in the observed factors and regime variable as well as the posterior distribution of the parameters. The dark grey subzone corresponds to the high regime and the light grey subzone corresponds to the low regime. Note that the scale of the axes is not linear.

topography (Virtanen et al., 1998) and it may be interesting to include topoclimatic variation in future analyses.

In contrast to EWT, the positive effect of the lake ratio on pest density is not yet well understood. Future studies could investigate whether climate, soil conditions and landscape fragmentation associated with lakes play a direct or indirect role in sawfly density. Probably the most parsimonious explanation for the lake ratio effect is that small mammals which are important predators of sawfly cocoons easily become extinct in the barren pine forests near lakes, and the recolonisation of this predator community is slow in landscapes fragmented by lakes (Hanski, 1990).

Further investigations are also required to understand the variation in the estimated transfer weights. These weights are probably too large to be solely explained by pest migration and so the contributions of predator migration and virus spread should be quantified.

The regime variable assigned to each municipality and each year is also a model output which could be analysed to better understand the dynamics. In particular, it may be useful to search for covariates, such as predator density, which are related to this regime variable (note that the occurrence of extreme winter temperature, used in this study, is not significantly correlated with the regime variable).

The two-regimes assumption and beyond. We used a modified Beverton-Holt model which fluctuates between two regimes: one under which the pest density is low and the other under which the pest density can be high. The regime variable determines the saturation level due to unobserved local conditions (e.g. predation pressure). The two-regimes assumption follows the suggestion made by Hanski (1990): the dynamics of *Neodiprion sertifer* should be studied in the context of metapopulations with alternative stable equilibria. However, the two-regime assumption is a simplification and the reality is certainly more like a continuum of regimes. Nevertheless, this assumption is commonly made in order to build tractable models and describe main changes in dynamical systems; see Iglesias and Labarta (2002), van Dijk and Franses (1999) and Wu et al. (2005) who discussed the two-regime (or two-state) approximation in physics, economics and chemistry, respectively.

In our case, identifying covariates linked with the restored regime variable, as proposed above, could help to refine the regime model. There seems to be some consensus among ecologists that the low density equilibrium of pine sawflies is controlled by the predation of cocoons by small mammals (e.g. Hanski, 1987; Larsson et al., 2000). On the other hand, the dynamics in the high densities are more likely controlled by interactions between the sawflies and host foliage and/or pathogens (Dwyer et al., 2004). Hence, in further studies, the Bernoulli distribution used to model the regime variable could be replaced by a (stochastic) function including a model of, or data on, predator dynamics, pathogen dynamics and foliage state.

Using data at various scales. The approximation made in the model consists of replacing the sawfly abundances in subunits by their probability distribution in each unit. Doing this yields a model which is tractable for performing statistical inference. The assumptions related to the probability distribution of abundances were based on qualitative ecological knowledge but to improve model construction and inference accuracy, data at finer scales could be used. For instance, pest abundances measured at the subunit resolution in some limited area (not in the whole study domain since it is not feasible) would help to specify and estimate the probability distribution of pest abundances in subunits. The mechanical-statistical framework that we built could be modified to account for various types of observations (occurrence and abundance) collected at various scales (unit and subunit). Abundance data observed in subunits could be handled by combining a second observation process with the present one. This possibility to use data at various scales is a significant advantage of the mechanical-statistical approach.

Acknowledgements. We thank Martti Varama and Tarmo Virtanen who compiled the data set about outbreak occurrence, Antti Pouttu for useful discussions and Karen Wiebe for constructive comments on the manuscript. This study was supported by the Academy of Finland (project 111156), the Finnish Forest Research Institute and the French National Institute for Agricultural Research.

References

Austarå, Ö. (1971). Cold hardiness in eggs of *Neodiprion sertifer* (geoffroy) (*Hym., Diprionidae*) under natural conditions. *Norsk Entomologisk Tidsskrift 18*, 45–48.

Berliner, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research 108*, 8776.

Buckland, S. T., K. B. Newman, L. Thomas, and N. B. Koesters (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling 171*, 157–175.

Campbell, E. P. (2004). An introduction to physical-statistical modelling using Bayesian methods. Technical Report 49, CSIRO Mathematical & Information Sciences, Australia.

Chilès, J.-P. and P. Delfiner (1999). *Geostatistics. Modeling Spatial Uncertainty.* New York: Wiley.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg (2002). A balanced view of scale in spatial statistical analysis. *Ecography 25*, 626–640.

Dwyer, G., J. Dushoff, and S. H. Yee (2004). The combined effects of pathogens and predators on insect outbreaks. *Nature 430*, 341–345.

Geritz, S. A. H. and E. Kisdi (2004). On the mechanistic underpinning of discrete-time population models with complex dynamics. *Journal of Theoretical Biology 228*, 261–269.

Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H. H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science 310*, 987–991.

Hanski, I. (1987). Pine sawfly dynamics: patterns, processes, problems. *Oikos 50*, 327–335.

Hanski, I. (1990). Small mammal predation and the population dynamics of *Neodiprion sertifer*. In A. D. Watt, S. R. Leather, M. D. Hunter, and N. A. C. Kidd (Eds.), *Population Dynamics of Forest Insects*, pp. 253–264. Andover: Intercept.

Iglesias, Ò. and A. Labarta (2002). Magnetic field scaling of relaxation curves in small particle systems. *Journal of Applied Physics 91*, 4409–4417.

Juutinen, P. (1967). Zur bionomie und zum vorkommen der Roten Kiefernbuschhornblattwespe (*Neodiprion sertifer* Geoffr.) in Finland in den Jahren 1959–65. *Communicationes Instituti Forestalis Fenniae 63*, 1–129.

Juutinen, P. and M. Varama (1986). Ruskean mäntypistiäisen (*Neodiprion sertifer*) esiintyminen Suomessa vuosina 1966–83. *Folia Forestalia 662*, 1–39.

Larsson, S., B. Ekbom, and C. Björkman (2000). Influence of plant quality on pine sawfly population dynamics. *Oikos 89*, 440–450.

Larsson, S. and O. Tenow (1979). Utilization of dry matter and bioelements in larvae of *Neodiprion sertifer* Geoffr. (Hym., Diprionidae) feeding on Scots pine (*Pinus sylvetris* L.). *Oecologica 43*, 157–172.

Larsson, S. and O. Tenow (1984). Aeral distribution of a *Neodiprion sertifer* (Hym., Diprionidae) outbreak on Scots pine as related to stand condition. *Holarctic Ecology 7*, 81–90.

Neuvonen, S., K. Saikkonen, and E. Haukioja (1990). Simulated acid rain reduces the susceptibility of the European pine sawfly (*Neodiprion sertifer*) to its nuclear polyhedrosis virus. *Oecologica 83*, 209–212.

Rivot, E., E. Prévost, E. Parent, and J. L. Baglinière (2004). A Bayesian state-space modelling framework for fitting a salmon stage-structured population dynamic model to multiple time series of field data. *Ecological Modelling 179*, 463–485.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. New York: Springer.

Saikkonen, K. and S. Neuvonen (1993). European pine sawfly and microbial interactions mediated by the host plant. In M. R. Wagner and K. F. Raffa (Eds.), *Sawfly life history adaptions to woody plants*, pp. 431–450. Orlando: Academic Press.

Saikkonen, K., S. Neuvonen, and P. Kainulainen (1995). Oviposition and larval performance of European pine sawfly in relation to irrigation, simulated acid rain and resin acid concentration in Scots pine. *Oikos 74*, 173–282.

Soubeyrand, S., G. Thébaud, and J. Chadœuf (2007). Accounting for biological variability and sampling scale: a multi-scale approach to building epidemic models. *Journal of the Royal Society Interface 4*, 985–997.

van Dijk, D. and P. H. Franses (1999). Modeling multiple regimes in the business cycle. *Macroeconomic Dynamics 3*, 311–340.

Virtanen, T., S. Neuvonen, and A. Nikula (1998). Modelling topoclimatic patterns of egg mortality of *Epirrita autumnata* (Lep., Gemometrida) with Geographical Information System: predictions in current climate and in scenarios with warmer climate. *Journal of Applied Ecology 35*, 311–322.

Virtanen, T., S. Neuvonen, A. Nikula, M. Varama, and P. Niemelä (1996). Climate change and the risks of *Neodiprion sertifer* outbreaks on Scots pine. *Silva Fennica 30*, 169–177.

Wei, G. C. G. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

Wiegand, T., F. Jeltsch, I. Hanski, and V. Grimm (2003). Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *OIKOS 100*, 209–222.

Wikle, C. K. (2003a). Hierarchical models in environmental science. *International Statistical Review 71*, 181–199.

Wikle, C. K. (2003b). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology 84*, 1382–1394.

Wikle, C. K. and L. M. Berliner (2005). Combining information across spatial scales. *Technometrics 47*, 80–91.

Wu, W., D. L. Noble, and A. J. Horsewill (2005). The correspondence between quantum and classical mechanics: an experimental demonstration of the smooth transition between the two regimes. *Chemical Physics Letters 402*, 519–523.

A Implementation of the MCMC algorithm

Here, we show how the MCMC algorithm was implemented for inferring the pine sawfly dynamics. We assume that the reader is familiar with MCMC methods (Robert and Casella, 1999). A block-acceptance strategy based on the decomposition property of the posterior distribution was used to update the unknowns.

From the section giving the Bayesian formulation of the model, the posterior of the hidden processes $\bar{\mathbf{S}}$ and $\bar{\mathbf{Z}}^{(h)} = \{H_{it} : i = 1, \ldots, I, t = 1, \ldots, T\}$ and the parameter vector θ is proportional to

$$P(\bar{\mathbf{S}},\bar{\mathbf{Z}}^{(h)}, \theta \mid \mathbf{Y}, \bar{\mathbf{Z}}^{(o)}) \propto$$
$$P(\theta) \prod_{i=1}^{I} \left(\prod_{t=1}^{T} P(Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t, \theta) P(H_{it} \mid \theta) \right) \left(P(S_{i0}) \prod_{t=1}^{T-1} P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_{\mathbf{t}}, \theta) \right).$$

The term $E_c(\bar{S}_{it})$ does not depend on sawfly densities and factors of municipalities which are at a distance greater than $\delta = 20$km from municipality i because the transfer weight between two municipality i and k is assumed to be zero when the inter-municipality distance Δ_{ik} is greater than δ. So, the distributions of Y_{it} and $\bar{S}_{it}$ can be written as follows:

$$\begin{aligned} P(Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t, \theta) =& P(Y_{it} \mid \bar{\mathbf{S}}_{\partial_i,t-1}, \bar{\mathbf{Z}}_{\partial_i,t}, \theta) \\ P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_{\mathbf{t}}, \theta) =& P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{\partial_i,t-1}, \bar{\mathbf{Z}}_{\partial_i,t}, \theta) \end{aligned}$$

where ∂_i is the set of municipalities such that $\Delta_{i,k} < \delta$ $(k = 1, \ldots, I)$ and $\bar{\mathbf{S}}_{\partial_i,t}$ (resp. $\bar{\mathbf{Z}}_{\partial_i,t}$) is the set of sawfly densities (resp. factors) at time t for municipalities in ∂_i. Thus, when one updates $\bar{S}_{i,t-1}$ or H_{it} (i.e. the hidden component of $\bar{Z}_{i,t}$), only the probabilities $P(Y_{kt} \mid \bar{\mathbf{S}}_{\partial_k,t-1}, \bar{\mathbf{Z}}_{\partial_k,t}, \theta)$ and $P(\bar{S}_{kt} \mid \bar{\mathbf{S}}_{\partial_k,t-1}, \bar{\mathbf{Z}}_{\partial_k,t}, \theta)$ for k in ∂_i may change and, consequently, must be computed.

Based on this remark, we performed the following updating sequence at each iteration of the MCMC algorithm.

1. For each municipality i in $\{1, \ldots, I\}$, update in block the sawfly densities $\{\bar{S}_{i,t-1}, t = 1, \ldots, T\}$ and the hidden factors $\{H_{it}, t = 1, \ldots, T\}$ as follows:

    - Draw the candidate values $\bar{S}^*_{i,t-1}$ and H^*_{it} $(t = 1, \ldots, T)$ from the proposal $\prod_{t=1}^T Q_S(\cdot \mid \bar{S}_{i,t-1}) Q_H(\cdot \mid H_{it})$, where $\bar{S}_{i,t-1}$ and H_{it} are the current values of the sawfly density and the hidden factor, $Q_S(\cdot \mid \bar{S}_{i,t-1})$ is a gamma distribution with shape parameter $\bar{S}_{i,t-1}/0.002$ and scale parameter 0.002, and $Q_H(\cdot \mid H_{it})$ is a Bernoulli distribution with success probability 0.95 if $H_{it} = 1$ and 0.05 if $H_{it} = 0$.

    - Replace the current values by the candidate values with probability

    $$\min\left\{1, \frac{\Lambda_i^* P(\bar{S}^*_{i0})}{\Lambda_i P(\bar{S}_{i0})} \prod_{t=1}^T \frac{P(H^*_{it} \mid \theta) Q_S(\bar{S}_{i,t-1} \mid \bar{S}^*_{i,t-1}) Q_H(H_{it} \mid H^*_{it})}{P(H_{it} \mid \theta) Q_S(\bar{S}^*_{i,t-1} \mid \bar{S}_{i,t-1}) Q_H(H^*_{it} \mid H_{it})}\right\},$$

    where

    $$\Lambda_i = \prod_{k \in \partial_i} \prod_{t=1}^T P(Y_{kt} \mid \bar{\mathbf{S}}_{\partial_k,t-1}, \bar{\mathbf{Z}}_{\partial_k,t}, \theta) \prod_{t=1}^{T-1} P(\bar{S}_{kt} \mid \bar{\mathbf{S}}_{\partial_k,t-1}, \bar{\mathbf{Z}}_{\partial_k,t}, \theta)$$

    and Λ_i^* is equal to Λ_i except that $\bar{S}_{i,t-1}$ and H_{it} $(t = 1, \ldots, T)$ are replaced by $\bar{S}^*_{i,t-1}$ and H^*_{it}.

2. Update in block the parameter vector θ as follows:

    - Draw the candidate subvector θ^* from the proposal $Q(\cdot \mid \theta)$ where θ is the current subvector and $Q(\cdot \mid \theta_1)$ is a multivariate Gaussian distribution with mean vector θ and variance matrix the diagonal matrix whose diagonal elements are 0.01^2.

- Replace the current vector by the candidate vector with probability

$$\min\left\{1, \frac{\Phi^* Q(\theta \mid \theta^*)}{\Phi Q(\theta^* \mid \theta)}\right\},$$

where

$$\Phi = P(\theta)\prod_{i=1}^{I}\left(\prod_{t=1}^{T} P(Y_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_t, \theta) P(H_{it} \mid \theta)\right)\left(\prod_{t=1}^{T-1} P(\bar{S}_{it} \mid \bar{\mathbf{S}}_{t-1}, \bar{\mathbf{Z}}_{\mathbf{t}}, \theta)\right)$$

and Φ^* is equal to Φ except that θ is replaced by θ^*.

Remark: In the algorithm, the parameters were all defined such as their supports were $\mathbb{R}$. Thus θ was defined as $\theta = \{\alpha, \beta, \gamma, \log(\omega), \text{logit}(\kappa), \log(\eta)\}$