THE METROPOLIS-HASTINGS ALGORITHM, A HANDY TOOL FOR THE PRACTICE OF ENVIRONMENTAL MODEL ESTIMATION: ILLUSTRATION WITH BIOCHEMICAL OXYGEN DEMAND DATA

L'ALGORITHME DE METROPOLIS-HASTINGS, UN OUTIL PUISSANT POUR L'ESTIMATION DE MODELES ENVIRONNEMENTAUX. APPLICATION A LA MODÉLISATION DE DONNÉES DE DEMANDE BIOLOGIQUE EN OXYGENE.

TORRE F. [(1)], BOREUX J.-J. [(2)], PARENT E. [(3)]
(1) Department of Statistics, University of Warwick, Gibbett Hill Road, Coventry CV4 7AL, UK, tel. : 44 2476 524 630.
(2) Fondation Universitaire Luxembourgeoise (FUL), 185 avenue de Longwy, 6700 Arlon, Belgique, tel. : 32 63 230 811
(3) Laboratoire de Gestion du Risque En Sciences de l'Eau, ENGREF, 19 av du Maine, 75732, Paris, CEDEX 15, France, tel. : 33 1 45 49 89 30 – fax :33 1 45 49 88 27

Abstract

Environmental scientists often face situations where: (i) stimulus-response relationships are non-linear; (ii) data are rare or imprecise; (iii) facts are uncertain and stimulus-responses relationships are questionable.

In this paper, we focus on the first two points. A powerful and easy-to-use statistical method, the Metropolis-Hastings algorithm, allows the quantification of the uncertainty attached to any model response. This stochastic simulation technique is able to reproduce the statistical joint distribution of the whole parameter set of any model. The Metropolis-Hastings algorithm is described and illustrated on a typical environmental model: the biochemical oxygen demand (BOD). The aim is to provide a helpful guideline for further, and ultimately more complex, models. As a first illustration, the MH-method is also applied to a simple regression example to demonstrate to the practitioner the ability of the algorithm to produce valid results.

Key words: Parameter uncertainty, Markov Chain Monte Carlo sampling, Bayesian inference, Non linear modelling, BOD, Metropolis-Hastings algorithm

Résumé

Les données environnementales ont le plus souvent trois caractéristiques: (I) les relations entre stimuli et réponses sont non linéaires; (ii) les données sont rares ou imprécises; (iii) les faits sont incertains et les relations stimuli-réponse sont mal établies.

Dans cet article, nous nous sommes concentrés sur les deux premiers points. L'algorithme de Metropolis-Hastings (MH) est une méthode efficace et simple à utiliser qui permet de quantifier l'incertitude de la variable de réponse pour une grande variété de modèles. Il s'agit d'une technique de simulation stochastique capable de reproduire la distribution jointe de l'ensemble des paramètres d'un modèle. L'algorithme MH est décrit et illustré à l'aide d'un modèle courant en environnement, celui de la demande biochimique en oxygène (DBO). Le lecteur trouvera une description pratique de l'algorithme sur un cas simple (mais non linéaire) lui permettant d'envisager sereinement une application à des modèles plus complexes. Dans un premier temps, une application à un cas de régression linéaire simple permet de se rendre compte de la conformité des résultats avec ceux obtenus par les méthodes statistiques classiques.

Mots-clés: Incertitude autour de la valeur d'un paramètre, Méthodes de Monte-Carlo par chaînes de Markov, Inférence Bayésienne, Modèles non-linéaires, Demande Biologique en Oxygène (DBO), Algorithme de Metropolis-Hastings.

1. Introduction

Environmental quantitative problems can be categorised according to two main aspects: the knowledge before experimentation (model) and the experimental results (or data). Cases with many data points and a high level of knowledge can be dealt with using classical statistics. Situations with many data points but a low level of knowledge require exploratory data analysis. Cases with few data and a high level of knowledge present important difficulties. These difficulties are further increased when only few facts are known about the phenomenon. The Bayesian approach allows these last two cases to be handled.

In many cases in environmental science, experiments provide only a small amount of data with a large variability. In addition, the stimulus-response relationships are non-linear in most cases (Kuczera (1983); Van Straten and Keesman (1991))

Due to their ease of use, linear models are often broadly applied by scientists, whatever the behaviour expressed by the environmental system. It may be known that a non-linear model would fit better but as specific statistical tools remain relatively undeveloped, linear modelling is still applied. In linear regression, the slope and intercept parameters have well described interpretations and the procedures to obtain the parameter or prediction probability distributions are well known. This is not the case for almost all non-linear models that often lead to intractable mathematical expression or high-level numerical analysis beyond the scope of most environmental science practitioners. In addition, conventional tests for non linear modelling generally rely on asymptotic properties: such tests are consequently appropriate only when large samples of data are available.

To illustrate the use of non-linear modelling, we consider data from Marske and Polkovski (1972) concerning the biochemical oxygen demand (BOD) against time, using a non-linear model based on an exponential decay with a fixed rate constant.

BOD measurement is a useful assessment of the quality of wastewaters and is one of the main criterion for public health national institutes (Kiely (1997)). The BOD of wastewater is the amount of oxygen necessary for micro-organisms to decompose the carbonaceous materials that are subject to microbial decomposition (oxidation of nitrogen compounds is neglected). The BOD value usually reported is the amount of oxygen consumed in milligrams per litre of water or wastewater over a period of five days at 20 °C under laboratory conditions. The BOD test is more completely described in Kiely (1997), p.304). The data from Marske and Polkovski (1972) concern a six-day BOD test on a stream water sample. The data are averaged values of two analyses per day (Tab. 1 and Fig. 1).

{Table 1 & Figure 1}

Assuming the rate of decomposition of organic matter is proportional to the amount of organic matter available, the BOD against time tends to an asymptotic value called the ultimate BOD, and the following non-linear regression model is often used:

$$BOD(t) = BOD_u\left[1 - \exp(-K_{20} \cdot t)\right] + \varepsilon_t, \quad t \in \{1,2,3,4,5,7,9,11\} \qquad \textbf{(1.1)}$$

where $BOD(t)$ is the biochemical oxygen demand at time t, ε_t stands for independent normal errors with constant variance σ^2, BOD_u is the ultimate BOD, K_{20} denotes the constant rate of the organic matter decomposition ($t \in \{1,2,3,4,5,7,9,11\}$), at 20°C.

Parameters of interest for the assessment of water quality are $BOD(5)$, BOD_u and K_{20}, and it may be useful to know the credible intervals for these parameters. For instance, an urban wastewater directive defines that $BOD(5)$ of discharged urban water should not be over 25 mg/L (directive 91/271/EEC). How can the risk of exceeding this limit be evaluated? For a second example, given that the BOD rate coefficient K_{20} should not exceed 0.10 in a river, what is the probability that a studied river is not dramatically polluted? These questions may be answered by the application of the Metropolis-Hastings algorithm (MH-algorithm).

Although simple, this statistical model cannot be solved analytically. It exemplifies the numerical integration problems that non-linear regression models lead to. As a result, researchers have looked for other mathematical tools, e.g. neural networks for which Monte Carlo based estimation techniques also apply (Chen *et al.* (1990)).

New interest in many older algorithms, based on simulation rather than numerical approximations, has arisen from the development and increased accessibility of powerful computer systems. The MH algorithm is a Monte Carlo Markov Chain method, designed to simulate any target probability distribution (Metropolis *et al.* (1953); Hastings (1970)).

Recently, a number of reference books have been written on MCMC-methods (Gamerman (1997); Robert (1996); Gelman *et al.* (1995); Tanner (1992)) and two methods are often described: the Gibbs sampler and the MH-algorithm.

The Gibbs sampler (Geman and Geman (1984), see also Casella and Georges (1992) for a review) is often presented for academic purposes as a special case of the MH-algorithm (Gelman *et al.* (1995), p. 328). It involves sampling from several conditional distributions (equal to the number of parameters in a given model), all of which should be available in an explicit form for stochastic simulation. This, therefore, is a major drawback in the use of the Gibbs algorithm. Normally, neither the posterior joint distribution nor the conditional distributions of the parameters from an environmental model can be derived explicitly, unless by *ad hoc* formulation of model structure and prior density. For an intermediate case, Ritter and Tanner (1992) have proposed a hybrid algorithm between the Gibbs sampler and the MH-algorithm.

On the contrary, the MH-algorithm is entirely general with no such restrictions. As will be described in this paper, this approach is indeed so generally applicable and easy to use that the only limitation to the class of candidate models for a given data set now appears to be the modeller's imagination.

Gelman, Carlin *et al.* (1995) and Brooks (1998) give thorough discussions of Markov chain sampling algorithms. Gilks *et al.* (1996) present many examples of Markov chain applications. Gelman, Carlin *et al.* (1995) illustrate the MH-algorithm applied to the coagulation time of blood drawn from animals randomly allocated to four different diets. Credible intervals were deduced for seven parameters. Tanner (1992) used a simple genetic linkage model and provided a one-parameter distribution simulation.

However, the bibliography about applications of MH-algorithm to model estimation in the domain of environmental sciences is quite poor. An incursion into the domain of rainfall-

runoff modelling can be found in Kuczera and Parent (1998). In this paper, we focus on a case study of physical -especially environmental - model parameters. The BOD example presents all the aspects of variability and uncertainty that may be encountered in the environmental sciences and that can be easily quantified by MH techniques.

The MH-algorithm is particularly well suited for parameter estimation in a Bayesian approach (DeFinetti (1937); Lindley (1972)). It can simulate almost any joint posterior distribution of model parameters, when the likelihood function of the vector of parameters, a prior distribution and the data set are known. The algorithm explores the parameter definition domain, using a specific exploration strategy. At each step, it accepts or rejects the proposed values, according to a specific rule. The algorithm appears to be a homogeneous and positive Markov chain that converges under mild conditions to the desired limiting posterior distribution, namely the posterior distribution of the parameter set.

2. Presentation of the Objective

2.1. Components of Any Model and Notation

A model is in fact a three stage intellectual construction:

1. hypotheses on the relationships between stimuli and response described in terms of mathematical equations,
2. ideas about the model parameter values that a skilled practitioner or an experienced researcher may already have in mind before any data is obtained,
3. error term. The last term represents non reducible uncertainties due to measurement errors, etc.

The first stage includes reducible uncertainties that influence the unobservable variables – i.e. parameters – introduced by the hypotheses. The art of modelling is the choice of realistic hypotheses. Note that, in the first part, hypotheses reduce uncertainty as far as possible, by enforcing a model structure that is kept throughout the process of modelling.

Let us denote $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_k)$ as the (multi-dimensional) parameter of any model where θ_1, θ_2 and θ_k are the components of the parameter. The components are real numbers. For example, in the BOD model, the parameter is a vector of $k = 3$ elements, i.e. $\boldsymbol{\theta} = (BOD_u, K_{20}, \sigma)$. We ask the readers, particularly those unfamiliar with statistical formulations, to keep in mind the analogy between the theoretical explanation and the meaning in terms of the BOD model.

2.2. Bayesian vs. Classical Approach

The debate between classical and Bayesian approach is well illustrated by the way in which the two approaches interpret the nature of parameters.

In classical statistics, parameters are regarded as fixed and deterministic, but unknown. A 95 percent confidence interval means that taking n samples would give n confidence intervals and that 95 percent of these would contain the true value of the parameter.

According to the Bayesian point of view, parameters are uncertain and do not have any fixed "real" and "true" value. Possible parameter values may, of course, be suggested, and conditional lines of reasoning can be explored with the help of models. In the Bayesian setting, probability distributions are the mathematical tool used to encode both the uncertainty associated with non-observable quantities, and to express the natural variability of the observable variables. In this context, credible intervals refer only to the data at hand and

do not involve any imaginary infinite repetition of sample drawing. A 95 percent credible interval simply means that the probability that the parameter belongs to the credible interval is 95 percent, i.e. the odds of betting that the parameter may take a value within the interval are 95 to 5.

We will adhere to the Bayesian paradigm in this paper. Firstly, the classical definition may be confusing: in evidence, many scientists and several handbooks do provide a definition of the classical confidence interval close to the Bayesian credible one (Lecoutre (1997)). Secondly, it appears strange to require n samples to be considered, when only one is known or even exists.

2.3. Prior and Posterior Distribution of a Parameter

In the Bayesian approach, the uncertainty associated with a parameter $\boldsymbol{\theta}$ is represented by a probability distribution. The uncertainty always refers to a state of knowledge, which should be mentioned by a conditioning argument in the probability distribution of the parameter. We will denote by H the modeller's background state of information, which encompasses all hypotheses and existing knowledge before collecting data. The parameter probability density function $P(\boldsymbol{\theta}|H)$ associates a degree of belief to each possible value of the parameter $\boldsymbol{\theta}$. As with any probability density function, this function is positive and the hyper-volume under the curve equals one. The marginal probabilities are the well-known univariate distributions of each component of the vector $\boldsymbol{\theta}$.

The prior distribution $P(\boldsymbol{\theta}|H)$ of $\boldsymbol{\theta}=(\theta_1, \theta_2 \ldots \theta_n)$ takes into account our physical interpretation of each parameter before gathering any experimental data. This prior knowledge may be encoded probabilistically by various methods: Berger (1985) describes 10 various ways to elicit priors. He discerns location parameters from scale parameters and explains how to give them non-informative priors. Vague knowledge about location

parameters should be associated with a uniform prior distribution, but with a given scale parameter σ, it should lead to a σ^{-1}-shape prior distribution. A non-informative prior distribution gives no preference for any vector in the parameter definition domain. However, it can be noted that a non-informative prior gives information on the parameter limit values (fig. 2a).

{Figure 2}

The posterior distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_n)$ describes our probabilistic judgement of plausible values for $\boldsymbol{\theta}$, taking into account the experiment (the data set y) and the prior knowledge (likelihood of the model $P(\mathbf{y}|\boldsymbol{\theta}, H)$ and parameter prior distribution $P(\boldsymbol{\theta}|\mathbf{y}, H)$). It is therefore written as $P(\boldsymbol{\theta}|\mathbf{y}, H)$. Bayes (1763) formula works as an information processor that updates the prior density function $P(\boldsymbol{\theta}|H)$ into the posterior $P(\boldsymbol{\theta}|\mathbf{y}, H)$:

$$P(\boldsymbol{\theta}|\mathbf{y}, H) = \frac{P(\mathbf{y}|\boldsymbol{\theta}, H) \cdot P(\boldsymbol{\theta}|H)}{\int_{\theta} P(\mathbf{y}|\boldsymbol{\theta}, H) \cdot P(\boldsymbol{\theta}|H) \cdot d\boldsymbol{\theta}} \qquad \textbf{(2.1)}$$

Bayesian statisticians interpret this theorem as a relevant mechanism to provide a rational solution of how to learn from the data $\mathbf{y}$ about the quantity of interest $\boldsymbol{\theta}$.

Under mild technical conditions, the posterior distribution function is generally ”sharper” than the prior one (Gelman, Carlin *et al.* (1995)), which is intuitively appealing since one expects the posterior probabilistic statement of belief about $\boldsymbol{\theta}$ to be more “precise than” the prior one (fig. 2b).

The predictive posterior distribution reads:

$$P(\mathbf{z}|\mathbf{y}, H) = \int_{\theta} P(\mathbf{z}|\boldsymbol{\theta}, H) \cdot P(\boldsymbol{\theta}|\mathbf{y}, H) \cdot d\boldsymbol{\theta} \qquad \textbf{(2.2)}$$

This reflects how the chances of obtaining further data $\mathbf{z}$ from the same phenomenon can vary with reference to the initial state of information H and the collected data $\mathbf{y}$, once

possible variations of the parameter $\boldsymbol{\theta}$ have been "integrated". An example of the use of predictive distributions for validating a model of daily precipitations can be found in Chaouche and Parent (1999).

2.4. Objective

In the Bayesian setting, the only problem is the evaluation of the posterior joint distribution of the parameter $(\theta_1, \theta_2, ...\theta_d)$, once experiment results are given and prior knowledge is assessed (fig. 3). This can be theoretically achieved by evaluating Bayes formula . In very specific cases, the so-called conjugate situation, prior and posterior functions belong to the same family of functions (see for instance, Robert (1992) for a general theory of conjugation for the exponential family, or Box and Cox (1973) for practical application to normal models). In such cases, the prior is simply revised into the posterior by updating the coefficients characterising the conjugate class of probability distributions. Unfortunately, apart from these mathematically convenient situations, solving equation (2.2) is problematic due to the evaluation of the integral in the denominator (remember that $\boldsymbol{\theta}$ can be multidimensional) and the integration required for the predictive density function may be numerically infeasible. For a long time, the evaluation of integrals has limited Bayesian inference to rather unrealistic ad hoc student book examples.

{Figure 3}

These computational difficulties may now be overcome using the Metropolis-Hastings algorithm. In other words, Bayesian inference in practice – i.e. getting the parameter posterior distribution – can be performed without difficulty for any non-linear model and any prior distribution.

3. Description of the Metropolis-Hastings Algorithm

Any student in statistics knows that a positive homogeneous Markov chain converges to a limiting probability distribution. In practice, this means that if one looks at the values generated by a given Markov chain sufficiently far from the simulation origin, the successively generated values will be distributed with stable frequencies stemming from a fixed probability distribution. Such a series will behave as a pseudo sample of this limiting probability distribution and any statistical quantity of interest such as the density function and the various moments can be evaluated from the generated pseudo sample.

The MH-algorithm solves the inverse problem: i.e. how to design a Markov chain that converges to a given probability distribution, for instance $P(\cdot|\mathbf{y},H)$. The idea is to generate pseudo-samples of size *N*, with *N* large, $(\boldsymbol{\theta}^{(0)},\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)},...,\boldsymbol{\theta}^{(i)},...,\boldsymbol{\theta}^{(N)})$ with a Markov chain converging to the limiting distribution $P(\boldsymbol{\theta}|\mathbf{y},H)$. Note that, for practical purposes, a Markov chain can be viewed as nothing more than a computer program (with a random function call) applied in a loop that gives $\boldsymbol{\theta}^{(i)}$ as output for the i^{th} iteration, and uses this as the only input for iteration i+1 to produce $\boldsymbol{\theta}^{(i+1)}$ and so on. This chain is homogeneous if the program in the loop does not change as the iterations proceed. The chain is positive if any value belonging to the domain of variation of $\boldsymbol{\theta}$ can always be reached at random in the next loops, whatever the starting values may be.

We describe now the definition of a jump from a given value $\boldsymbol{\theta}^{(i)}$ to a candidate parameter value $\boldsymbol{\psi}$ and the rules leading to the acceptance or the rejection of $\boldsymbol{\psi}$.

3.1. MH jump specification

Let us suppose that the algorithm has just generated $\boldsymbol{\theta}^{(i)}$ after iteration i. At iteration stage i+1, a candidate parameter value $\boldsymbol{\psi}$ is sampled from a fixed symmetric multivariate

jump probability distribution $J\left(\boldsymbol{\psi}\middle|\boldsymbol{\theta}^{(\mathbf{i})}\right)$. The jump distribution $J\left(\cdot\middle|\boldsymbol{\theta}^{(\mathbf{i})}\right)$ may depend on $\boldsymbol{\theta}^{(\mathbf{i})}$ and is used to explore the surroundings of $\boldsymbol{\theta}^{(\mathbf{i})}$. Many versions of the algorithm have been developed using various jump distributions, as will be discussed below. In general, it is required that $J(.|.)$ be symmetric, i.e. $J\left(\boldsymbol{\psi}\middle|\boldsymbol{\theta}^{(\mathbf{i})}\right)=J\left(\boldsymbol{\theta}^{(\mathbf{i})}\middle|\boldsymbol{\psi}\right)$ in order that the transition $\boldsymbol{\theta}^{(\mathbf{i})}\rightarrow\boldsymbol{\psi}$ shall be given the same probability as the reverse transition $\boldsymbol{\psi}\rightarrow\boldsymbol{\theta}^{(\mathbf{i})}$.

3.2. Acceptance-Rejection rule

The candidate value $\boldsymbol{\psi}$ for the next term is or is not added to the chain according to the following acceptance-rejection rule:

- Generate a [0,1]-uniform random value *r*
- Compute the ratio *q* as:

$$q\left(\boldsymbol{\theta}^{(\mathbf{i})},\boldsymbol{\psi}\right)=\frac{P\left(\boldsymbol{\psi}\middle|\mathbf{y},H\right)}{P\left(\boldsymbol{\theta}^{(\mathbf{i})}\middle|\mathbf{y},H\right)}$$

  Note that, since the denominator of Bayes formula giving $P\left(\boldsymbol{\psi}\middle|\mathbf{y},H\right)$ and $P\left(\boldsymbol{\theta}^{(\mathbf{i})}\middle|\mathbf{y},H\right)$ is fortunately the same normalising constant, it does not need to be computed, since $q\left(\boldsymbol{\theta}^{(\mathbf{i})},\boldsymbol{\psi}\right)$ can be directly evaluated as

$$q\left(\boldsymbol{\theta}^{(\mathbf{i})},\boldsymbol{\psi}\right)=\frac{P\left(\mathbf{y}\middle|\boldsymbol{\psi},H\right)\cdot P\left(\boldsymbol{\psi}\middle|H\right)}{P\left(\mathbf{y}\middle|\boldsymbol{\theta}^{(\mathbf{i})},H\right)\cdot P\left(\boldsymbol{\theta}^{(\mathbf{i})}\middle|H\right)}$$

- If $q\left(\boldsymbol{\theta}^{(\mathbf{i})},\boldsymbol{\psi}\right)\geq r$ then jump to $\boldsymbol{\psi}$, that is set $\boldsymbol{\theta}^{(\mathbf{i+1})}=\boldsymbol{\psi}$. However if $q\left(\boldsymbol{\theta}^{(\mathbf{i})},\boldsymbol{\psi}\right)<r$, then remain at the former position, which means that the algorithm output is $\boldsymbol{\theta}^{(\mathbf{i+1})}=\boldsymbol{\theta}^{(\mathbf{i})}$. $\boldsymbol{\theta}^{(\mathbf{i+1})}$ is then used to start the next loop of the algorithm in the exploration step and so on.

However counter-intuitive it may seem, it can be shown that this stochastic algorithm converges to the stationary distribution, $P\left(\boldsymbol{\theta}^{(\mathbf{i})}\middle|\mathbf{y},H\right)$. Due to the so-called ergodic property,

this convergence occurs for any symmetric exploration function, providing it does not prevent the Markov chain from becoming homogeneous and positive. Proofs are given in Gelman *et al.* (1995) or in Brooks (1998).

The random feature in the acceptance-rejection rule is the key point of the MH-algorithm: if $\boldsymbol{\psi}$ is *a posteriori* more likely than $\boldsymbol{\theta}^{(i)}$, it will always be accepted since $q\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\psi}\right)$ will then be greater than 1 and r is always less than 1. However, in the opposite case, the algorithm can leave the regions of high posterior probability where it mostly gravitates, and jump to regions of lower posterior probability. This way, the algorithm browses randomly the whole domain of θ looking for potential local second order modes of $P\left(\boldsymbol{\theta}|\mathbf{y}, H\right)$.

3.3. Exploration options

Independent Walk Algorithm (IW mode)

{Figure 4}

Each candidate $\boldsymbol{\psi}$ is drawn according to the prior distribution (fig 4a). The fixed symmetric multivariate jump probability distribution $J\left(\boldsymbol{\psi}|\boldsymbol{\theta}^{(i)}\right)$ is therefore $J\left(\boldsymbol{\psi}|\boldsymbol{\theta}^{(i)}\right) = P\left(\boldsymbol{\psi}|H\right)$. The candidate values are accepted or refused according to the decision rule. This is the simplest procedure to obtain a convergent Markov Chain, but it can be dramatically slow.

Random Walk Algorithm (RW mode)

Each candidate $\boldsymbol{\psi}$ is drawn around the last term of the chain without any preference concerning the direction (fig. 4b). Let k represent the number of parameters in a given model,

$\mathbf{I_k}$, a $k \times k$ diagonal matrix with diagonal terms equal to 1 and *s* an adaptive scaling factor used to maintain an acceptable jump. The chosen jump function $J\left(\boldsymbol{\psi} \middle| \boldsymbol{\theta}^{(i)}\right)$ is a hyper-spherical multinormal probability distribution $J\left(\cdot \middle| \boldsymbol{\theta}^{(i)}\right) = N\left(\boldsymbol{\theta}^{(i)}, s \cdot \mathbf{I_k}\right)$. The distribution is centred on $\boldsymbol{\theta}^{(i)}$, the last term of the chain, so probability decreases with the span of exploration. The distribution is non correlated, so every direction of movement has the same probability. This version of the algorithm can be very slow.

Forced Walk Algorithm (FW mode)

Each candidate $\boldsymbol{\psi}$ is drawn around the last term of the chain with a directional preference (fig. 4c). The chosen jump function $J\left(\boldsymbol{\psi} \middle| \boldsymbol{\theta}^{(i)}\right)$ is a multivariate normal probability distribution $J\left(\cdot \middle| \boldsymbol{\theta}^{(i)}\right) = N\left(\boldsymbol{\theta}^{(i)}, s \cdot \mathbf{V}\right)$. The exploration distribution is centred on $\boldsymbol{\theta}^{(i)}$, the last term of the chain but its variance $s \cdot \mathbf{V}$ is defined from the observed covariance matrix of the chain. Candidates will, therefore, be generated in the averaged directions of former terms of the chain.

Although theoretical results have been demonstrated for a homogeneous Markov chain, in practice the algorithm is tuned periodically after a series of sub-runs to increase the speed of convergence. This can be obtained by periodically updating the variance **V** of the jump distribution according to the previous results, therefore adapting to “successful” search directions. Changing the scaling factor s can be interpreted as adjusting the average length of exploration.

Tuning in this manner can also be found in other random search techniques, such as simulated annealing. Here, adjustment of the “temperature” coefficient governs the global "agitation" of the algorithm in order to explore adequately the neighbourhoods of successive iterations.

This last version of the algorithm is more efficient but the chain tends to slow down and to stay around local optimum when posterior parameter distributions are very sharp. Further developments have been proposed to shake the chain and browse other regions of the definition domain of the parameters (Robert, 1996)

3.4. Implementation issues

How to modify the strength of a jump?

The choice of the best modification method is made by checking all available options against mathematically well-known problems. Gelman *et al.* (1995) based their answer by running the MH-algorithm with a multivariate normal model including from 1 to 50 parameters. They concluded that the main criterion to define a suitable value for the jump strength is the observed acceptance rate computed after a given number of iterations. Thus, a high acceptance rate means that candidates resulted from too small jumps around a main distribution mode, and would lead to an overestimation around this mode. On the contrary, a low observed acceptance rate implies too large jump values as if the algorithm was in "a bog without seeing the relief". Gelman *et al.* (1995) show that the initial jump strength must be $\frac{2.4}{\sqrt{d}}$ with d equal to the number of parameters in the model. After a given number of iterations, the jump strength may be modified in order to constrain the observed acceptance rate to stay between 0.23 (d>5) and 0.44 (d=1).

How to assess the Markov chain convergence rate?

The rate of convergence to $P(\cdot|H,\mathbf{y})$ is still under research. For the time being, only empirical answers can be given when judging how long the algorithm should run to sufficiently approximate the limiting distribution. K parallel sequences of the algorithm with

different possible starting points can be launched. After a sufficient number of iterations, the K sequences should stem from the same $P(\cdot|H,\mathbf{y})$ limiting distribution. One can then test that the K sub-samples belong statistically to the same population by various parametric and non-parametric tests. Gelman, Carlin *et al.* (1995) describe one such test based on the R statistic which compares the variability of generated parameters within and across the K sequences.

Is there any limiting distribution that the MH-algorithm can reach?

Another important issue is often overlooked; namely that the existence of the limiting distribution $P(\cdot|H,\mathbf{y})$ should be mathematically proved. The algorithm can be run when the posterior is improper (due to a non-integrable combination of likelihood and improper prior), and may even "generate" samples with good-looking bell shaped histograms! However, the results would be spurious. This problem may be avoided by using priors with a finite domain for $\boldsymbol{\theta}$.

3.5. MCMC and MH Software

The first author has developed a Fortran routine of the FW mode algorithm with convergence rate assessment and jump strength updating, initially designed for use in paleo-ecology (Guiot & al., 2000), but applicable to any model. A program in Matlab for the FW mode algorithm was used to check that the well known linear regression results belongs to the credible regions obtained by the MH-method. Both pieces of software operate in a two-stage process. Firstly, the user has to parameterise the algorithm according to the model, which, unfortunately, requires programming knowledge. Secondly, the modified program must be compiled and run. The linear regression example was run on the Matlab program, and the BOD example ran on the Fortran program.

Non-programming level software is available elsewhere. Carlin and Louis (1996), p.327) propose a useful "software guide" detailing the main software related to Bayesian methods in general and, in particular, MCMC methods. Most of this software is free and available electronically via the World Wide Web. They emphasise a product named BUGS – Bayesian inference Using Gibbs Sampling – developed at the MRC Biostatistics Unit at the University of Cambridge, and initially described by Gilks *et al.* (1992). Dedicated primarily to Gibbs sampling, a recent version includes a Metropolis-within-Gibbs option for multi-modal distributions sampling (see Section 1). Several platform versions are available on the universal resource location (URL) http://www.mrc-bsu.cam.ac.uk. The Microsoft Windows 95/NT version allows statistical models to be constructed simply by drawing graphs, eliminating traditional computer syntax altogether.

4.Applications

We now apply the MH-algorithm to two examples.

The first example presents simulated data from a linear model for which the routine statistical analysis is performed and the posterior distributions are explicit. Our aim here is mainly to reassure the practitioner by demonstrating that the MH procedure will exhibit the same well known results.

The second example is based on a BOD non-linear model for which the statistical analysis was previously considered as intractable from both the classical and the bayesian perspective (at least by environmentalists with a standard background in statistics). Our objective for this example is to illustrate that a Bayesian analysis of many environmental models can nowadays be carried out without the slightest numerical difficulty.

4.1. MH algorithm on a linear example

Consider the following model (simple linear regression of y with respect to x) $y = \theta_1 + \theta_2 \cdot x + \varepsilon$ where ε is drawn from a normal distribution with mean 0 and standard deviation θ_3

Parameters θ_1, θ_2 and θ_3 represent respectively the intercept, the slope and the error standard deviation around the regression line. Using the parameter values $(\theta_1 = 0, \theta_2 = 2, \theta_3 = 1)$, the 10 data were simulated using the normal random generator of the MATLAB statistical toolbox (table 2).

Regression coefficient estimation (by maximum likelihood) gives the following observed values: $\theta_1 = -0.788$ and $\theta_2 = 2.143$ (fig. 5). These values will be compared to their position in the resulting MH algorithm marginal distributions.

{Figure 5}

The theoretical distributions of each parameter are also computed for comparison. Box and Tiao. (1973) proved that, when a so-called non-informative prior is elicited for the prior parameter distribution, the following results hold for the posterior distributions:

Let $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ be the usual least squares estimates for $(\theta_1, \theta_2, \theta_3)$, obtained as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_{10} \end{pmatrix} \qquad \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad \hat{\theta}_3 = \sqrt{\frac{1}{8}\left\|\mathbf{y} - \mathbf{X}\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}\right\|^2}$$

{Figure 6}

- Unconditional on the knowledge of (θ_1, θ_2) the posterior distribution of θ_3 is such that the variable $z = (10-2)\left(\frac{\hat{\theta}_3}{\theta_3}\right)^2$ is a chi-square random variable with (10-2) degrees of freedom. Figure (6a) plots the posterior distribution of θ_3 , both theoretical (*based on a*

change of variable from the distribution of z) and as obtained by 1000 runs of MH algorithm (after a burn-in sequence of 9000 runs that were discarded)

- Conditional on the knowledge of θ_3, (θ_1,θ_2) has a bivariate normal distribution centred on $(\hat{\theta}_1,\hat{\theta}_2)$ with a variance covariance matrix: $\theta_3\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}$. Unconditional on the knowledge of θ_3, (θ_1,θ_2) has a Student distribution centred on $(\hat{\theta}_1,\hat{\theta}_2)$ with a scale parameter of: $\hat{\theta}_3\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}$ and 8 degrees of freedom. Figures (6b) and (6c) describe the Student marginal distributions for θ_1 and θ_2 both with their analytical expression and by drawing histograms from the MH pseudo-sample.

Visual inspection shows a close agreement between the theoretical results and those obtained from the MH algorithm and that observed regression coefficient for $\theta_1(-0.788)$ and $\theta_2(2.143)$ are modal values.

4.2. MH Algorithm on BOD example

As stated above, the statistical analysis of this environmental problem was previously considered as intractable from both a classical and the Bayesian perspective. We start the analysis by distinguishing the prior knowledge from the posterior knowledge. The prior knowledge is what is known before an experiment: primarily, the definition of a model to describe the BOD versus time measurements. It also includes the physical interpretation of the model parameters. The data collected from a particular site improve this knowledge – particularly the knowledge related to that site. We try to illustrate how the prior degree of belief is integrated by the MH-algorithm to produce posterior knowledge.

4.2.1. The BOD model and the environmental context

As previously established (see Section 1), the BOD against time measurements are usually described by the following model:

$$BOD(t) = BOD_u\left[1 - \exp(-K_{20} \cdot t)\right] + \varepsilon_t$$

where $BOD(t)$ is the biochemical oxygen demand at time t. ε_t stands for independent normal errors with constant variance, BOD_u is the ultimate BOD, K_{20} denotes the constant rate of the organic matter decomposition ($t \in \{1.2.3.4.5.7.9.11\}$, at 20°C), and σ stems from the natural experimental variability when collecting the data.

Note that, by introducing the variable $BOD_r(t)$ as the remaining BOD over time, and a starting value $BOD_u(t)$ there is an equivalent expression of the model as a continuous first order reaction :

$$\frac{dBOD_r}{dt}(t) = -K_{20} \cdot BOD_r(t)$$

where $BOD_r(t) = BOD_u(t) - BOD(t)$ and $BOD_r(t=0) = BOD_u$.

We hope that the last remark demonstrates the wide range of applications covered by this kind of model. It has already been used on the same data in some statistical handbooks. Berthouex and Brown (1994) looked for an optimal experiment design as a compromise between the number of measurements and the size of the credible region for the parameters. Bates and Watts (1988) studied BOD data and several other data sets related to non-linear regression modelling, and built credible regions according to the different approximation methods. Ritter and Tanner (1992) and Tanner (1992) illustrate an improved Gibbs sampler on the BOD data. However, they intentionally take non realistic prior distributions for BOD_u

and K_{20} so that the conditional distributions needed for the Gibbs sampler are available in closed form.

In this section, we take the point of view of a practitioner, and describe how to obtain a proper joint posterior distribution for BOD_u, K_{20} and σ.

The main water quality assessment indicators are $BOD(5)$ but also BOD_u and K_{20}. BOD_u and K_{20} are parameters of the model. $BOD(5)$ is not a model parameter but is a less expensive way to quantify the BOD level of a given water sample. The range of $BOD(5)$ and BOD_u values is from 0 to 5 mg/L in rivers. K_{20} lies between 0.05 and 0.40 day^{-1}.

4.2.2. Building a posterior distribution

We first define prior distributions and the likelihood function for each of the three parameters.

Informative or non-informative prior distributions:

The choice between an informative and a non-informative prior is dependant upon what is known about the distributions of BOD_u, K_{20} and σ.

Uniform prior distribution for BOD_u between 0 and 5. There is no prior preference for any values between 0 and 5. We specify a non-informative distribution, that is to say, a uniform distribution for BOD_u, this parameter appearing as a location parameter:

$$P(BOD_u|H) = \alpha_1 \qquad \textbf{(4.1)}$$

Uniform prior distribution for K_{20} between 0.05 and 0.40. The same reasoning lead to the same type of prior :

$$P(K_{20}|H) = \alpha_2 \qquad \textbf{(4.2)}$$

Non informative prior distribution for σ. The prior distribution for the scale parameter σ has to be σ^{-1}-shaped (Berger (1985)). So, we consider:

$$P(\sigma|H) = \frac{\alpha_3}{\sigma} \quad \textbf{(4.3)}$$

From equations , and , a non informative prior distribution is obtained for $\boldsymbol{\theta} = (BOD_u, K_{20}, \sigma)$:

$$P(BOD_u, K_{20}, \sigma|H) = \alpha_1 \cdot \alpha_2 \cdot \frac{\alpha_3}{\sigma} \quad \textbf{(4.4)}$$

We need only to remember for next steps that:

$$P(BOD_u, K_{20}, \sigma|H) \propto \frac{1}{\sigma} \quad \textbf{(4.5)}$$

Likelihood function of the model :

Let us denote:

$$\forall t \in \{1,2,3,4,5,7,9,11\}, BOD_{est}(t) = BOD_u \cdot (1 - \exp(K_{20} \cdot t))$$

the prediction associated with time t. The model can be expressed as:

$$t \in \{1,2,3,4,5,7,9,11\} \begin{cases} BOD_{est}(t) = BOD_u \cdot (1 - \exp(K_{20} \cdot t)) \\ BOD(t) = BOD_{est}(t) + \varepsilon_t \\ \varepsilon_t \leftarrow N(0, \sigma) \end{cases}$$

or

$$t \in \{1,2,3,4,5,7,9,11\} \begin{cases} BOD_{est}(t) = BOD_u \cdot (1 - \exp(K_{20} \cdot t)) \\ BOD(t) \leftarrow N(BOD_{est}(t), \sigma) \end{cases}$$

According to BOD_u, K_{20} and σ values, a wide range of models are possible. For this reason, we need a quantitative criterion to evaluate the quality of each model. Our model

quality criterion is based logically on the ability of the models to fit observed values $(t, BOD(t))_{t\in\{1,2,3,4,5,7,9,11\}}$. This is what we call the *likelihood function*. The likelihood function for a non-linear regression model is the probability distribution function of a normal random variable :

$$\forall t \in \{1,2,3,4,5,7,9,11\}, P(BOD(t)|BOD_u, K_{20}, \sigma, H) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(BOD(t) - BOD_{est}(t))^2}{2\sigma} \right]$$

If we assume independence between observations, the likelihood function becomes:

$$P(BOD(1),...BOD(11)|BOD_u, K_{20}, \sigma, H) = \prod_t P(BOD(t)|BOD_u, K_{20}, \sigma, H) \quad \textbf{(4.6)}$$

Note the presence of the classical square sum of errors around the regression curve in the likelihood function expression. Logically, the likelihood of the model will increase when the sum of squares of errors will decrease for any given positive value of σ.

Posterior joint distribution of the parameters:

The posterior joint distribution is known up to a factor, as it is expressed in Bayes' formula as a product of the former quantities expressed in and :

$$P(BOD_u, K_{20}, \sigma | BOD(1),..., BOD(11), H) \propto P(BOD(1),..., BOD(11)|BOD_u, K_{20}, \sigma, H) \cdot P(BOD_u, K_{20}, \sigma)$$

The purpose of the MH-algorithm is to accept or reject successive BOD_u, K_{20} and σ values according to the posterior joint distribution evaluation.

Let us denote $\boldsymbol{\theta}^{(\mathbf{i})} = \left(BOD_u^{(i)}, K_{20}^{(i)}, \sigma^{(i)}\right)$ as the vector of parameter after i iterations and $\boldsymbol{\psi} = \left(BOD_u^{(c)}, K_{20}^{(c)}, \sigma^{(c)}\right)$ a candidate vector. The algorithm described in a former section is applied with an acceptance/rejection rule based on:

$$q\left(\boldsymbol{\theta}^{(\mathbf{i})}, \boldsymbol{\psi}\right) = \frac{P\left(BOD_u^{(c)}, K_{20}^{(c)}, \sigma^{(c)} | \mathbf{y}, H\right)}{P\left(BOD_u^{(i)}, K_{20}^{(i)}, \sigma^{(i)} | \mathbf{y}, H\right)}$$

4.2.3. MH-algorithm in action

The MH-algorithm initialises five independent Markov chains with random values taken from the prior thresholds (Table 3). It runs 2,500 times in RW mode and 11,500 times in FW mode before reaching convergence in the last 1,000 iterations. Convergence is tested by simultaneously launching five independent sequences. The acceptance rate for the last 1,000 iterations equals .31 and is not far from the value recommended by Gelman et al (1995) (0.36 when *d* is 3)

Figure 7 traces the five sequences of values for each parameter along 13,001 iterations.

{Figure 7}

From iteration 1 to 2500, the five chains are in a RW mode. The chains browse the whole definition space with no evidence of convergence. From iteration 2,501 to 13,001, the five chains are in a FW mode. Each chain evolves in a main direction indicated by the averaged direction of the previous values. The leading direction is updated every 60 iterations. This part of the chart is characterised by two phases:

- from iteration 2,501 to 6,500, a phase of transition with important changes resulting from several direction updates;
- from iteration 6,500 to the end of sequences, a phase of relative stabilisation around optimal values.

From the last 1000 iterations, 1 in 10 are saved so 100 MC terms are retained from each sequence. Finally, the posterior distribution simulation is based on the 500 last iterations from each sequence.

4.2.4. Posterior marginal distributions

{Figure 8}

The algorithm gives 500 observed vectors (BOD_u, K_{20}, σ). From this, posterior marginal distributions are easily computed by taking into account the whole range of values for each parameter (figure 8).

90-percent credible intervals for BOD_u, K_{20} and σ are built from the posterior marginal distributions, giving the following results:

$$I(BOD_u) = [2.315, 2.774],\ I(K_{20}) = [0.165, 0.235],\ I(\sigma) = [0.046, 0.119].$$

A first interpretation of these results is that the sampled river is highly polluted ($K_{20} > 0.12$, Krenkel and Novotny (1980)).

4.2.5. Predicted values distributions

{Figure 9}

For any given value of t and for each of the 500 resulting vectors (BOD_u, K_{20}, σ) the prediction formula can be applied in the following way:

(i) compute $BOD_{est}(t) = BOD_u \cdot (1 - \exp(K_{20} \cdot t))$

(ii) sample a centered normal random value with σ standard deviation

(iii) add the two terms

This gives 500 values of $BOD(t)$ for one given value of t. This was carried out for $t \in \{1,2,3,4,5,7,9,11\}$ and corresponding histograms were built. A further histogram was added for t=20 as a forecast (fig. 9). For the available observations, one can easily observe the agreement between the distribution and the observed values of y.

90-percent credible intervals have been built for BOD(5) and BOD(20): $I(BOD(5)) = [1.4248, 1.7232]$ and $I(BOD(20)) = [2.2461, 2.6894]$. Consequently, $BOD(5)$ is

almost certainly over 1mg/L, which is a common threshold for clean water. $BOD(20)$ credibility interval leads to the same diagnosis. A more interesting discussion is on the comparison of $I(BOD_u)$ with $I(BOD(20))$ and with every credible interval of t>20, e.g. $I(BOD(1000))$:

$$I(BOD(20)) = [2.246, 2.689]$$
$$\ldots$$
$$I(BOD(1000)) = [2.279, 2.816]$$
$$I(BOD_u) = [2.315, 2.774]$$

As BOD_u represents the "ultimate BOD", $I(BOD(t))$become closer and closer to $I(BOD_u)$ as t tends to plus infinity. Further, BOD_u is unaffected by the uncertainty in K_{20}. On the contrary, for any finite t value, $BOD(t)$depends on K_{20} as well as BOD_u, so any uncertainty about K_{20} will result in an extra uncertainty about $BOD(t)$. Consequently, $I(BOD(t))$ will always be wider than $I(BOD_u)$ and the credible region $I(BOD(1000))$ entirely contains $I(BOD_u)$.

The BOD example illustrates the extent to which MH-algorithm allows the environmental scientist to define credible intervals for BOD_u, K_{20} and σ, as well as any non observed $BOD(t)$ value. Credible intervals for observed $BOD(t)$ values are performed and compared with observed values to check that MH pseudo-sample has converged to the target distribution. Several tools are now available to facilitate (jump and direction updates in a FW mode of the algorithm) and to assess (CVG rate periodical calculations) the convergence.

Many other factors affect the BOD : temperature significantly affects the reaction rate and also slightly influences the ultimate BOD ; extreme pH values in wastewaters may interfere in the experimentation.... These factors and many others (Krenkel and Novotny,

1980) could be integrated in a more complex BOD model. The MH-algorithm could be used for this model, with no supplementary difficulties.

Finally, a useful feature is the ability to limit the influence of outliers on credible intervals. Distributions with a relatively strong mode and short tails, e.g. Gaussian distributions are particularly outlier-sensitive. Gelman, Carlin *et al.* (1995) suggest modelling outliers in the error term by using long-tailed distributions such as the Student's distribution.

5. Conclusion

In this paper, we have stressed the usefulness and tractability of the MH-algorithm for statistical analysis of environmental models in the Bayesian context. It has been shown that MH-algorithm can be easily implemented and overcome the technical difficulties often associated with rare data, imprecise knowledge and non-linear effects. Of course, no technique can ever overcome conceptual modelling complications and all results given in this paper are subject to the basic constraints of scientific reasoning, i.e. the ability to design a (realistic) model and the will to deliver quantified statements in addition to qualitative judgements.

Based on this work, it should be possible to analyse other environmental models with more parameters and systematically check the sensibility of the analysis to the prior – i.e. how the part stemming from the environmentalist's expertise may influence the conclusion of the study. An interesting aspect of the MH algorithm is the easy generation of predictive values, meaning that once the environmentalist has assembled a model, data and prior knowledge, various conditional future scenarios can be obtained. Moreover, within the Bayesian setting, inference and prediction are achieved without recourse to the usual asymptotic approximations needed in conventional analysis of non-linear models.

6. References

Bates, D. M. and Watts, D. G. (1988). "Nonlinear Regression Analysis and Its Applications". New York: John Wiley & Sons.

Bayes, T. (1763). “An essay towards solving a problem in the doctrine of chances”. *Phil. Trans. Roy. Soc* **53.** 370-418. Reprinted, with an introduction by georges Barnard, in 1958 in Biometrika, 45:293-315.

Berger, J. O. (1985). "Statistical Decision Theory and Bayesian Analysis". New York: Springer Verlag.

Berthouex, P. M. and Brown, L. C. (1994). "Designing experiments to estimate parameters in non linear models". In *Statistics for environmental engineers*, eds. P. M. Berthouex and L. C. Brown. Boca Raton, USA: Lewis Publishers.

Box, G. E. P. and C., T. G. (1973). "Bayesian inference in statistical analysis". Reading, Massachusetts, USA: Addison-Wesley.

Brooks, S. P. (1998). “ Markov chain Monte Carlo and its application,”. *The Statistician, Journal of the Royal Statistical Society, Series D.* **47.**1: 69-100.

Carlin, B. P. and Louis, T. A. (1996). "Bayes and empirical Bayes methods for data analysis". London, England: Chapmann & Hall.

Casella, G. and Georges, E. I. (1992). “Explaining the Gibbs Sampler”. *The American Statistician* **46.**3: 167-174.

Chaouche A., Parent E.(1999). "Inférence et validation bayésiennes d’un modèle de pluies journalières en régime de mousson". Hydrological Sciences Journal 44(2):199-220.

Chen, S., *et al.* (1990). “Nonlinear system identification using neural networks”. *Int. J. Control.* **51.**6: 1215-1228.

DeFinetti, B. (1937). "La prévision, ses lois logiques, ses sources subjectives". Paris: Herman.

Gamerman, D. (1997). "Markov Chain Monte Carlo - Stochastic Simulation for Bayesian inference": Chapman & Hall Eds.

Gelman, A., *et al.* (1995). "Bayesian Data Analysis". London, UK: Chapman & Hall Eds.

Geman, S. and Geman, D. (1984). “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6.** 721-741.

Gilks, W. R., *et al.*, Eds. (1996). "Practical Markov Chain Monte Carlo". New York., Chapman-Hall.

Gilks, W. R., *et al.* (1992). “Software for the Gibbs sampler”. *Computing Science and Statistics,* **24.** 439-448.

Guiot, J., Torre, F., Jolly D., Peyron O., Boreux J.-J. and Cheddadi R. (2000). “Inverse vegetation modelling by Monte Carlo sampling to reconstruct palaeoclimates under changed precipitation seasonality and CO_2 conditions: application to glacial climate in Mediterranean region”. *Ecological modelling, **127**, 119-140.*

Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their application”. *Biometrika* **57.** 97-109.

Kiely, G. (1997). "Environmental engineering". Berkshire, England: Irvin / McGraw-Hill.

Krenkel, P. A. and Novotny, V. (1980). "Water quality monitoring". London, UK: Academic Press.

Kuczera, G. (1983). “ Improved parameter inference in catchment models: 1. Evaluating parameter uncertainly”. *Water Resources Research* **19.**5: 1151-1162.

Kuczera G., Parent E. (1998). "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm", Journal of. Hydrology (211)**1**-4:. 69-85.

Lecoutre, B. (1997). “C'est bon à savoir! Et si vous étiez un bayésien qui s'ignore!”. *Modulad* **18.** 81-87.

Lindley, D. V. (1972). “Bayesian statistics. A review”. CBMS-NSF Regional Conference Series in Applied Mathematics 2. *SIAM* (Eds). v + 83 pp.

Marske, D. and Polkovski, B. (1972). “Evaluation of methods for estimating biochemical oxygen demand parameters”. *Journal of Water Pollution Control Federation* **44.**10: 1987-2000.Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953). “Equations of state calculations by fast computing machines”. *J. Chem. Phys.* **21.** 1087-1092.

Ritter, C. and Tanner, M. A. (1992). “Facilitating the Gibbs Sampler : The Gibbs Stopper and the Griddy-Gibbs Sampler”. *Journal of the American Statistical Association* **87**(419): 861-868.

Robert, C. (1992). "L'analyse statistique bayésienne". Paris, France: Economica.

Robert, C. (1996). "Méthodes de Monte-Carlo par chaînes de Markov". Paris, France: Economica.

Tanner, M. H. (1992). "*Tools for statistical inference: Observed data and data augmentation methods*": Springer-Verlag.

Van Straten, G. and Keesman, K. J. (1991). “Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example”. *J. Forecasting* **10.** 163-190.

Table 1 : Biochemical Oxygen Demand vs. Time. Data from Marske & Polkovski (1972)

Table 2 : Artificial data simulated using the normal random generator of the MATLAB statistical toolbox (see text for the model definition).

Table 3 : Initial values for the five independent Markov chains generated by MH algorithm in the BOD data example. BOD_u is sampled from an uniform (0-5) distribution, K_{20} from an uniform (0.05-0.40) and σ from an uniform (0.001-1000).

Figure 1 : Scatterplot of the BOD data and possible fittings by three different curves. These curves are from the BOD model with varying parameter values.

Figure 2 : Prior and Posterior Joint Distribution of (θ_1, θ_2). a - Uniform two-dimensional prior in [-20, 50] x [-2, 6] ; b - 80%, 90%, 95% and 99.9% Normal likelihood contours for (θ_1, θ_2) labelled by approximate frequency coverage using the F statistic. From Tanner & al. 1993.

Figure 3 : MH-algorithm inputs and outputs using Bayesian approach and terminology.

Figure 4 : Independent, Random and Forced Walk Version of the algorithm. Algorithms are ordered according to their efficiency in simulating the posterior joint distribution of (θ_1, θ_2). Numbers indicate iteration rank and a particular (θ_1, θ_2) value ; some are circled (accepted), others are struck through (rejected) ; a - Independent Walk Algorithm defines the Markov Chain MC1 = (1, 2, 2, 2, 5, 5, 5, 8, 9) ; b - Random Walk Algorithm defines MC2 = (1, 1, 3, 4, 4, 6, 7, 8, 9, 10, 11) ; c - MC3 = (1,..n, n+1, n+2, n+3, n+4, n+4, n+6), Forced Walk Algorithm changes from Random Walk Algorithm after n iterations. Repetition means that a jump has been rejected and that the MC stays in the same place.

Figure 5 : scatterplot of the data $(x_i, y_i)_{i=1,10}$ and the line with equation $y = -0.78800 + 2.14327 \cdot x$

Figure 6 : Theoretical and MH algorithm distributions for the three parameters involved. The curves and the histograms show respectively the theoretical and the MH algorithm distributions.

Figure 7 : Five simultaneous Markov Chains simulating the three dimensional parameter (BOD_u, K_{20}, σ) ; Each graph plots the parameter component value against the iteration number from 0 to 13,000. Arrows indicates the starting values of each chain as shown in table

3. The windows show the part of each sequence where the algorithm converges to the posterior distributions of BOD_u, K_{20} and σ.

Figure 8 : Marginal posterior distribution of BOD_u, K_{20} and σ. P5, P50 and P95 are respectively the 5^{th}, 50^{th} (or median) and 95^{th} percentile of each distribution.

Figure 9 : y prediction for observed values x_i ; dashed lines represent observed y values and solid bars the distribution of y predictions. P5, P50 and P95 are respectively the 5^{th}, 50^{th} (or median) and 95^{th} percentile of each distribution.