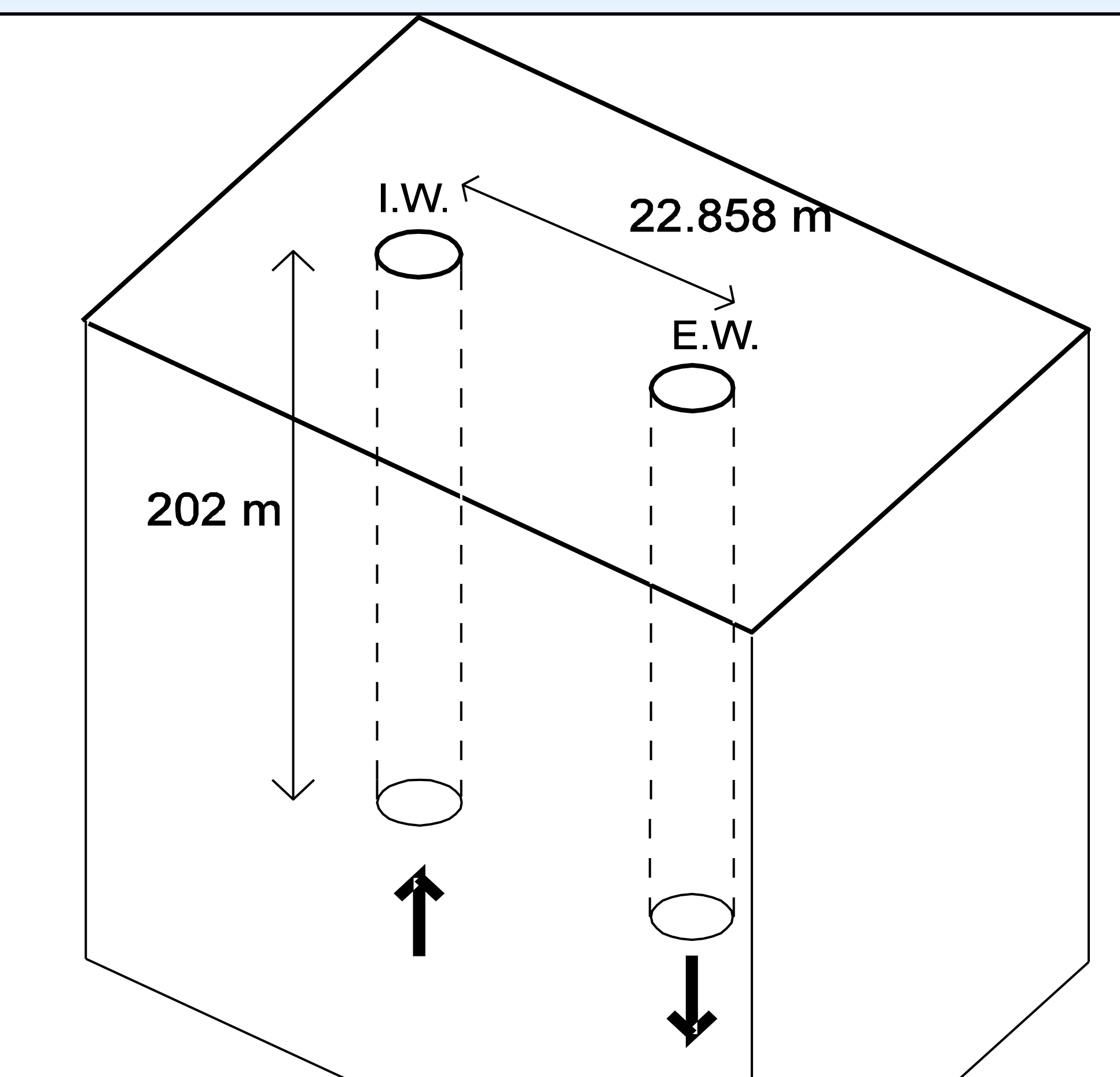
Estimating Different Regimes in a Tracer Breakthrough Curve with Bayesian Statistics

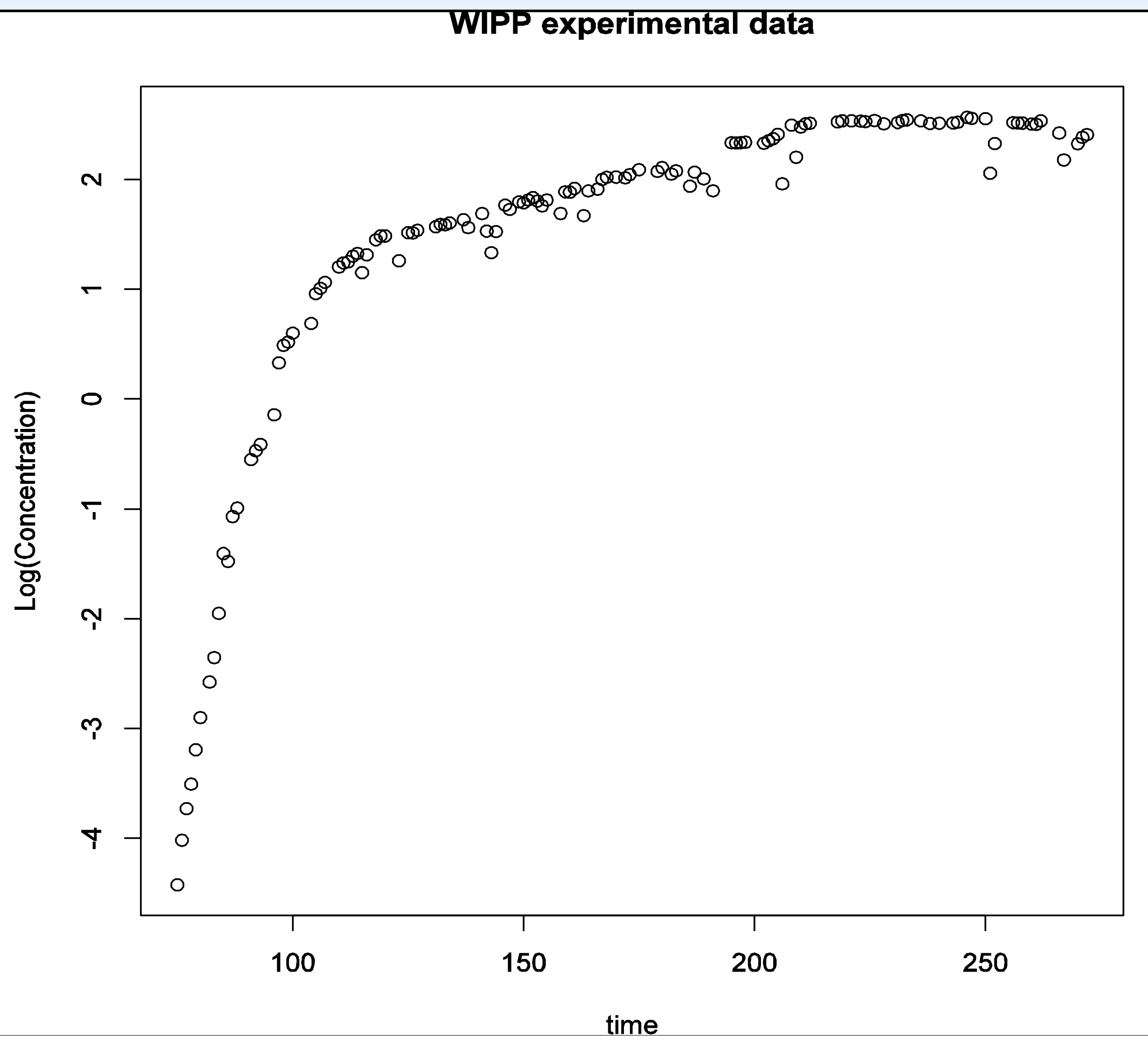
*Bruno Mendes*T (mendes@ams.ucsc.edu), David Draper**

(*)University of California, Santa Cruz, Applied Math & Statistics Depart. Mail Stop SOE2,1156 High Street Santa Cruz, CA, USA, (T) Centro de Geofisica de Evora, Evora, Portugal



Introduction

The authors have been working with a data set produced in an early experiment conducted at the Waste Isolation Pilot Plant site [Gonzales, 1984]. Previous investigators have worked with this data set as if the physical conditions during the experiment were stable, but Gonzales acknowledged in the original study that this may not be true. We propose using Bayesian statistics and the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm [Green, 1995] to estimate the number and duration of regimes that were present in the experiment. Our hypothesis is that for reasons particular to the experiment, the pumping in the extraction well changed during the study, so the experiment is not a single steady-state, but a collection of several steady-state regimes. We use a simple 1-dimensional transport model to explain the breakthrough curve for each steady-state regime, and the number and duration of different regimes are included as free parameters to be inferred from the data. RJMCMC simulation provides an approximation to posterior probability distributions for the number of change points, time of occurrence of these changes, and also probability distributions for the physical parameters that characterize each pumping regime. We will also show that this problem can be seen as a variable-selection problem, and that the method can be readily applied to other situations where variable selection is an ambition of the modelers.



Description of method

The Bayesian paradigm allows us to use probability distributions to describe uncertainty about parameters of a mathematical model. The statistical model can be represented by:

(Measured concentration)$_i$=(deterministic model output)$_i$ + (*stochastic error)$_i$*,

where i=0,..., 124 days, concentration is measured in kg/l. The deterministic model is described below

$$\frac{\partial C}{\partial t} = v\frac{\partial C}{r\partial r} + D\frac{\partial^2 C}{\partial r^2}, v \text{ is the velocity of water and } D \text{ the dispersivity}$$

The stochastic errors above are assumed to be normaly distributited with mean 0 and homoscedastic, with a standard deviation σ which is to be inferred from the data.

The deterministic model can applied to different sub-intervals of the total duration of the experiment with different pairs (ν_j,D_j), defining different regimes as the experiment develops.

The satistical model above works as a meta-model, ie. it encompasses all the deterministic sub-models and experimental error (represented by σ).

Our main goal is to solve the inverse problem in order to infer from the data the number and location of change-points, and the values of the pairs (ν_j,D_j), **for each sub-model j.**

In terms of Bayesian statistics, this problem translates to finding the posterior distributions for p(θ|data), where θ is the vector (ν_j,D_j, σ_j). Formally, p(θ |data) is calculated by

$$p(\theta \mid \text{data}) = \frac{\text{f}(\text{data} \mid \theta)\, p(\theta)}{\iiint \text{f}(\text{data} \mid \theta)\, p(\theta) d\theta}$$

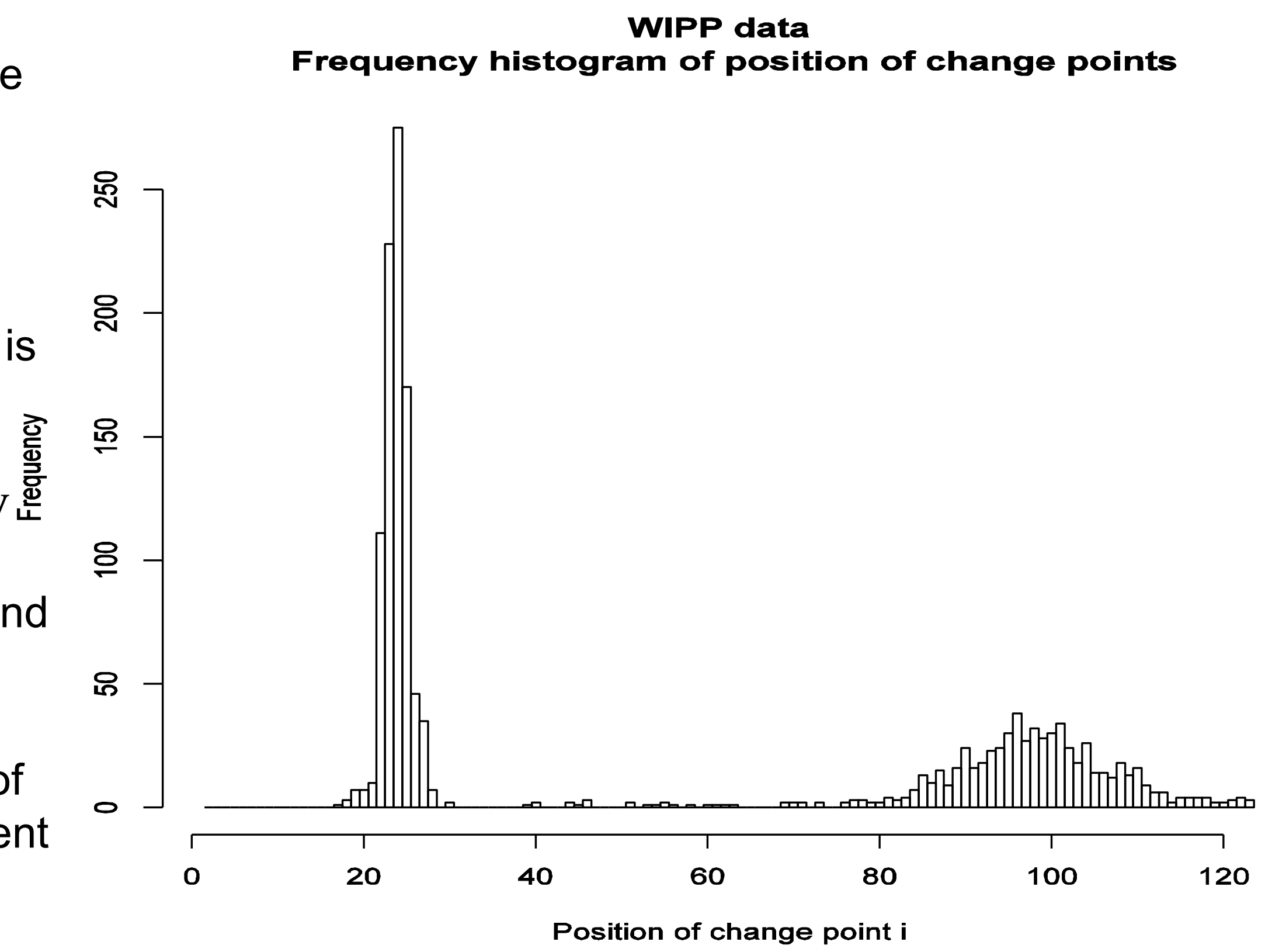
where f(.) is known as the likelihood function. The integrals are multi-dimensional and difficult to solve in closed form. There is hope, though, in approximating these distributions by simulation using *Markov Chain Monte Carlo* methods. In particular, since the very number of parameters *j* is unknown and to be inferred from the data, one needs to use the method known as *reversible jump Markov chain Monte Carlo -RJMCMC* [Green, 1995].

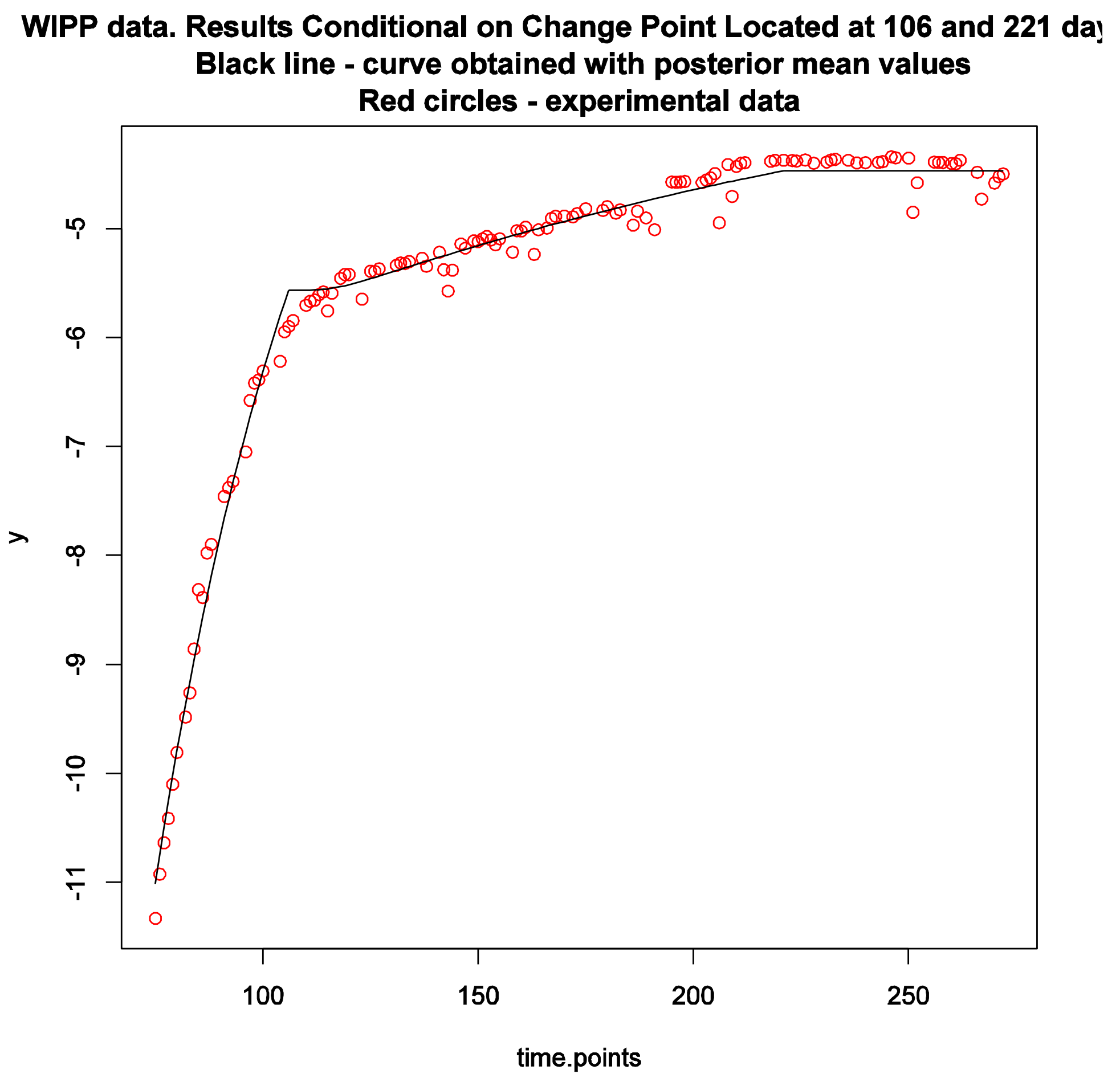
This problem can also be seen as a variable selection problem, i.e., how many variables do we really need to best describe the data? The RJMCMC method has an in-built procedure to penalize over-fitting. This is because on the one hand a high number of variables implies an increase in dimensionality of the parameter's space and this usually dilutes the posterior probability distribution associated with that region. On the other hand, the simulation procedure penalizes acceptance of moves to parameter space regions with low posterior ratios, basically censuring in a natural way moves to high dimensional parameter regions.

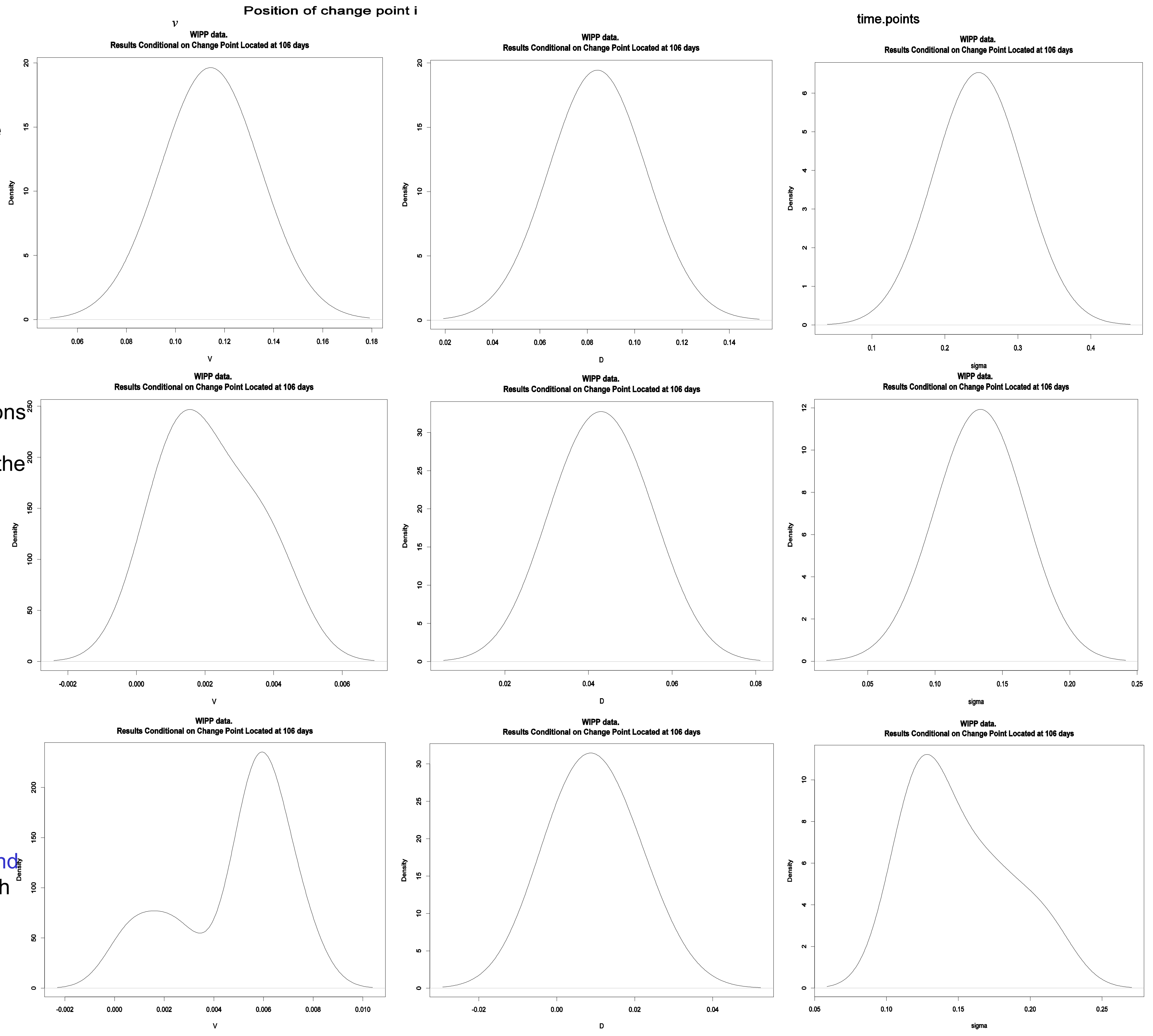
RESULTS (figures on the right)

The upper left figure represents the estimates for the location of the change-points. The evidence is very strong for the existence of three different regimes. The second, third and fourth rows of figures represent the marginal posterior distributions for (ν_j,D_j, σ_j). for each regime with units m/day, m^2/day and kg/l, respectively. Finally, the upper-right figure represents the analytical curves obtained by using the mean of each of the marginal posterior distributions as the point estimate for (ν_j,D_j, σ_j) superimposed on the data concentration in a log-scale. The posterior means are, respectively (0.11428, 0.084340, 0.24611) for branch 1, (0.002112, 0.04311, 0.13288) for branch2 and 0.004813, 0.009393,0.15048) for branch3.

Results







Conclusions

The RJMCMC method basically “compares” changes of values of concentration, from one time point to another, with the estimated value of σ (which is an estimate of experimental error), so the model’s capability of detecting change-points depends on how good our model is at estimating that quantity. Although our statistical method is a basic one, it does an excellent job at identifying the two major change-points. There are some additional candidates for change-points that are visually enticing; these additional candidates are difficult to infer with our simple statistical model, but if we improve our statistical model for the experimental error σ we believe we would be able to distinguish more subtle change-points in the data. We are currently investigating a way to measure the precision with which our method can detect change-points and also studying ways to improve it.

From the geophysics point of view, there is the important point of justifying these regime-changes (even though they are quite obvious). We believe the main reason for changes in the hydrology of the experiment arose most likely from the unusual duration of it (almost one full year) and also from some problems reported on instrumentation by Gonzalez [1984].

We have shown how Bayesian statistical methods can be applied to a problem in hydrogeology not only to infer values of physical parameters from experimental data, but also to infer possible change of regime points in the data. In this particular experiment, we identified 3 different regimes with different physical parameters associated with each of them.

Acknowledgements

This work was supported by Fundação para a Ciência e Technologia under grant SFRH/BPD/14406/2003

References

Gonzales D, Bentley C (1984). Field test for effective porosity and dispersivity in fractured dolomite: the WIPP, Southeastern New Mexico. In Groundwater Hydraulics, Rosenshein JS, Bennett GD (editors), Washington DC: American Geophysical Union, 207–221.

Green P (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82, 711–732.