FUNCTIONAL DATA ANALYSIS IN GROUNDWATER MODELING

BRUNO MENDES[1,2] DAVID DRAPER[1]

[1]Faculdade de Engenharia, Universidade Católica Portuguesa
[2]Department of Applied Mathematics and Statistics, University of California, Santa Cruz

Abstract

In groundwater contamination studies uncertainties are a constant presence. We have in previous studies classified the different sources of uncertainty one can encounter in such studies [Draper et al, 1999] and we propose a framework to tackle them. We have proposed to use probability to describe those uncertainties and that there are four main sources of uncertainty that affect observed dose of contaminant in the biosphere, with a conditional probabilistic structure relating them: scenario uncertainty, structure uncertainty, parametric uncertainty and predictive uncertainty. We have been developing work in all of these types of uncertainty and the present work will focus on scenario uncertainty. The set of scenarios we used was developed by [Prado et al, 1998], and consist of different sets of hydrogeological setups for a deep underground storage system for high-level nuclear waste repository. We use a set of differential equations to model the behavior of this system in case of an accidental leak of radioactive waste. This model produces data of different kinds, one of them is a collection of data points that can be seen to approximate a continuous curve of observed radioactive dose versus time at a particular point in the biosphere. In this paper we describe statistical methods that are useful when the outcome of interest is an entire function rather than just a single numerical summary of the function. Functional Principal Component Analysis is performed on the curves in order to find the curve's main modes of variability, also ANOVA-like calculations are made were we identify the effects of alternative scenarios on expected dose. We performed functional linear regression of the program's input parameters on the whole dose curve. It is shown that the application of these innovative techniques brings about new important information on the uncertainties we should expect from computer simulations in this field; we noted that scenario effects can account for as much as 40 times increase in the uncertainty of predicted doses.

1. Introduction

A key issue in the consolidation process of nuclear fuel cycle is the safe disposal of radioactive waste. At present, deep geological disposal based on multibarrier concept is considered the most promising option (which consists in a deep underground chamber within which radioactive materials such as spent fuel rods are entombed in layers of concrete and other barriers). The containment capability of this concept ultimately depends on the reliability of mechanical, chemical and physical barriers offered by the geological formation itself. In spite of worldwide efforts for the last three decades, physico-chemical

behavior of disposal system over geological time scales (hundreds or thousands of years) is far from known with certainty.

1.1. **Sources of uncertainty.** With partners in Italy (A. Saltelli), Spain (P. Prado), and Sweden (A. Pereira), we were involved from 1996 to 1999 in a European Commission project, `GESAMAC`, which aimed to capture most or all relevant sources of uncertainty in predicting what would happen if disposal barriers were compromised in future by processes like geological faulting, human intrusion, climatic change. One of the ultimate goals of that project was to forecast outcomes including radiologic dose for people on earth's surface as function of time depending on several parameters: how far disposal chamber is underground, what are the hydro-geologic conditions of the underground geologic system, and many other (totaling in some scenarios 16 physical parameters). We developed an uncertainty framework for `GESAMAC` and other similar projects, with six ingredients:

- **Past observables** D, if available, would consist of readings on radiologic dose under laboratory conditions relevant to those likely to be experienced in geosphere (no accidents to date of type whose probabilities we are assessing).
- **Future observables** y^* consist of dose at given location L, t years from now, as L and t vary.
- **Scenarios** X detail different sets of likely geosphere conditions at location L and time t, as result of human intrusion, faulting, climate.
- **Structural possibilities** S include different combinations of chemical processes (sorption, matrix diffusion), different sets of partial differential equations (PDEs) to model them.
- **Parametric uncertainty** arises because precise values of relevant physical constants appearing in PDEs are unknown. Note that parameters may be specific not only to structure but also to scenario (an early ice-age climatic scenario would have certain chemical constants driving it, whereas worst-case geologic fracture scenario would be governed by different constants);
- **Predictive uncertainty** will be as speculative as past data, and might be based on things like discrepancies between (a) actual, predicted lab results, extrapolated to field conditions or (b) actual, predicted failure rates in other related industries (e.g., nuclear power generation).

1.2. **Set of alternative scenarios.** The set of scenarios we used was developed by [Prado et al, 1998], below we present a brief explanation of them.

- **Reference (Ref) Scenario**. Assumes that the present conditions will be maintained in the future.
- **Fast Path (FP) Scenario**. The distance from the repository to the surface is reduced considerably (due to erosion, drilling, etc)
- **Additional Geosphere (AG) Scenario**. Involves changes dealing with an additional layer in the geosphere (accumulation of debris left behind by a retreating glacier, geological dynamics that creates a longer rock pathway, etc).
- **Glacial Advance (GA) Scenario**. Tries to predict the conditions when the next glacial age comes (conditions typical of an advancing glacier).

- **Environmental Induced Changes (EIC) Scenario**. Changes induced by tectonic movement of the rock, or changes induced by human activities in the hydrologic characteristics of the area surrounding the disposal site.
- **Human Disposal Errors (HDE) Scenario**. This includes errors done during repository construction, the disposal of the wastes or any other activities connected with the repository operation.

1.3. **Deterministic model.** Eguilior and Prado [Prado and Eguilior, 1996] have produced a software package, `GTMCHEM`, which solves a coupled differential equation that models the transport and dispersion of a contaminant in a underground hydrological system.

$$\frac{\partial C_i}{\partial t} = -V\frac{\partial C_i}{\partial X} + D\frac{\partial^2 C_i}{\partial X^2} + SoSi, \tag{1}$$

where C represents concentration (mols/m^3), t is time (yr), X is the space coordinate (m), V is the groundwater velocity (m/yr), D is the hydrodynamic dispersion (m^2/yr), $SoSi$ is the source/sink term in which the chemical reactions are included.

2. Uncertainty Propagation: Calculations

Conditional on a particular choice of scenario, structure, and parameters, `GTMCHEM` produces deterministic output for radioactive dose that may be compared with actual observables (if any) to assess predictive uncertainty. With the six ingredients listed above, the goal in uncertainty propagation is to produce two types of predictive distributions: **scenario-specific** and **aggregate**. The only hope of doing this in a way that captures all relevant sources of uncertainty is a fully Bayesian analysis [Draper, 1995, Draper et al, 1999]. In the Bayesian approach past data D (if any) are known; future observable outcome(s) y^* are unknown, and to be predicted; and we must behave as if the sets X and S of possible scenarios and structures are known. Then the **scenario-specific** predictive distribution $p(y^*|S, x, D)$ for y^* given D, S, and a particular scenario x is

$$p(y^*|S, x, D) = \int_S \int_\Theta p(y^*|\theta_S, S, x)\, p(\theta_S|S, D)\, p(S|x, D)\, d\theta_S\, dS,$$

and the **aggregate** predictive distribution $p(y^*|S, X, D)$ for y^* given D, S, and X is

$$p(y^*|S, X, D) = \int_X p(y^*|S, x, D)\, dx. \tag{2}$$

Here $p(y^*|\theta_S, S, x)$ is the conditional predictive distribution for y^* given specific choices for scenario, structure, parameters, and $p(\theta_S|S, D), p(S|x, D)$, and $p(x|D)$ are posterior distributions for parameters, structure, and scenario (respectively) given past data. Each of these posterior distributions depends on prior distributions in the usual Bayesian way.

2.1. **GESAMAC results.** We used Monte Carlo methods to approximate the relevant integrals presented above and to simulate in a way that fleshes out all four sources of uncertainty. With a given set of inputs `GTMCHEM` produces an entire time trace of predicted dose values as a function of t. It is of central scientific interest to study the variation within and between scenarios of the entire dose function $D(t)$.

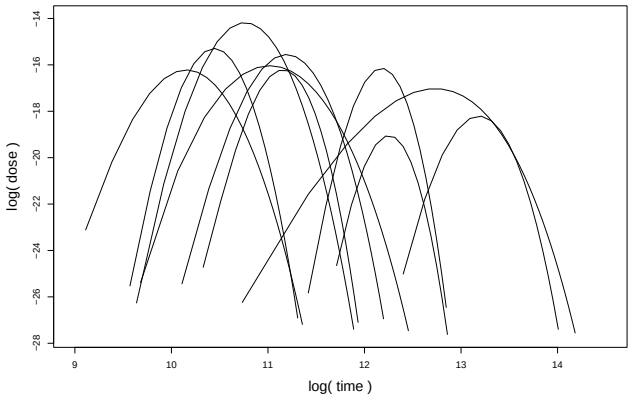

FIGURE 1. log(*Dose*) *against* log(*time*) *for I–129,* ***REF*** *scenario, 10 randomly chosen curves.*

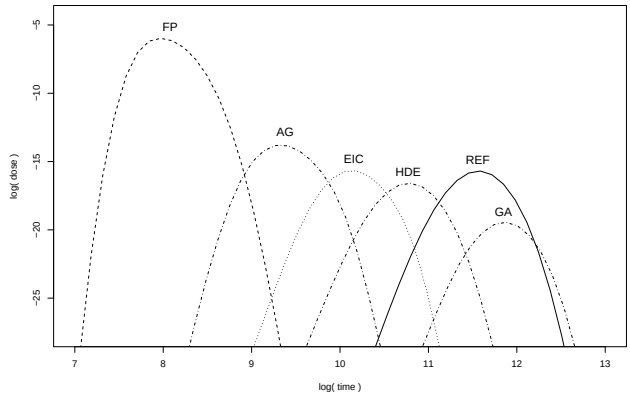

FIGURE 2. *Smoothed functional mean* log(*Dose*) *against* log(*time*) *for I–129, all six scenarios.*

3. Functional Data Analysis

One way to examine the entire dose curve $D_i(t)$ is through functional data analysis (FDA) [Ramsay and Silverman, 1997]. In this approach one starts by choosing a set of basis functions ϕ_k (e.g., Fourier, polynomial, B-splines, wavelets, and others) and representing $D_i(t)$ by a basis function expansion of the form

$$D_i(t) = \sum_{k=1}^{K} c_{ik}\, \phi_k(t). \tag{3}$$

This representation of $D_i(t)$ can then be used to perform the FDA analogues of principal components analysis, analysis of variance (within and between scenarios), and regression of log dose on the inputs (i.e., sensitivity analysis on the whole dose curve, not just its maximum, as is usually done). We used an ensemble of freeware FDA functions [Ramsay,1996] to perform all the relevant calculations.

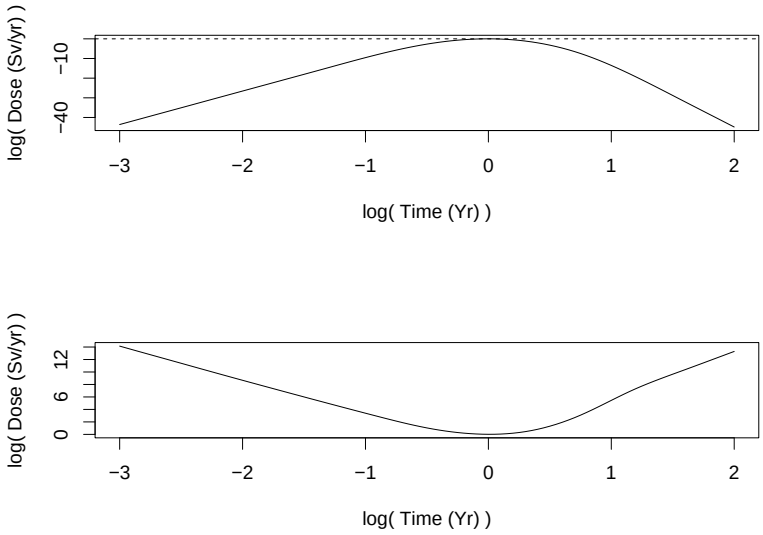

FIGURE 3. *Mean dose function and the standard deviation function, for the* `Ref` *scenario.*

Given a data matrix $\{x_{ij}\}$, where i indexes observations and j indexes variables, the idea underlying ordinary multivariate principal components analysis (PCA) is dimensionality reduction through construction of linear combinations or scores

$$f_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n. \tag{4}$$

For example, the first principal component is defined by the weight vector $\beta_1 = (\beta_{11}, \ldots, \beta_{p1})'$ for which the scores

$$f_{i1} = \sum_{j=1}^{p} \beta_{j1} x_{ij} \tag{5}$$

have the largest possible mean square $\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2$ subject to the constraint $\sum_{j=1}^{p} \beta_{j1}^2 = 1$, and further principal components maximize mean squares in directions orthogonal to the previous ones (still subject to norm constraints). This is an eigenanalysis problem involving extraction of eigenvalues and eigenvectors from the sample covariance matrix. This idea can be directly generalized to functions, with $D_i(t)$ replacing x_{ij}; the scores become

$$f_i = \int \beta(t) D_i(t)\, dt, \tag{6}$$

and the first principal component, or **harmonic**, is now a **weight function** $\beta_1(t)$ chosen to maximize $\frac{1}{n}\sum_{i=1}^{n} f_{i1}^2$ subject to the **constraint** $\int \beta_1(t)^2 = 1$. Ramsay and Silverman (1997) convert the continuous functional eigenanalysis problem into an approximately equivalent matrix eigenanalysis by discretizing the relevant functions.

In Figure 4 we present the results of the functional PCS for the Reference scenario (the analysis for all the other scenarios gave similar results). Harmonic 1 represents horizontal

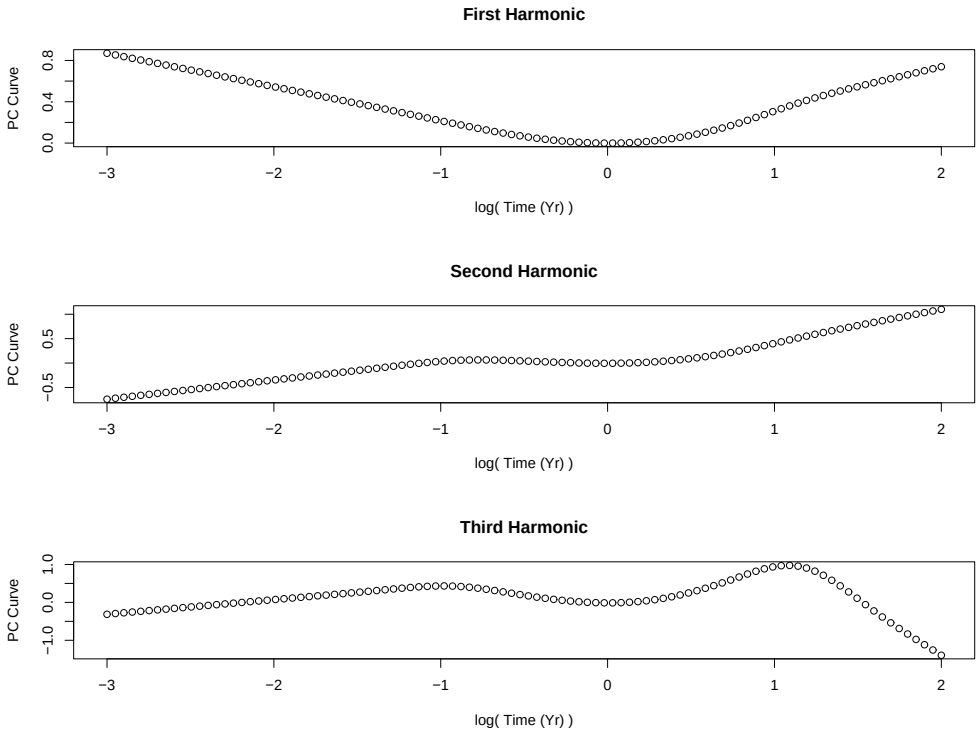

FIGURE 4. *The three first principal components for the* `Ref` *scenario.*

scaling in both directions around the mean shape, i.e., the standard deviation in the theoretical underlying approximate Gaussian; harmonic 2 represents horizontal scaling to the right of the max (asymmetry, which is not predicted by the Gaussian).

We fit a functional linear model to the data from all six scenarios with dummy variables for scenario membership; values in Figure 6 represent differences between each non-Reference scenario and the Reference scenario(functional equivalents of the scenario main effects in a one-way ANOVA), together with the Reference scenario as a baseline. The main observations are the fact that no matter what time point one choose there are always significant differences between each non-reference scenario and the reference scenario, this shows explicitly that assuming constance of present conditions in waste repositories could lead to significant errors in predicting the behavior of the system even if the geological conditions are altered only slightly (as represented by scenario EIC, for instance).

We also fit a functional regression model to each scenario data, regressing the log(dose) curves on the vectors of their inputs; Figure 7 shows the functional analogue of the regression coefficients. We observe that some variables (like containment time and the retention coefficient of geological layer 2) have little effect on the shape of the log(dose) curve; but the effects of other variables are difficult to interpret.

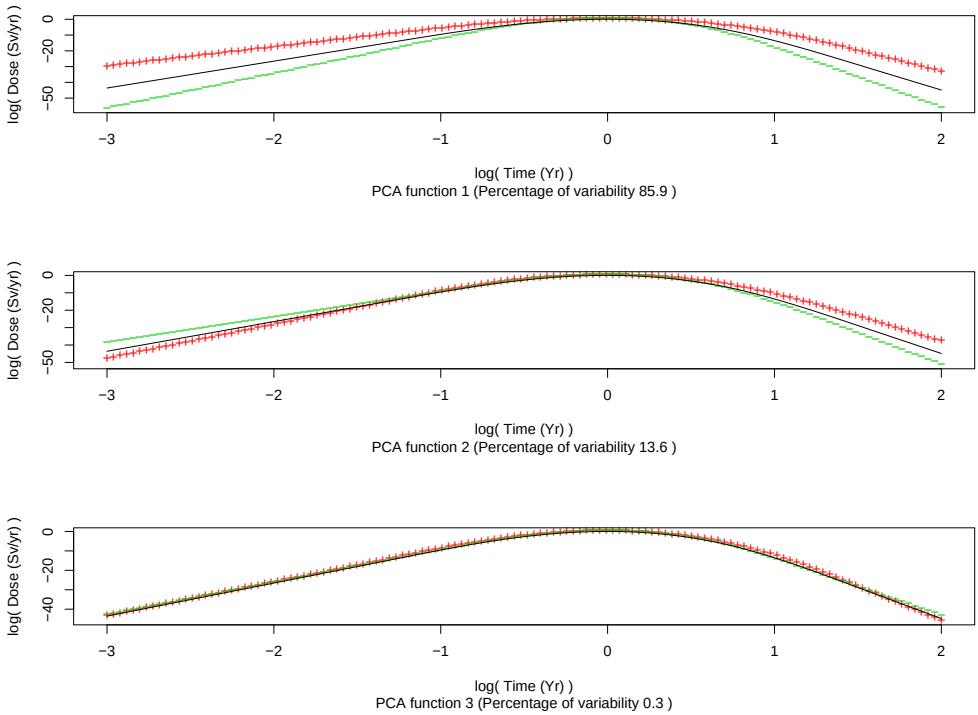

FIGURE 5. *Mean function and mean function plus/minus 20% of the first, second and third Principal Components, for the* `Ref` *scenario.*

REFERENCES

Prado et al, 1998. P. Prado, A. Saltelli and S. Eguilior,Level E/G, Test Case Proposal for GESAMAC,Centro de Investigaciones Energeticas Medioambientales y Tecnologicas- Instituto de Medio Ambiente CIEMAT-IMA, 1998, CIEMAT/DIAE/550/55900/04/98

Ramsay and Silverman, 1997. J. Ramsay and B. Silverman, Functional Data Analysis, Springer, 1997

Ramsay, 2001. J. Ramsay, Matlab and S-PLUS functions for functional data analysis, McGill University, 2001

Draper, 1995. D. Draper,Assessment and Propagation of Model Uncertainty, J. R. Statist. Soc. B, 1995, 57(1), 45-97

Draper et al, 1999. D. Draper, A. Pereira, P. Prado, A. Saltelli, R. Cheal, S. Eguilior, B. Mendes and S. Tarantola,Scenario and Parametric Uncertainty in GESAMAC: A Case Study in nuclear Waste Disposal Risk Assessment, Computer Physics Communications, 1999, 117, 142-155

Prado and Eguilior, 1996. P. Prado and S. Eguilior, GTMCHEM Computer Code- User Manual, CIEMAT, 1996, CIEMAT-IMA-550-55D18-96

Ramsay,1996. Jim Ramsay, 1996, MATLAB, R and S-PLUS Functions for Functional Data Analysis, `www.psych.mcgill.ca/faculty/ramsay/ramsay.html`

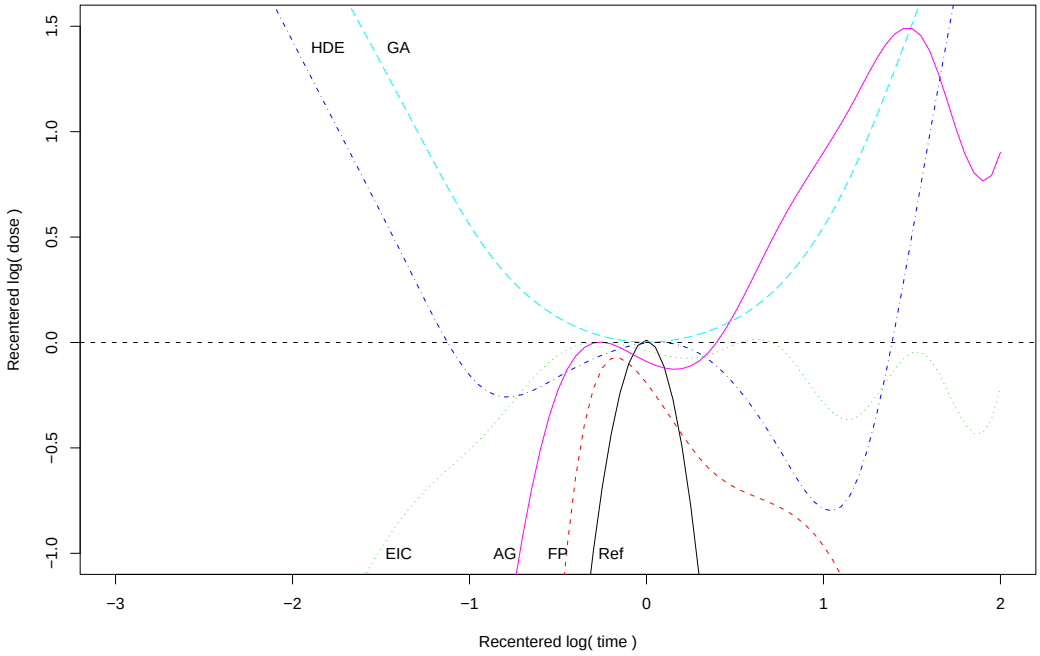

FIGURE 6. *Functional main effects for scenario, I–129 data.*

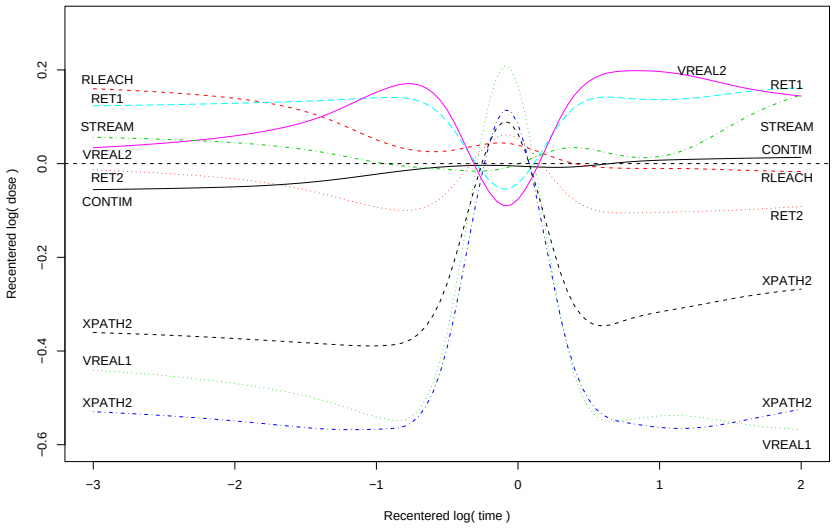

FIGURE 7. *Functional regression coefficients,* ***REF*** *scenario.*