Large-scale simultaneous hypothesis testing in monitoring carbon content from French soil database - A semi-parametric mixture approach

Didier Chauveau, Nicolas P.A. Saby, Thomas G. Orton, Blandine Lemercier, Christian Walter, Dominique Arrouays

Large-scale simultaneous hypothesis testing in monitoring carbon content from French soil database – A semi-parametric mixture approach

Didier Chauveau[1], Nicolas P.A. Saby[2], Thomas G. Orton[2], Blandine Lemercier[3], Christian Walter[3], Dominique Arrouays[2],

[1]Université d'Orléans and CNRS UMR 7349, France
[2]INRA, US1106 Unité Infosol, F-45000 Orléans, France
[3]UMR SAS, INRA, Agrocampus, Rennes, France

November 28, 2013

Abstract: Investigating the information of the French National Soil Tests database for soil monitoring produces multiple hypothesis testing problems with hundreds or thousands of test responses to consider simultaneously. A largely used concept of error control in such multiple testing is the expected proportion of falsely rejected hypotheses, or False Discovery Rate (FDR). A related notion of local FDR (ℓFDR) can be appropriately represented by considering that the observed p-values come from a two-components mixture model where the component corresponding to the null hypothesis is known. In this work, we explore different solutions for FDR estimation. In particular, we introduce a specific version of a semi-parametric Expectation-Maximization (EM) algorithm for ℓFDR estimation, and compare it to classical ℓFDR estimation using parametric mixtures, and conventional FDR approaches. The performances of the different models for estimating the FDR and related criteria are first illustrated on the results of simulated multiple comparison tests. These approaches are then applied to soil carbon content monitoring on our database. The results show that not taking into account the FDR estimation can lead to over-estimation of the number of *cantons* (locations) subject to a significant change. However, we have detected large numbers of significant changes in the database that occured during the time period of this study. Globally, losses in organic

carbon are observed in Northern France, along the Atlantic coastal regions, and to a less extend for the data collected over the North-Eastern regions. The OC increases are more scattered over the territory. We also use the data to estimate the minimum number of samples needed at each period to detect a given change.

Keywords: Carbon content, EM algorithms, FDR, finite mixture, France, semi-parametric mixtures, soil monitoring.

1 Introduction

Political awareness that soil is threatened by increasing pressures has been rising for several years (European Commission, 2002). Indeed, the demand for soil information is increasing continuously (Richer de Forges and Arrouays, 2010). Although rates of soil degradation are often slow and only detectable over long timescales, they are often irreversible. Therefore, monitoring soil quality is essential in order to detect adverse changes in their status at an early stage (Morvan et al., 2008). There are many generic issues that must addressed when establishing and operating such monitoring systems. Of particular importance is the requirement to detect change in soil over relevant spatial and temporal scales with adequate precision and statistical power (Arrouays et al., 2008; Desaules et al., 2010; van Wesemael et al., 2011; Arrouays et al., 2012). Analysing the results from existing soil measurement exercises, such as operational soil testing by farmers or fertiliser suppliers is one potential option for detecting large temporal trends in soil characteristics. Therefore, legacy soil data testing results have often been used to assess temporal changes at national and regional levels (eg., for phosphorus by Skinner and Todd (1998) in England and Wales, Cahoon and Ensign (2004) in eastern North Carolina (USA), Wheeler et al. (2004) in New Zealand, Lemercier et al. (2008) in France, Reijneveld et al. (2010) in The Netherlands; and for carbon, Saby et al. (2008) in France; Reijneveld et al. (2009) in The Netherlands). However, the conclusions drawn using these kinds of data may be subject to several sources of bias that are inherent in a non-controlled sampling strategy.

Since the exact locations of the carbon content observations are not known, the only geographical information available is the name of the town of the sampled location. A simple statistical framework consists then in gathering the observations per group of commune (called canton) and time period to perform inference tests. In this framework, the null hypothesis

could be either that the means or the medians of the two groups are equal. We assume here that the locations are distant enough, so that measurements can be considered as independent (i.e. we do not consider a spatial correlation in the present work). Hence, investigating the information of the French National Soil Tests database for soil monitoring produces multiple hypothesis testing problems with hundreds or thousands of test responses to consider simultaneously. In this multiple inference setup, the unguarded use of single-inference procedures results in an increased false positive rate among the simultaneous tests. The consideration of a global type I error level (say $\alpha = 5\%$) when several tests are performed simultaneously, has been considered first in the well-known context of analysis of variance (Anova), where several levels of a factor are compared two-by-two, leading to n simultaneous tests. It is easy to check that, if n independent tests with level of significance α are applied simultaneously, the achieved FamilyWise Error Rate (FWER), that is the probability of observing at least one false rejection among the n tests, is $1-(1-\alpha)^n$ which quickly increases with n and is already $\approx 99\%$ for $n = 100$. The historical approach since the early 1950s, called the Bonferroni approximation procedure (see e.g. Benjamini and Hochberg (1995) and references therein) consists in applying each test at a level α/n, resulting in a FWER lesser than α. However, Bonferroni-type procedures appear to be too conservative when n gets large because α/n gets too small, leading to too few rejections (as will be illustrated in Section 4). Since then, the point of view on the problem has changed, focusing in the number (or ratio) of erroneous rejections instead of the question wether any error was made, as for the FWER. In this vein, the most popular and largely used concept of error control in such multiple testing inference is nowaday the expected proportion of falsely rejected hypotheses among the simultaneous tests, or False Discovery Rate (FDR, Benjamini and Hochberg, 1995).

Several statistical algorithms have been proposed in the literature for estimating the FDR, the recent and unified procedure based on a nonparametric approach from Strimmer (2008) appearing to be one of the current standard for practitioners. An alternative approach to FDR estimation consists in viewing the problem as the statistical estimation of the parameters of a finite mixture model. Parametric mixtures for FDR estimation have been considered in, e.g., McLachlan et al. (2006), and semiparametric versions have also been considered recently, such as in Robin et al. (2007). The alternative approach that we propose here belongs to that last framework, but takes advantages of very recent developments of statistical algorithms for estimating semi- or nonparametric mixtures. All these standard and new

models will be detailed in Section 1.2. In particular, we implement and compare new ideas pertaining to recent semi- and non-parametric extensions of the well-known EM (Expectation-Maximization) algorithm commonly used for estimating parametric mixtures. Some specific ingredients coming from the multiple testing framework can be specified within the EM algorithm machinery.

When monitoring soil quality, the estimation and control of the FDR is of particular importance as false negative or false positive may lead to unwarranted recommendations or to irrelevant policies (see, e.g., Lark et al., 2007). Taking the example of organic carbon, such false positive or negative results may lead to errors in national carbon accounting, or to wrong identification of areas at risks where protection policies should be implemented. The basic motivations we have in this paper are thus twofold: (i) To improve analysis of soil carbon content monitoring, in particular by applying recent and new multiple testing procedures and FDR estimation and control, on datasets issued from the French National Soil Tests database. This application field is different than micro-array techniques for which the FDR is more commonly applied. For instance, the proportion of null hypothesis in the population is about 50%, smaller than what is usually observed in micro-array setup (e.g., 80% in the gene experiment example studied in Robin et al. (2007)). Our second objective is (ii) to compare recent methods for FDR estimation, such as the challenging nonparametric approach available in the fdrtool package (Strimmer, 2008), to approaches taking full advantage of an underlying mixture model assumed for the vector of p-values. Note that the alternative FDR control procedures that we develop here are implemented in the software package mixtools (Benaglia et al., 2009b) for the R statistical software (R Development Core Team, 2010), and are available in the last public version.

1.1 The French National Soil Tests Database

The results of the soil analyses compiled in the French National Soil Test Database are a consequence of requests from farmers and landowners for help to improve the management of their crops and pastures. Samples were taken from topsoil horizons of cultivated or pasture fields, but the specific reasons individual farmers requested soil analysis are not known and therefore the sampling strategy could not be controlled. In each sampled field, 10–15 subsamples of the ploughed layer (or the 0 to 30 cm topsoil layer in the case of pasture) were collected with a hand auger and the extracted soil mixed to provide a composite sample. Samples were sent to laboratories certified

by the French Ministry of Agriculture and the results from standardised analytical procedures were stored in the national database.

In the study area, 1,220,039 soil samples were analysed for OC content between 1990 and 2004. All samples were analysed by the wet chromic acid digestion method, utilising an excess amount of potassium dichromate with concentrated sulphuric acid at boiling-point. After the addition of concentrated phosphoric acid, the excess dichromate was back-titrated with ferrous ammonium sulphate. This method is known not to measure all the organic carbon present but only that which is easily oxidisable, and recovery factors have been recently studied by De Vos et al. (2007) who recommend the use of dry combustion. This was not possible because the study used historical data from samples which were not archived. The OC that is not recovered by wet chromic acid digestion is mainly composed of highly resistant compounds and thus our estimates of change are likely to be little affected by the use of this wet digestion. Moreover, in previous studies we showed that wet chromic acid digestion recovered on average 95% of the OC estimated by dry combustion on various French soils (Jolivet et al., 1998; Arrouays et al., 2001). The raw data are spatially aggregated at the cantonal level (in France, a *canton* is an administrative unit combining several municipalities). They are also aggregated per time period, 1990–1997 and 1998–2005. In previous work, Saby et al. (2008) applied a different approach regarding splitting the data over time by using 3 time periods. This was adapted to the smallest region under study, where more data were avaible than for the present Nationwide study. In this work, we then used only 2 periods to optimise the number of data per canton. We refer to these data as Carbon data in the following of this paper.

1.2 Mixture models in multiple testing FDR evaluation

In multiple testing, we observe a sample of p-values $\boldsymbol{p} = (p_1, \ldots, p_n)$, where each individual observation p_i corresponds to the critical probability of the ith test, for which either the null hypothesis H_0 is true (not significant or not interesting), or the alternative hypothesis H_1 is true (interesting i.e. rejection of H_0). Since it is not observed whether each hypothesis is true or false, we are in the general framework of statistical inference from missing data: the critical probabilities can be viewed as a sample of observations in $[0, 1]$ coming from a finite mixture model.

Finite mixtures give a flexible way to model a wide variety of random observations (see, e.g., McLachlan and Peel, 2000). In such models, we assume that n independent measurements $X_1, \ldots, X_n$ are observed such that each

X_i comes from one of m possible sub-populations, called *component* distributions. Importantly, the component index, 1 though m, is not observed along with X_i. It is common to associate to X_i the unobserved indicator vector $Z_i = (Z_{i1}, \ldots, Z_{im})$ where $Z_{ij} = 1$ if observation i comes from component j (hence Z_i have $(m-1)$ zero's and one 1). It is assumed that, unconditional on X_i, each $\mathbb{P}(Z_{ij} = 1) = \lambda_j$, the proportion of individuals from component j in the population, and that the probability density function (pdf) of an observation given that it belongs to component j is some density f_j. Thus, the pdf of an observation from the mixture is $g(x) = \sum_{j=1}^{m} \lambda_j f_j(x)$, where the λ_j's are strictly positive and sum to unity. The vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ represents the missing data associated to the n observations.

Until recently, only *parametric* mixture models have been considered and investigated, where parametric means that the f_j's come from a (generally single) parametric family indexed by some parameter ξ, i.e. $f_j(\cdot) \equiv f(\cdot\,; \xi_j)$, so that $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \xi_1, \ldots, \xi_m)$ is the (finite-dimensional) model parameter to be estimated from the data. Normal mixtures, where $f(\cdot\,; \xi_j)$ is the density of the Gaussian $\mathcal{N}(\mu_j, \sigma_j^2)$ with parameters $\xi_j = (\mu_j, \sigma_j^2)$, are the most commonly used in this category.

In multiple testing, we thus associate to the sample of p-values $\boldsymbol{p}$ the unobserved indicator variables $\boldsymbol{Z}$ where Z_i will be the unobserved indicator of the true hypothesis associated with the ith test. For ease of notation we will index the components by $\{0, 1\}$, according to the hypotheses, i.e. f_j is the pdf of the p-value conditional of H_j being true, $Z_i = (1, 0)$ if H_0 is true, and $Z_i = (0, 1)$ if H_1 is true, i.e. if H_0 is rejected (interesting ith test). Hence, p_i typically comes from a two-components mixture model with pdf

$$g_{\boldsymbol{\theta}}(p) = \lambda_0 f_0(p) + (1 - \lambda_0) f_1(p), \tag{1}$$

where $\boldsymbol{\theta} = (\lambda_0, f_0, f_1)$ is the model parameter. Theoretically, given that H_0 is true, the p-value is uniformly distributed over $[0, 1]$,

$$(p_i | \text{“not interesting”}) = (p_i | H_0 \text{ true}) = (p_i | Z_{i0} = 1) \sim \mathcal{U}_{[0,1]},$$

whichever test is actually performed, so that $f_0(p) = \mathbb{I}_{[0,1]}(p)$ is known and the model parameter reduces to $\boldsymbol{\theta} = (\lambda_0, f_1)$, which makes an important difference between usual mixture model estimation and the present multiple tests setting. The density of $(p_i | Z_{i1} = 1) \sim f_1$ may be parametrically specified, as in usual mixture modelling. In the multiple testing setup for instance, Beta distributions have been considered, leading to the Beta-Uniform mixture model (see, e.g. Robin et al., 2007). However, since often in practice

very little is known about the distribution under the alternative hypothesis, we propose here to relax the parametric assumption on f_1.

Mixing parametric and nonparametric estimates in the FDR context is not new (see Robin et al., 2007, and references therein). In other contexts, extensions of mixture models where the component densities (f_j's) are semi- or non-parametrically specified have been considered quite recently in the statistical literature (Hall and Zhou, 2003; Hall et al., 2005). Semi- and non-parametric models and associated EM-like estimation algorithms have recently been proposed, in Bordes et al. (2007) for the univariate case, Benaglia et al. (2009a) and Levine et al. (2012) for multivariate and fully nonparametric situations. We extend here these algorithmic ideas to the multiple testing estimation problem. For such a mixture with one component known, the theoretical question of statistical identifiability has been adressed in Bordes et al. (2006), where it is proved that identifiability is satisfied at least if f_1 is semi-parametrically specified, as will be detailed later on.

The two-components mixture model given by equation (1) is assumed, e.g., in Strimmer (2008) and Robin et al. (2007). However, one difference with our approach is that the fdrtool estimation method in Strimmer (2008) is grounded on an empirical (nonparametric) estimate of the model cumulative density function (cdf). It does not rely on the missing data aspect induced by the mixture, unlike any EM-based strategy does. In particular, an EM-like algorithm delivers, together with estimates of the mixture parameters (in this case (λ_0, f_1)), estimates of the *posterior probabilities* that each p-value comes from each component. In the present FDR context we can estimate directly the so-called local FDR (ℓFDR, Efron et al., 2001) from these posteriors

$$\ell\text{FDR}(p_i) = \mathbb{P}(\text{“not interesting”}|p_i) = \mathbb{P}(Z_{i0} = 1|p_i),$$

for p_i's leading to rejection of H_0. The FDR can then be computed as the average of the $\ell\text{FDR}(p_i)$'s over all the rejected p_i's. Robin et al (2007) approach is very close to the semi-parametric EM approach we present here, except that they estimate λ_0 separately, and then estimate f_1 using a weighted kernel density estimate (wkde). It appears that this wkde itself is much in the spirit of the density estimation step of our semi-parametric EM-like strategy, following Benaglia et al. (2009a).

Another difference is that these recent FDR control approaches have been compared, e.g. on Robin et al (2007) and Strimmer (2008), on synthetic p-values coming from possibly unrealistic mixtures, typically Beta-Uniform,

Uniform-Uniform, or Uniform-Truncated exponential distributions. The advantage being that in these cases, the pdf of p under H_1 is parametrically known in a simulation context. In an actual situation like in our setting considering the carbon data, the p-values come from two-sample tests that are of different types due to the underlying assumed distribution for the populations that have to be compared (soil carbon levels). Typically, when these can be assumed Gaussian, usual Student t-tests are performed, but when this Gaussian assumption is inappropriate, Mann-Whitney (MW) nonparametric (unpaired) tests are required. In addition, when rejection occurs, the distribution of the critical value under the alternative hypothesis depends on the actual difference (e.g., shift) between the two samples distributions, which is varying and unknown. Our approach relies on the assumption that the complexity of the pdf $(p|H_1 \text{ true}) \sim f_1$ is modelled by the flexibility brought by our nonparametric assumption for it, even though the pdf of $(p|H_0 \text{ true})$ remains $\mathcal{U}_{[0,1]}$.

Finally, note that it is common to consider the probit transform of the p-values since the known component pdf f_0 simply becomes then the density of the standard normal $\mathcal{N}(0,1)$. This is also a good choice in our case for technical reasons since one of our algorithms require kernel density estimates that behave better for observations on the real line instead of $[0,1]$. In the present paper we will in practice consider the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$ where $x_i = \text{probit}(p_i)$ for $i = 1, \ldots, n$.

1.3 EM algorithms

The EM algorithm, as defined in the seminal paper of Dempster et al. (1977), is more properly understood to be a class of algorithms, a number of which predate even the Dempster et al. (1977) paper in the literature. These algorithms are designed for maximum likelihood estimation in missing data problems, of which finite mixture problems are canonical examples because the unobserved Z_{ij}'s give an easy interpretation of missing data. Intuitively, EM algorithms replace the unfeasible maximization of the likelihood of the observed data by the maximization of a pseudo likelihood that resembles the likelihood of the complete data, which is itself easy to maximize. For a comprehensive and recent account of EM algorithms, we refer to McLachlan and Krishnan (2008); here we only briefly describe the finite mixture case.

We consider the *complete data* $((x_1, Z_1), \ldots, (x_n, Z_n)) = (\boldsymbol{x}, \boldsymbol{Z})$ associated with the actually observed sample denoted generically by $\boldsymbol{x}$ (which, for us, represents the vector of probit-transformed p-values from the n statistical tests). In parametric mixture models, the complete data log-likelihood

function is easy to maximize over $\boldsymbol{\theta}$. The EM algorithm takes advantage of this by iteratively maximizing instead (since $\boldsymbol{Z}$ is not observed), its expectation conditionnally on $\boldsymbol{x}$, under the assumption that the parameter driving the random behavior of $\boldsymbol{Z}$ at iteration t is the current value $\theta^{(t)}$. Computing this expectation usually denoted $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, reduces to the computation of the already introduced *posterior probabilities*

$$P_{ij}^{(t)} = \mathbb{P}_{\boldsymbol{\theta}^{(t)}}\left(Z_{ij} = 1 \,|\, X_i = x_i\right) = \frac{\lambda_j^{(t)} f(x_i; \xi_j^{(t)})}{\sum_{j'=1}^{m} \lambda_{j'}^{(t)} f(x_i; \xi_{j'}^{(t)})}, \tag{2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots m$. Each $P_{ij}^{(t)}$ is the posterior probability of individual i coming from component j, given the current $\boldsymbol{\theta}^{(t)}$ and the observed data x_i. The EM iteration $\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta}^{(t+1)}$ is defined by

1. `E-step:` `compute` $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ `i.e. compute the` $P_{ij}^{(t)}$`'s by (2)`
2. `M-step:` `set` $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Conveniently, the M-step for finite mixture models always looks partly the same: whatever the assumption on the f_j's, the updates to the mixing proportions are given by

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} P_{ij}^{(t)}, \quad j = 1, \ldots, m. \tag{3}$$

The updates for the f_j or the $\boldsymbol{\xi}_j$ parameters depend on the particular form of the component densities and are easy for, e.g., Gaussian mixtures. These will be precised below for our models.

2 Two mixture models for FDR estimation

2.1 Gaussian mixture with one component known

This first model that we consider is a 2-component parametric Gaussian mixture with component f_0 forced to $\mathcal{N}(0, 1)$, which is a is a slight difference with the Gaussian EM fit from McLachlan et al. (2006). The mixture pdf is then

$$g_{\boldsymbol{\theta}}(x) = \lambda_0 f_0(x) + (1 - \lambda_0)\phi(x; \mu_1, \sigma_1^2), \tag{4}$$

where $\phi(\,\cdot\,; \mu_1, \sigma_1^2)$ denotes here the pdf of the Gaussian $\mathcal{N}(\mu_1, \sigma_1^2)$, and the (parametric) model parameter is $\boldsymbol{\theta} = (\lambda_0, \mu_1, \sigma_1^2)$.

The standard EM algorithm for Gaussian mixtures is straightforward to implement (see, e.g., McLachlan and Krishnan, 2008) and we do not give it here for brevity. The specific version of this algorithm for model (4) turns out to be a very simple case of more general EM-type algorithms for mixtures with constrained parameters, investigated recently by Chauveau and Hunter (2013). Of course constraining here $\mu_0 = 0$ and $\sigma_0^2 = 1$ is simple enough that it does not really require these far more general extensions. We only mention it because these new algorithms are already coded in the current version of the mixtools package (Benaglia et al., 2009b), and we use it here to run model (4) as a "parametric benchmark" in sections 3 and 4.

2.2 Semiparametric mixture with one component known

The purpose of this model is, while keeping f_0 known to be $\mathcal{N}(0,1)$ as before, to relax the parametric assumption on f_1 as explained in Section 1.2. We rename f_1 simply f here, and assume that it is an even ($f(x) = f(-x)$) but otherwise unspecified density. With such partially nonparametric models, the question of identifiability arises. Basically, this statistical property insures that different values of the parameter must generate different probability distributions of the observable variables. It turns out that identifiability requires here an additional location-shift scalar parameter μ (Bordes et al., 2006). The model pdf becomes then

$$g_{\boldsymbol{\theta}}(x) = \lambda_0 f_0(x) + (1-\lambda_0) f(x-\mu), \tag{5}$$

with semi-parametric model parameter $\boldsymbol{\theta} = (\lambda_0, \mu, f)$. This identifiable mixture model may be viewed as a special case of the model for which Benaglia et al. (2009a, section 4.3) have proposed a semi-parametric EM-like algorithm (spEM), that was purely methodological and was called "EM-like" because it does not satisfy a provable convergence property like a genuine EM. We do not get here into details about this methodology, but this original algorithm happened to be very similar to a recent modified version that relies on more solid theoretical grounds, see Chauveau et al. (2011). The algorithm we propose here for our semi-parametric FDR framework is in addition specific for two reasons. First, it requires constraining f_0 to the normal density, and second, to adapt the M-step for the parametric part (λ_0, μ) and the nonparametric part f for which a weighted kernel density estimate (wkde) is combined with a symmetrization step:

spEM algorithm with one component known Let K be a kernel density function, and h a bandwidth chosen by the user or data-driven (see

Section 3). For a current value $\boldsymbol{\theta}^{(t)} = (\lambda_0^{(t)}, \mu^{(t)}, f^{(t)})$ of the parameter, the iteration $\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta}^{(t+1)}$ is as follows:

`E-step:` compute, for $i = 1, \ldots, n$

$$P_{i0}^{(t)} = \frac{\lambda_0^{(t)}}{\lambda_0^{(t)} f_0(x_i) + (1 - \lambda_0^{(t)}) f(x_i - \mu^{(t)})}, \quad P_{i1}^{(t)} = 1 - P_{i0}^{(t)}$$

`M-step 1:` parametric part

$$\lambda_0^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} P_{i0}^{(t)}, \quad \mu^{(t+1)} = \frac{\sum_{i=1}^{n} P_{i1}^{(t)} x_i}{\sum_{i=1}^{n} P_{i1}^{(t)}} \tag{6}$$

`M-step 2:` nonparametric density estimation

$$\hat{f}(u) = \frac{1}{nh} \sum_{i=1}^{n} P_{i1}^{(t)} K\left(\frac{u - x_i + \mu^{(t+1)}}{h}\right) \tag{7}$$

and symmetrize, i.e. set $f^{t+1}(u) = (\hat{f}(u) + \hat{f}(-u))/2$.

3 A simulation study

We illustrate the application of the two mixture models presented in Section 2, fitted by these specific EM and spEM algorithms, together with the fdrtool method (Strimmer, 2008). We compare the approaches on synthetic data simulated from known models, intended to represent a similar set-up to that presented by the actual Carbon data.

3.1 Methodology

We have simulated a multiple testing procedure with n two-samples comparison tests. Let us denote for a single case these two samples by $(Y_1^k, \ldots, Y_{n_k}^k)$ i.i.d.$\sim$ F^k, for $k = 1, 2$. For not interesting cases (H_0 true), $F^1 = F^2$, whereas for interesting cases we have simulated F^2 as the distribution F^1 shifted by some $\delta > 0$. For 36% of the test cases, parametric Student t-tests on Gaussian populations (F^1 normal) were simulated. For the remaining 64% cases, F^1 was set to a Student $t(5)$ heavy-tailed distribution, and unpaired nonparametric MW tests were performed. These 36% and 64% proportions correspond approximately to the observed proportions in the carbon data. Each comparison has been based on two small samples

($n_1 = 5$ and $n_2 = 7$) where, for interesting cases, the distribution of the second sample was shifted by $\delta = 3$ so that the tests were not too powerful, i.e. rejection of H_0 was not too obvious. We also tried random shifted values δ.

We ran $R = 500$ replications of $n = 2000$ such tests, with the proportion of true H_0 set to $\lambda_0 = 0.4$ and $\lambda_0 = 0.6$, i.e. much larger than in usual microarray experiments, but corresponding roughly to the possible observed proportion in the carbon data. Both EM and semi-parametric EM need an initial parameter value for their first E-step. It is a common practice to provide initial centers from which a binning of the data is performed using a K-means clustering algorithm. Since the *probit* transform results in normally-distributed x values for not interesting cases, and negative values for the small p-values corresponding to interesting cases, we set these initial centers to $(0, -3)$ for both algorithms, where the value -3 comes from the leftmost peak of the typical histogram of $\boldsymbol{x}$ we obtained.

For comparing the methods on replications, we focused on two criterion motivated by the end user concern: proportion of interesting/not interesting cases and control of the FDR.

Error on the proportion λ_0: The estimation of the proportion λ_0 of not interesting cases (H_0 true) estimated by each method on the rth replication is denoted $\hat{\lambda}_0^{(r)}$. We can then evaluate the empirical Mean Square Error for each method by

$$MSE(\lambda_0) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\lambda}_0^{(r)} - \lambda_0)^2.$$

Error on the target FDR level The error on the target level α which is supposed to be achieved by an FDR control procedure may also be evaluated in this simulation context. At replication r, let $\boldsymbol{p}^{(r)} = (p_1^{(r)} \leq \cdots \leq p_n^{(r)})$ be the n ordered p-values, and $\widehat{FDR}(p)$ be the FDR estimated by one of the competing methods for $p \in [0, 1]$. In practice, this estimate is used to define how many of the smallest observed p-value have to be rejected, in order to achieve an estimated error level smaller than α. Using the ordered p-values this number of rejections can be defined by the index

$$\hat{d}_\alpha^{(r)} = \hat{d}_\alpha := \max\left\{i \in \{1, \ldots, n\} : \widehat{FDR}\big(p_i^{(r)}\big) \leq \alpha\right\}, \tag{8}$$

where we will drop the superscript r below to avoid redundancy in some notations. This index corresponds to the largest ordered p-value before which the estimated FDR crosses the level α for the "last time" (note that

when this estimated $FDR(p)$ is an increasing function of p, (8) is just the index of the p-value after which the estimated FDR becomes greater than α). Now in our simulation setup, since we observe the complete data $(\boldsymbol{p}, \boldsymbol{Z})^{(r)}$, we can compute the true FDR at each replication,

$$FDR(p_i^{(r)}) = \frac{1}{i} \sum_{\ell=1}^{i} \mathbb{I}_{Z_{\ell 0}^{(r)}=1} \,,$$

from which we can evaluate the *actual* error $FDR(p_{\hat{d}_\alpha}^{(r)})$ achieved when the $\hat{d}_\alpha$ smallest p-values are rejected at replication r, so that

$$\Delta(\alpha) = \frac{1}{R} \sum_{r=1}^{R} \left(FDR\left(p_{\hat{d}_\alpha}^{(r)}\right) - \alpha \right)^2$$

can be viewed as a MSE on the target level α over all the replications.

3.2 Results

An illustrative output from a single replication result of our Monte-Carlo experiment is provided in Fig. 1, which displays the behavior of the three FDR estimates we compare, together with the true FDR, BH and raw-p-value approaches. The number of rejections for each method based on definition (8) is given in the legend of Fig. 1, together with the value from the true FDR. We can see that for this particular replication, fdrtool and spEM suggest similar numbers of rejections $\hat{d}_\alpha = 793$ and $\hat{d}_\alpha = 798$ for the target level $\alpha = 10\%$, both slightly conservative since the true FDR available on these simulated data leads to 817 rejections. This figure also shows graphically how these quantities $\hat{d}_\alpha$ and the error on α depend only on the p-values $\boldsymbol{p}$, the estimated FDR and the level α. The Benjamini-Hochberg (BH) procedure suggests $\hat{d}_\alpha = 680$ rejects, which results in an actual error $FDR(p_{680}) = 3.82\%$, that is too conservative. The crude approach based on the uncorrected p-values rejects the $\hat{d}_\alpha = 860$ smallest p-values, that is non-conservative (as expected) with an actual error $FDR(p_{860}) = 11.74\%$, larger than the desired $\alpha = 10\%$.

Global results of the experiment for $\lambda_0 = 0.6$ are displayed in Fig. 2 and Table 1. The semi-parametric EM (spEM) with component 1 constrained to the normal $\mathcal{N}(0, 1)$ returns better estimates than the Gaussian EM with the same constraint (which is not surprising), but also more precise estimates in terms of bias and variance, for both λ_0 and the target FDR of 10%, than fdrtool. We can see in Table 1 and the right-hand pane of Fig. 2 that

spEM is on average less conservative for the target level α than fdrtool, and considerably better than BH or the Gaussian EM. We observed similar results in our second experiment, based on $\lambda_0 = 0.4$ (results not presented for conciseness).

We also tried different experiments allowing the distribution F^2 for interesting cases to be shifted by a random variable uniformly distributed. This results in a more complicated component density for interesting cases, rendering it more difficult to recover. When the shift is not too small, e.g., $\delta \sim \mathcal{U}_{[3,5]}$, results are similar to those in Fig. 2. However, this aspect warrants further investigation since for other random shift settings such as $\delta \sim \mathcal{U}_{[1,5]}$ we cannot tell which method is preferable between spEM and fdrtool.

4 FDR estimation for the carbon data

In this section, we detail the numerical results on the carbon data. We applied our two mixture models together with the fdrtool method and the Bonferroni and BH (Benjamini and Hochberg, 1995) historical procedures that are expected to be too conservative. Numerically, the procedures are identical to the one used in Section 3.1, with respect to the initialization of the EM and spEM algorithms from an empirical binning of the data.

Figure 3 displays the estimated $FDR(\cdot)$'s together with their associated number of rejected smallest p-values $\hat{d}_\alpha$ as in Fig. 1, except that here we do not know the truth and of course cannot compute the error with respect to the level α. The range of the numbers of rejected hypotheses for a ten percent level α (i.e. less than ten percent of false positives within rejected H_0 among all the $n = 2714$ tests performed) is somewhat important. Indeed, the Bonferroni procedure suggests 317 rejections among the 2714 tests performed while the fdrtool method suggests 1339 rejections. Moreover, we should reject 1312 null hypothesis on the raw p-values, i.e. without controlling the FDR, 1041 with BH and 1144 with the spEM algorithm.

It is somewhat surprising that fdrtool and spEM seem to behave differently here than in the Monte-Carlo experiment of Section 3, where fdrtool appeared slightly more conservative than spEM. Moreover, the (normally non-conservative) raw p-values approach is here slightly more conservative than fdrtool with 1312 rejects. This comparison with the raw p-values suggests that the estimation given by fdrtool should be taken with caution for these particular data, that may be due to a more complex density of p under the alternative hypothesis.

These results show that not taking into account the FDR estimation can

lead to over-estimation of the number of cantons subjected to a significant change (for both increase or decrease). For instance, the difference between the spEM method and more classical one (Saby et al., 2008; Lemercier et al., 2008) concerns 168 cantons, that is 12.8 % of the area. This over-estimation represents about 16 % for the canton showing an OC increase, and 10% for an OC decrease. Given the need to accurately quantify SOC change for national carbon accounting, this difference is of great importance.

Fig 4 maps the results of three of the six procedures tested: raw p-values, fdrtool and spEM procedures. The Bonferroni and BH (Benjamini and Hochberg, 1995) historical procedures and gaussEM are not plotted. The spatial structures of the area where the tests suggest significant changes in the carbon content are quite similar on the three maps. This suggests that the spatially isolated cantons where significant changes are detected in the database are more expected to show a false positive test. *Globally, losses in organic carbon are observed in Northern France, along the Atlantic coastal regions, and to a lesser extent in the North-Eastern regions. The OC increases are more scattered over the territory.*

We also tried to estimate the minimum sample size required at each campaign in order to have a reasonable probability of detecting a given change in the carbon level of a canton between two time periods. For each canton $i = 1, \ldots, n$ here, denote by D_i the relative difference between the medians of the two samples (between the two time periods), and by S_i the smallest size of the corresponding two samples. We tried to fit an Analysis of Covariance model for the predictor D, response variable S, and a categorical factor with two levels, "evolution" or "no evolution" based on the decision rule at a 10% level of the spEM estimate of the FDR. For the numerical predictor, a linear regression model of a log transform, $\log(S_i) = \alpha \log(D_i) + \beta$ seemed appropriate. Fig. 5 gives (on the linear scale) the upper prediction intervals for typical levels 50%, ..., 90%, for the level "evolution" of the categorical predictor. As a matter of example, if we apply the mean yearly decline of 0.6% found by Bellamy et al. (2005) in England and wales over a 25 yr period (1978-2003), the cumulative relative difference between 1978 and 2003 is about 14% and therefore the 50% or 90% upper prediction interval of the sample sizes needed to detect the changes are 190 and 470, respectively.

5 Discussion

We applied a large scale Monte-Carlo experiment to compare the recent and standard fdrtool method with mixture models, and in particular the new

spEM approach. We have shown that, at least in some cases, spEM performs better in terms of FDR precision and number of rejected null hypotheses. This suggests that in real applications, it would be sensible to at least test both methods to see whether or not they agree. Indeed, for the carbon data, spEM and fdrtool did not agree. The somewhat bizarre results given by fdrtool (less conservative than the raw p-values themselves) can be viewed as an argument to claim that our spEM approach may be more robust in such a case based on a complicated dataset (this complexity may be induced here by p-values coming from two kinds of tests).

Some technical computational details deserves further work. In particular the bandwidth choice in the kernel density estimate involved within the spEM algorithm has been done using Silverman's rule (R default) on the leftmost part of the sample of the probit transform of the p-values, since the values $y < c$ for some $c < 0$ are mostly associated with the interesting cases (rejections). The tuning of c itself, empirically done here, may be improved.

In this paper, we have detected large areas in France where significant changes of carbon contents occurred during the fifteen years period of this study. Differences in soil carbon content were observed between our two sampling periods. However, the sampling design used does not allow to draw statistical conclusions Nationwide. Indeed, soil carbon changes have already been shown by various studies using numerous paired samples taken exactly at the same location, which is not the case in our study (e.g., Bellamy et al., 2005). Therefore, some effective changes might not have been detected using our dataset. Our conclusions might be affected by a bias inherent in the unsupervised sampling strategy associated with the French soil test database: farmers' intentions towards C testing could have changed over time. Moreover, sampling practices may have changed over time (e.g., sampling depth, numbers of subsamples, etc). The spatial and temporal trends described from the national soil test database should be compared with data from two other French soil survey programmes: first, the French Soil Monitoring Network (RMQS) Arrouays et al. (2002), to confirm these trends and assess biases, and second, the inventory programme (IGCS) to take into account the pedological context of soil test values. Indeed, the RMQS grid is more appropriate to a statistical assessment of changes over the national French territory, however it will take at least 10 years before the second sampling campaing is achieved. The IGCS data might be a useful tool since it includes older data (from the 1950's up to now) but it needs to address spatio-temporal correlations effects as the sampling date is related to soil mapping programmes progresses and therefore, the age of data from a given area may be very different to another one.

References

Arrouays, D., Deslais, W., and Badeau, V. (2001). The carbon content of topsoil and its geographical distribution in france. *Soil Use and Management*, 17:7–11.

Arrouays, D., Jolivet, C., Boulonne, L., Bodineau, G., Saby, N., and Grolleau, E. (2002). A new initiative in France: a multi-institutional soil quality monitoring network. *Comptes Rendus de l'Académie d'Agriculture de France*, 88(5):93–105.

Arrouays, D., Marchant, B. P., Saby, N. P. A., Meersmans, J., Orton, T. G., Martin, M. P., Bellamy, P. H., Lark, R. M., and Kibblewhite, M. (2012). Generic issues on broad-scale soil monitoring schemes: A review. *Pedosphere*, 22(4):456–469.

Arrouays, D., Morvan, X., Saby, N. P. A., Richer de Forges, A., Le Bas, C., Bellamy, P. H., Berengi Uveges, J., Freudenschuss, A., Jones, R. J. A., Kibblewhite, M. G., Simota, C., Verdoodt, A., and Verheijen, F. G. A. (2008). Environmental Assessment of Soil for Monitoring: Volume IIa Inventory & Monitoring. Technical Report EUR 23490 EN/2A, Office for the Official Publications of the European Communities, Luxembourg.

Bellamy, P. H., Loveland, P. J., Bradley, R. I., Lark, R. M., and Kirk, G. J. D. (2005). Carbon losses from all soils across England and Wales 1978-2003. *Nature*, 437:245–248.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51(11):5429–5443.

Bordes, L., Delmas, C., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statistics*, 33:733–752.

Cahoon, L. B. and Ensign, S. H. (2004). Spatial and temporal variability in excessive soil phosphorus levels in eastern North Carolina. *Nutrient Cycling Agroecosyst*, 69:111–125.

Chauveau, D. and Hunter, D. R. (2013). ECM and MM algorithm for mixtures with constrained parameters. Technical Report hal-00625285, version 2, HAL.

Chauveau, D., Hunter, D. R., and Levine, M. (2011). Estimation for conditional independence multivariate finite mixture models. Technical Report hal-00558834, version 1, HAL.

De Vos, B., Lettens, S., Muys, B., and Deckers, J. (2007). Walkley-black analysis of forest soil organic carbon: recovery, limitations and uncertainty. *Soil Use and Management*, 23(3):221–229.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Desaules, A., Ammann, S., and Schwab, P. (2010). Advances in long-term soil-pollution monitoring of Switzerland. *Journal of Plant Nutrition and Soil Science*, 173(4):525–535.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

European Commission (2002). Communication of 16 April 2002 from the Commission to the Council, the European Parliament, the Economic and Social Committee and the committee of the Regions: Towards a Thematic Strategy for Soil Protection. COM (2006)231.

Hall, P., Neeman, A., Pakyari, R., and Elmore, R. T. (2005). Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678.

Hall, P. and Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224.

Jolivet, C., Arrouays, D., and Bernoux, M. (1998). Comparison between analytical methods for organic carbon and organic matter determination in sandy spodosols of france. *Communications in Soil Science and Plant Analysis*, 29:2227–2233.

Lark, R. M., Bishop, T. F. A., and Webster, R. (2007). Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma*, 138:65–78.

Lemercier, B., Gaudin, L., Walter, C., Aurousseau, P., Arrouays, D., Schvartz, C., Saby, N. P. A., Follain, S., and Abrassart, J. (2008). Soil phosphorus monitoring at the regional level by means of a soil test database. *Soil Use And Management*, 24(2):131–138.

Levine, M., Hunter, D. R., and Chauveau, D. (2012). Maximum smoothed likelihood for multivariate maximum. *Biometrika*, 98(2):403–416.

McLachlan, G., Bean, R., and Jones, B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Morvan, X., Saby, N. P. A., Arrouays, D., Le Bas, C., Jones, R. J. A., Verheijen, F. G. A., Bellamy, P. H., Stephens, M., and Kibblewhite, M. G. (2008). Soil monitoring in Europe: A review of existing systems and requirements for harmonisation. *Science Of The Total Environment*, 391(1):1–12.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reijneveld, A., van Wensem, J., and Oenema, O. (2009). Soil organic carbon contents of agricultural land in the Netherlands between 1984 and 2004. *Geoderma*, 152(3-4):231. 0016-7061 doi: DOI: 10.1016/j.geoderma.2009.06.007.

Reijneveld, J., Ehlert, P., Termorshuizen, A., and Onema, O. (2010). Changes in soil phosphorus status of agricultural land in the Netherlands during the 20th century. *Soil Use and Management*, 26(27-64).

Richer de Forges, A. C. and Arrouays (2010). Analysis of requests for information and data from a national soil data centre in france. *Soil Use and Management*, 26:374–378.

Robin, S., Bar-Hen, A., Daudin, J.-J., and Pierre, L. (2007). A semiparametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics and Data Analysis*, 51.

Saby, N. P. A., Arrouays, D., Antoni, V., Lemercier, B., Follain, S., Walter, C., and Schvartz, C. (2008). Changes in soil organic carbon in a mountainous French region, 1990-2004. *Soil Use and Management*, 24(3):254–262.

Skinner, R. J. and Todd, A. D. (1998). Twenty-five years of monitoring pH and nutrient status of soils in England and Wales. *Soil Use and Management*, 14(3):162–169.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9.

van Wesemael, B., Paustian, K., Andren, O., Cerri, C. E. P., Dodd, M., Etchevers, J., Goidts, E., Grace, P., Katterer, T., McConkey, B. G., Ogle, S., Pan, G., and Siebner, C. (2011). How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO(2) fluxes in agricultural soils? *Plant and Soil*, 338(1-2):247–259.

Wheeler, D., Sparling, G., and Roberts, A. (2004). Trends in some soil test data over a 14-year period in New Zealand. *New Zealand Journal of Agricultural Research*, 47:155–166.

	spEM	Gaussian EM	fdrtool
$\bar{\lambda}_0$	0.603	0.64	0.609
$MSE(\lambda_0)$	0.000149	0.00341	0.000307
$\bar{\alpha}$	0.0928	0.0424	0.0868
$\Delta(\alpha)$	0.000124	0.00353	0.000350

Table 1: Mean and MSE's for λ_0 and α from 500 replications of $n = 2000$ tests, true $\lambda_0 = 0.6$.

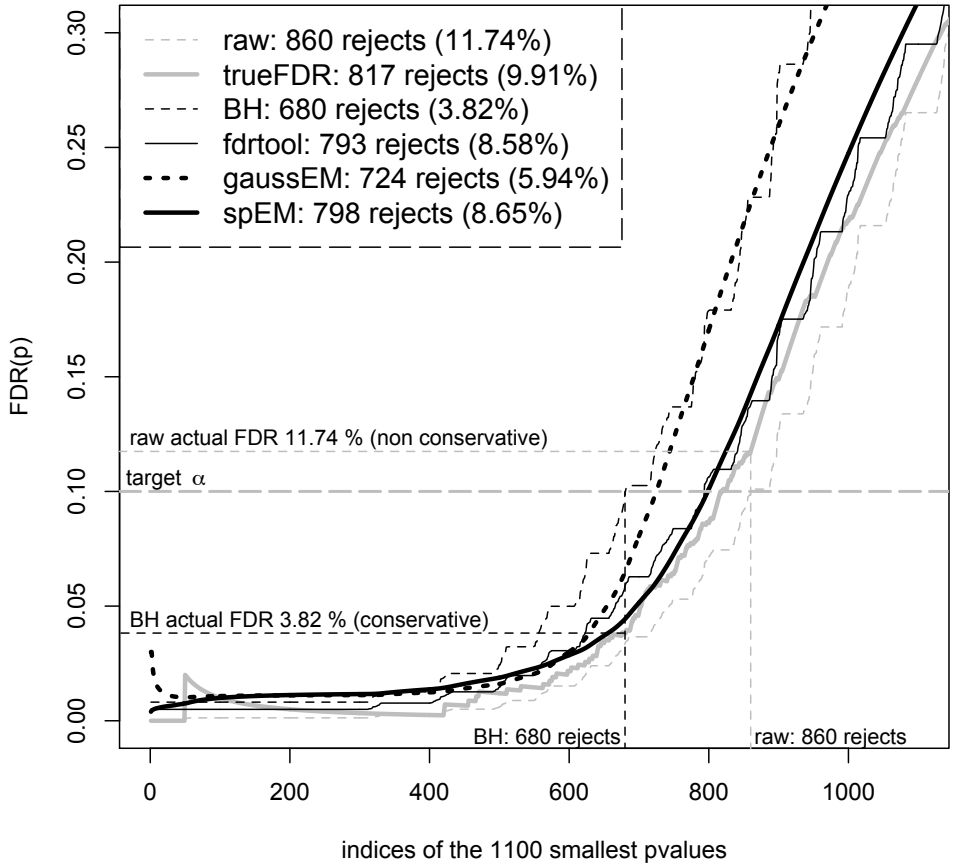

Figure 1: *Example of $FDR(p)$ estimates from five methods, for $n = 2000$ tests as detailed in Section 3.1. The number of rejections $\hat{d}_\alpha$ defined in* (8) *for the target level $\alpha = 0.1$, and actual error achieved by each method are given in the legend. Graphical illustration of computation of $\hat{d}_\alpha$ and error on α are given for the Benjamini-Hochberg (BH) conservative procedure (black dashed) and the non conservative approach using the raw p-values (grey dashed). For each method, $\hat{d}_\alpha$ corresponds to the index (vertical line) located where $\widehat{FDR}(\cdot)$ crosses the level α. Then the level of the horizontal line drawn from each true $FDR(p_{\hat{d}_\alpha})$ indicates the actual FDR achieved.*

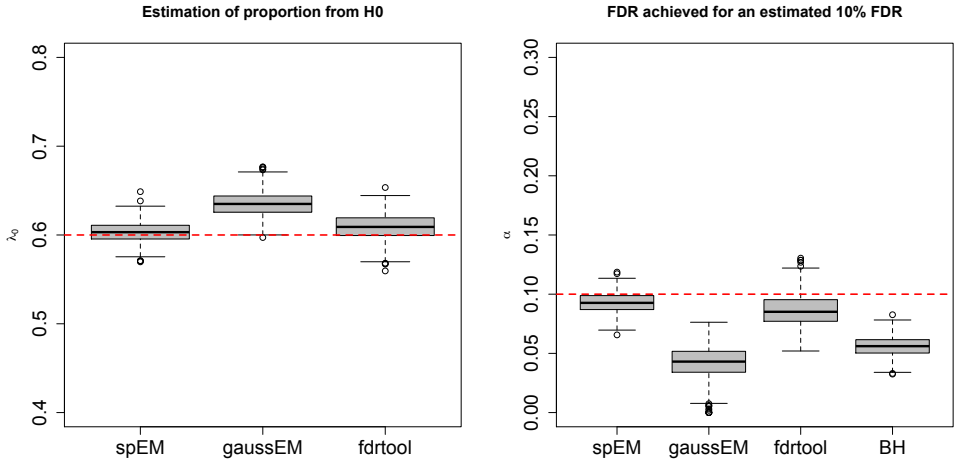

Figure 2: *Results from* $R = 500$ *replications of* $n = 2000$ *multiple tests, among which 36% are t-tests and 64% MW tests, for* $\lambda_0 = 0.6$ *and constant shift under* H_1*. Left: Boxplots of* λ_0 *estimates; Right: actual FDR after rejection of* $\hat{p}_\alpha^{(r)}$ *smallest p-vlues, for the three methods and the Benjamini-Hochberg procedure (BH). Horizontal dashed line are true values.*

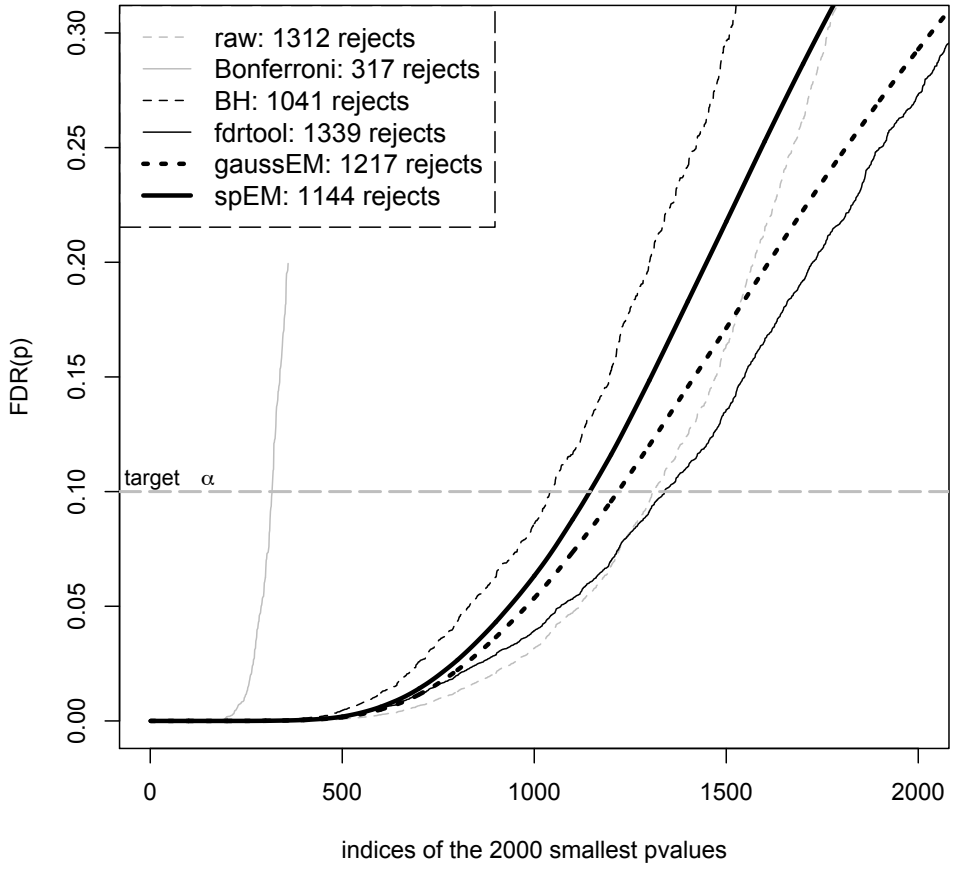

Figure 3: *FDR estimates from six approaches for the carbon data. Bonferroni and BH (Benjamini and Hochberg, 1995) historical procedures are also displayed for comparison.*

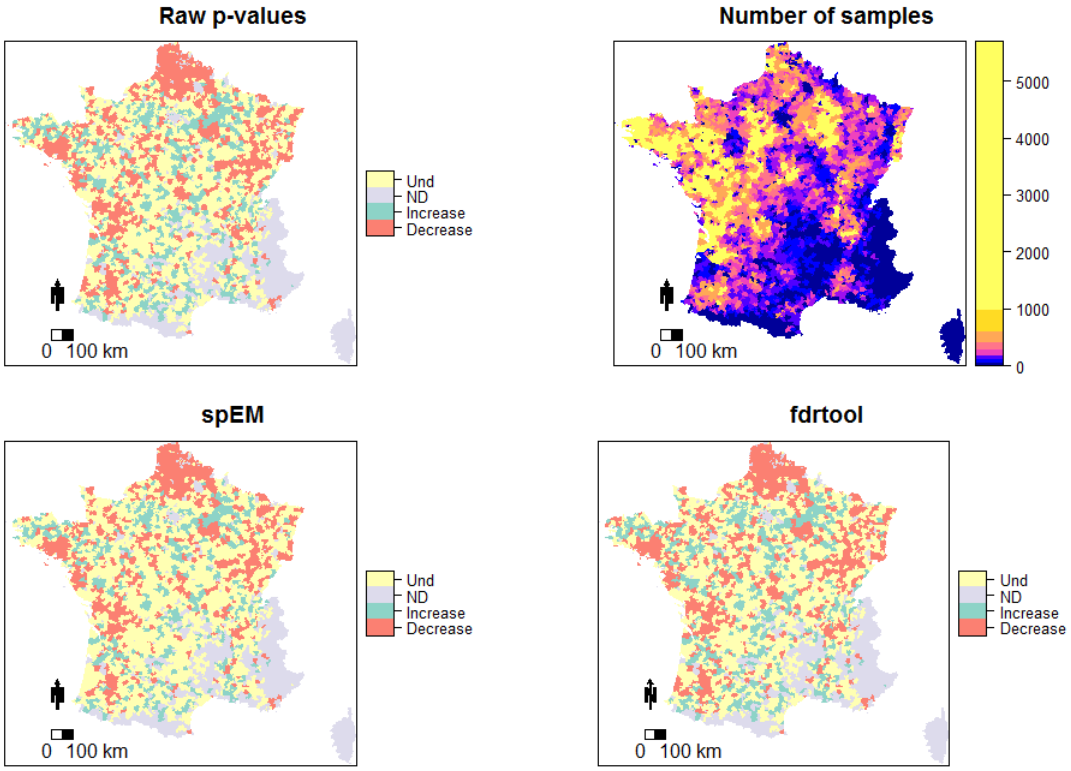

Figure 4: *Maps of the results of the multiple tests procedures at the cantonal level for a target level* $\alpha = 0.1$. *The upper left map corresponds to the raw p-values. The upper right map corresponds to the number of samples used in the procedure. The lower left map uses the FDR estimated by spEM. The lower right uses the FDR estimated by fdrtool. Und corresponds to the cantons where H0 cannot be rejected. ND corresponds to the cantons where there are not enough data to compute the test.*

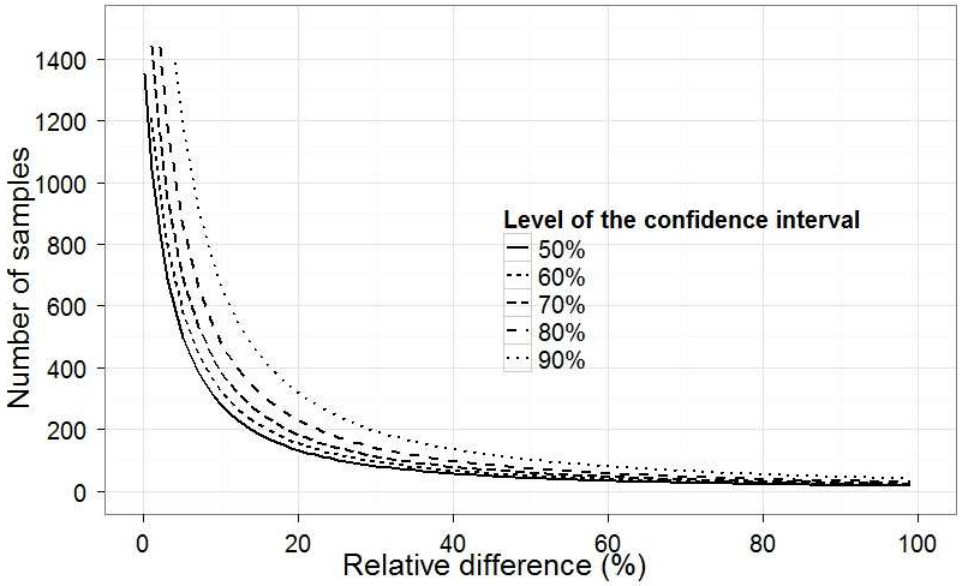

Figure 5: *Upper prediction intervals of levels 50%,. . . ,90%, for the minimim sample sizes needed at each campaign to detect a given change expressed in median relative difference between two time intervals, at a 10% FDR level.*