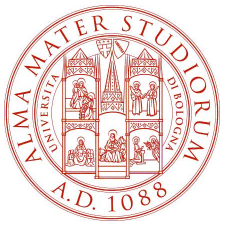


Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

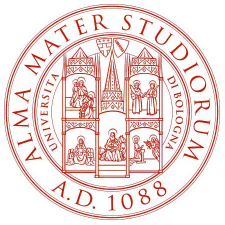
Alma Mater Studiorum - Università di Bologna

Multiple testing in spatial epidemiology: a Bayesian approach

Massimo Ventrucci

Dipartimento di Scienze Statistiche "P. Fortunati"

Marzo 2009



Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI ciclo

Multiple testing in spatial epidemiology: a Bayesian approach

Massimo Ventrucci

Coordinator
Professor Daniela Cocchi

Tutor
Professor Daniela Cocchi
Co-tutor
Professor Marian Scott

Settore Disciplinare
SECS-S/01

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2009

Contents

Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor Daniela Cocchi for the continuous support demonstrated during the last three years, the helpful suggestions and advices she gave me.

I am deeply indebted to my co-tutor Professor Marian Scott who helped me to approach many practical problems during my experience at the Statistics Department of the Glasgow University, providing me a constant support.
A special thanks goes also to Fedele Greco, Francesca Bruno, Carlo Trivisano, Michele Scagliarini, Rossella Miglio Clarissa Ferrari and Fabio Di Narzo who shared ideas about my work in many meetings. I cannot forget to especially thank Fedele, I very much appreciated time and energy he put into our fruitful discussions.
I had a very special moment in Glasgow and I would like to thank all professors, researchers, post-graduates and friends with whom I shared ideas and had very happy moments. Thus, thank you to Maria, Raul, Laura, Irene, Sara, Giorgio, Erik, Paolo and Jude.

A special thought goes to all friends met in the last three years in Bologna, everyone played a role in some sense and I cannot forget it. I cannot also forget my PhD colleagues, so thanks to Davide, Clarissa, Ida, Anna, Maria Serena, Elena.
A special thanks goes to Christian who can make me sure to have a good friend. Thank you to Federica for exactly the same reason.
Another special thought goes to my best friends in my village, who always wait for me to going back home, whereas I rarely do it. Also thanks, perhaps, to whom wait for me no longer.
In conclusion, I cannot forget to thank my family for the support they always provide me: Liliana, Giorgio, Stefano, Antonietta, Lino and finally Tommaso who cannot yet read these pages.

Massimo Ventrucci

Introduction

In this work we propose a new approach for preliminary epidemiological studies on Standardized Mortality Ratios (SMR) collected in many spatial regions. A preliminary (also called descriptive) analysis in this field aims to formulate hypotheses to be investigated via individual epidemiological studies that avoid bias introduced by aggregated analyses.

Starting from collecting disease counts y_i, calculating expected disease counts e_i by means of reference population disease rates, and assuming each area i count is distributed as a Poisson with mean $(e_i \cdot r_i)$, an $SMR_i = \frac{y_i}{e_i}$ is derived as the ML estimate of the parameter r_i, that is the relative risk for the disease under examination in area i. Such estimators have high standard errors when referred to small areas, i.e. areas where the expected count e_i is low either because of the small number of people living in the area or the rarity of the disease under study. Therefore, the presence of small areas yields maps of ML relative risk point estimates that are discontinuous; when the expected count is very low (even lower than 1) a huge SMR value may be caused by the occurrence of few disease cases. As a result a map of SMRs will tend to only highlight risk in poorly populated areas. If we undertake a hypothesis testing inferential approach, so evaluating the null hypothesis of absence of risk ($H_{0i} : r_i = 1$) against the alternative of a higher risk ($H_{1i} : r_i > 1$) in each area i by means of p-values computed with Poisson c.d.f., we meet the opposite problem: the test is more likely to be significant (more powerful) in non-small areas than in small areas, hence a map of p-values will tend to only highlight risk in high population areas.

Disease mapping models providing maps of smoothed relative risk estimates and other techniques for screening disease rates on the map, that aim to detect possible high-risk areas, have been proposed in the literature according to the classical and the Bayesian paradigm. Our proposal approaches this issue through a decision-oriented method: we want to evaluate many null hypotheses focusing on multiple testing control, without however leaving the "preliminary study" perspective. More precisely, we implement a multiple testing procedure that controls the False Discovery Rate (FDR), i.e. the number of falsely rejected null hypotheses (false discoveries) divided by the number of rejected null hypotheses (discoveries). This quantity is largely used to address multiple comparisons problems in the field of microarray data analysis but it is not usually employed either in

testing many hypotheses on a large SMRs dataset or in disease mapping applications, that are not concerned with testing hypotheses but only with point estimation of true relative risk values. Controlling the FDR means providing an estimate of the proportion of false discoveries for a set of discoveries, where a discovery is a declared high-risk area.

The presence of small areas and of positive spatial correlation between risks, that are frequently encountered in practice, create difficulties in applying p-value based traditional methods for FDR control/estimation (Benjamini and Hochberg, 1995; Storey, 2003) because the necessary distributional assumptions on the p-values do not generally hold. More precisely, the p-values cannot be assumed as independent when spatial correlation between risks is expected; furthermore they are not identically distributed under the null hypothesis as $U(0,1)$ when the population underlying the map is non-homogeneous, counts are sparse, and hence over-dispersion is expected.

The Bayesian paradigm offers a way to overcome the inappropriateness of p-value based methods. In the present work we propose a hierarchical full Bayesian model for FDR estimation in a testing framework where many null hypotheses of absence of risk are evaluated on the observed SMRs. We want to focus on cases where SMRs are collected in small areas and risks are spatially correlated, i.e. in cases where there is a lack of fit of the usually assumed Poisson model for independent counts. We will use concepts of Bayesian modeling for disease mapping, referring in particular to the Besag York and Mollié model (Besag York and Mollié, 1991) often used in practice for its prior assumptions flexibility w.r.t the distribution of risk parameters $\boldsymbol{r} = (r_1, ..., r_N)$ in the whole map. The borrowing of strength between prior and likelihood typical of a hierarchical Bayesian model takes advantage of evaluating the test in a given area i by means of all observations in the map under study ($\boldsymbol{y} = (y_1, ..., y_N)$) rather than just by means of the observation in the given area (y_i). This can improve the power of the test in small areas and addresses more appropriately the spatial correlation issue that suggests that relative risks are closer in spatially contiguous regions.

In practice, the proposed model aims primarily to make the practitioner able to declare a number of areas as high-risk areas (i.e. to reject a number of null hypotheses) controlling a desired level of FDR fixed a priori. Another peculiarity is its capability to still provide posterior estimates of relative risk values, that are the inferential target of the Besag York and Mollié model. As regards the primary aim, we can obtain an estimate of the False Discovery Rate through MCMC estimation of each area specific posterior probability that the null hypothesis is true, denoted as $\pi_i = P(H_{0i}|\boldsymbol{y})$. To be precise, we will focus on controlling the expected FDR conditional on data (Broet et al., 2004), denoted as $\widehat{FDR}$. This quantity can be worked out given any set of $\widehat{\pi}_i$'s by computing the empirical mean over them; the key point is that each $\widehat{\pi}_i$ is an estimate of the type I error probability relative to the rejecting H_{0i}. Thus, we can consider this set of posterior probabilities as relative

to the areas declared at high-risk, and $\widehat{FDR}$ as an estimate of the proportion of false discoveries which can occur in making such declarations.

The most interesting aspect of the work is the capability of the model to provide a non-arbitrary decision rule for rejecting null hypotheses that is based on the knowledge of $\widehat{FDR}$; we call such rules "$\widehat{FDR}$ based decision (or selection) rules". In a formal sense, a decision rule is defined as a function of the $\widehat{\pi}_i$'s and a threshold t_π, such that if $\widehat{\pi}_i \leq t_\pi$, then H_{0i} (i.e. $r_i = 1$) is rejected in favor of the alternative H_{1i} (i.e. $r_i > 1$). By applying for instance an $\widehat{FDR} = c$ based rule, where c is the pre-fixed FDR level, the practitioner can select as many as possible areas such that the $\widehat{FDR} \geq c$. The sensitivity and specificity of such rules depend on the goodness of estimation of the FDR. On this note, what is required in order to achieve a control is a "conservative" FDR estimation (Storey, 2002), that is $\widehat{FDR} \geq$ true FDR.

A simulation study to evaluate the model performance in FDR estimation in terms of accuracy, sensitivity and specificity of the decision rule, and not least the performance in goodness of estimation of relative risks, was set up. We chose a real map from which we generated several spatial scenarios whose simulated disease counts vary according to the spatial correlation degree, the size of the areas, the number of areas where the alternative hypothesis is true (HR areas) and the risk level in the HR areas. For each dataset (in total 100) of each scenario (in total 54) the model was ran using BRugs package (version 0.4 - 1) that implements OpenBUGS version 3.0.2. The main aim of the simulation is evaluating which FDR levels are conservatively estimated (i.e. not under-estimated) by the model in each scenarios, focusing the interest in small areas and spatially correlated risks scenarios.

An application to real data is finally presented to show the two kinds of maps that the method can produce: a map of posterior relative risk estimates and a map of highlighted high-risk areas given a pre-chosen value of $\widehat{FDR}$.

The plan of the work is as follows. In chapter 1 the basic concepts of FDR and multiple testing are illustrated. Chapter 2 introduces the spatial epidemiological case study, the motivation of the work and the multiple hypothesis setting based on SMRs. In chapter 3 we discuss the characteristics of the proposed model to estimate the expect FDR conditional on data. Chapter 4 describes features and results of the simulation study and gives application to a real dataset.

Chapter 1

Basic concepts on Multiple Testing

In evaluating a null hypothesis H_0 we need a decision rule $d(z(y), t_y)$, function of a summary statistics of data $z(y)$ and a threshold t_z, to decide whether or not H_0 can be rejected. In Neyman-Pearson theory, for instance, this function is ruled out such that the *type I* error probability (probability of rejecting the null hypothesis when it is true) cannot be greater than a pre-chosen level α. Traditionally, the most typical case is when the null hypothesis is evaluated by means of a p-value dependent on a chosen summary statistics of data $z(y)$ and on distributional assumption for $z(y)$; a p-value is calculated as the probability of randomly occurring values at least as unlikely as the observed $z(y)$. Rejecting the null hypothesis when the calculated p-value is less than the threshold $t_{p-value} = \alpha$ guarantee the control of the *typeI* error probability in doing a test for evaluating a single null hypothesis. When tests are performed on many, say m, null hypothesis $\{H_{01}, ..., H_{0m}\}$, and a set of m p-values $\{p_1, ..., p_m\}$ are available, it is very likely that at least one of them is lower than α even if all null hypothesis are true. Precisely, if tests are independent, $P(\text{at least one of } \{p_1, ..., p_m\} < \alpha | H_{01}, ..., H_{0m}) = 1 - (1 - \alpha)^m$ that for $m = 10$ tests becomes around 0.4. Thus, if in a multiple testing set up one employs the same decision function $d(z(y), t_{p-value})$ used for evaluating a single test, the probability of at least one *typeI* error is greater than the pre-chosen level α.

Therefore, every time we make evaluations about a multiplicity of null hypotheses we need to control a global error related in some sense to the *typeI* errors that can occur. Table 1.1 shows all possible outcomes from a multiple testing procedure and suggests a variety of global error measures which could be controlled in practice. Most traditional methods aim to control the quantity $P(V \geq 1)$, that is a multiple testing global error measure correspondent to the *type I* error rate in the single hypothesis testing set up. It is called Family Wise Error Rate ($FWER$), and it is the probability of obtaining at least on false positive.

In general, several techniques and different approaches have been proposed regarding the kind

	Accept null	Reject null	Total
Null true	U	V	m_0
Null false	T	S	m_1
	$m - R$	R	m

Table 1.1: Possible outcomes from testing m null hypothesis

of error measure considered, the inferential approach to estimate or control it (see section 1.2), the interpretation of probability on which inference is based (see section 1.1), the particular context the multiple tests are conducted under. In this chapter we do not discuss multiple testing methods in general, but just focus on methods for controlling the False Discovery Rate (FDR), a particular global error measure that we believe fruitful in the case of study of spatial epidemiology under examination. Motivations for this choice are explained in chapter 2.

1.1 False Discovery Rate

In this work we shall focus on the False Discovery Rate, that is the proportion of false discoveries (or false positives, or number of null hypotheses wrongly rejected) among all the discoveries (or positives, or number of null hypotheses rejected):

$$FDR = \frac{V}{R} \tag{1.1}$$

Note it is a random quantity (neither bayesian nor frequentist) where both numerator and denominator are unknown short of having determined a decision rule for rejecting null hypotheses. It is worth noting that the authors who introduced the False Discovery Rate (Benjamini and Hochberg, 1995) called FDR the quantity $E\left[\frac{V}{R}\right]$. We will instead denote FDR as simply the fraction between false discoveries and discoveries, following the terminology of Genovese and Wassermann (2003). The FDR as a global error measure is frequently employed in the field of microarray data analysis where multiple comparison problems arises in the identification of differentially expressed genes among a large number of observed gene expressions.

Several authors introduced p-value based methods that "adjust" the procedure for rejecting hypotheses such that in average the expected FDR is lower than a pre-specified error (Benjamini and Hochberg, 1995; Storey, 2002). We refer to such methods as "frequentists" since they control the expected value of such a global error, taking the expectation over repeated experiments. Other proposals follow a Bayesian perspective and consider the null hypotheses as random variables, taking the expectation over them conditionally on the observed data. We will not review the several

methods proposed in literature but focus primarily on the methods based on posterior probabilities of the null hypothesis (see section 1.3) rather than p-values (see section 1.2). For a methodological review of Bayesian proposals see Berry and Hochberg (1999); for a decisional theoretical approach see Muller et al. (2006), while for an application of the FDR estimation through a hierarchical Bayesian modeling framework in microarray data see Newton et al. (2004) and Broet et al. (2004).

1.2 Frequentist methods

Traditional methods for addressing FDR control are based on the knowledge of p-values and make use of frequentist arguments to demonstrate the control. Benjamini and Hochberg (1995) achieve the control of the expectation of the FDR as defined in (1.1), i.e. $E(FDR) = E\left(\frac{V}{R}\right)$. In Benjamini and Hochberg (BH) procedure a control of $E(FDR)$ is obtained by rejecting as many hypotheses as possible such that $E(FDR)$ is lower than a pre-specified value α. Such p-value based procedure allows for rejecting all null hypotheses for which $p_i \leq t_{p-value} \equiv p_{(j)}$ where:

$$j = max\left\{0 \leq i \leq m : p_{(i)} \leq \alpha\frac{i}{m}\right\}, \tag{1.2}$$

and $0 \equiv p_{(0)} < p_{(1)} < ... < p_{(m)}$ denote the ordered p-values. BH demonstrated that $E(FDR) \leq \alpha$ regardless of how many null hypotheses are true and regardless of the distribution of the p-values under the alternative hypothesis.

Storey (2002) focus on a "conservative" estimation of the expected positive False Discovery Rate ($pFDR$) given a threshold $t_{p-value}$:

$$pFDR = E\left(\frac{V}{R}|R > 0\right) \tag{1.3}$$

A conservative estimation as intended by the author is such that:

$$E(\widehat{pFDR}(t_{p-value})) \geq E(pFDR(t_{p-value})). \tag{1.4}$$

There is here a change of perspective in that the control is achieved by estimation of the $pFDR$ for fixed monotonic rejection regions (or monotonic sets of p-values) rather than by prefixing the level of FDR and work out the rejection region as in BH procedure. Storey (2003) uses a Bayesian argument for ruling out a non-parametric estimator for the $pFDR$ conditional on $t_{p-value}$. He provides estimators of a quantity called q-value that can be work out for each observed p-value. Briefly, the q-value* relative to a given p-value* corresponds to the $pFDR$ estimates conditional on the rejection of all p-values lower than p-value*; for details see Storey (2003). Thus, building a set of crescent-ordered p-values we can get an estimate of the $pFDR$ by calculating the q-value

relative to the highest observed p-value. Storey also shows a connection between his method and BH procedure, demonstrating that for the same level of FDR a higher number of null hypotheses can be rejected, gaining hence more sensitivity. The method makes however stronger assumption than BH, the most relevant being assumptions which yields the "Bayesian interpretation" of the $pFDR$ (Storey, 2003). Suppose m identical hypothesis tests aiming to evaluate null hypotheses $\{H_{0i}, ..., H_{0m}\}$ are performed considering the summary statistics $\{z(y_1), ..., z(y_m)\}$ and the rejection region Γ. Let us assume that $(z(y_i), H_{0i})$, for $i = 1, ..., m$, are i.i.d. random variables with marginal distribution:

$$[z(y_i)|H_{0i}] = H_{0i} \cdot F_0 + (1 - H_{0i}) \cdot F_1 \tag{1.5}$$

where F_0 and F_1 are the distributions of $z(y_i)$ respectively under the null and alternative hypothesis, and $H_{0i} \sim Bernoulli(pr_{H_0})$. Then it follows that:

$$pFDR(\Gamma) = P(H_{0i} = 1|z(y_i) \in \Gamma) \tag{1.6}$$

(the notation $[a|b]$ to mean the distribution of the random variable a conditional on b will be used in the following mostly when analytic expressions are introduced; when a more compact notation helps the comprehension we will use the classic $\sim$).

The result holds for all $i = 1, ..., m$ and regardless of m. Moreover it is valid even if we consider p-values instead of summary statistics. Storey proposed non-parametric estimators of $pFDR(\Gamma)$ by stressing the reasonable assumption that p-values distribution under H_0 is $Uniform(0, 1)$ so achieving an estimator for the overall probability of the null hypothesis (calculated over the whole set of tests). The attempt to estimate such an overall probability is what allows the gain in power with respect to the BH procedure. In fact, considering a generalization of the BH result (Genovese and Wasserman, 2003) we see that the BH procedure assures that $E(FDR) \leq a \cdot \alpha \leq \alpha$, where a is the overall probability of the null hypothesis, that is implicitly equal to 1 in the BH procedure. Indeed, in the case where the $pFDR$ estimator is obtained by using the most conservative estimation of the overall probability of the null hypothesis (i.e. $a = 1$), BH procedure and Storey's method are equivalent.

1.3 Bayesian methods

The FDR is a ratio where both the numerator and the denominator depend on a decision about the set of null hypotheses. We can generally define such a decision rule for H_{0i} with the indicator function $d_i(\cdot) = I(H_{0i}$ is rejected).

Suppose $H_{0i} = 1$ indicates the unknown null hypothesis is true and $H_{0i} = 0$ that the alternative is true, hence the true FDR can be formally expressed by:

$$FDR = \frac{\sum_i H_{0i} \cdot d_i}{D} \tag{1.7}$$

where $D = \sum_i d_i$ and $d_i = 1$ if the decision rule is such that H_{0i} is rejected. Frequentist methods use decision rule of the form $d_i = I(z(y_i) \leq t_z)$, hence based on a summary statistics $z(y_i)$ of data and a critical value t_z. Decision rules of this form (or equivalently based on p-values) are intuitive but not necessarily optimal as observed by Muller et al. (2006).

The Bayesian methods we will focus on consider H_{0i} a binary random variable (equal to 1 when it is true) and allow us to determine the decision on the ith null hypothesis by the posterior probability that the null hypothesis itself is true, that is:

$$\pi_i = P(H_{0i} = 1|data) = E(H_{0i}|data). \tag{1.8}$$

note this quantity is conditional on the observed data. Muller et al. (2006) discusses decision rules of the form $d_i(\pi_i, t_\pi)$.

Considering frequentist expectation of the FDR, i.e. expectation over hypothetically repeated experiments, we need to consider expectation over a ratio of random variables since the decision $d_i(z(y_i))$ is a function of the data and appears in both numerator and denominator of the ratio. Under a Bayesian perspective the discussion simplifies because, looking at (1.7), the only unknown quantity is the unknown H_{0i} in the numerator, D being determined by a decision rule $d_i(\pi_i, t_\pi)$ that is conditional on data. The FDR here is a function of the π_i's (and of a threshold t_π for the π_i's) which are posterior probabilities conditional on the observed data. Thus, considering each conditional expected value of H_{0i}, i.e the π_i's defined in (1.8), we derive the expected FDR conditional on data by:

$$E(FDR|data) = \frac{\sum_i \pi_i d(\pi_i < t_\pi)}{D} \tag{1.9}$$

where the expectation is relative to H_{0i}. This quantity is often used for addressing multiple comparisons problems in microarray data analysis. Following the mixture assumption of Storey (1.5) and introducing exchangeability (instead of i.i.d.) assumption on summary statistics, authors (Newton et al., 2004; Broet et al., 2004) proposed fully hierarchical Bayesian models for estimating the expected FDR conditional on data. Such models can provide an estimate of each π_i via MCMC computation as a Monte Carlo mean over a sample of realizations from the respective posterior distribution $[H_{0i}|data]$. Given the estimates $\widehat{\pi}_i$'s, an estimate of the expected FDR conditional on data is provided for any set of discoveries of cardinality D by:

$$\widehat{E(FDR|data)} = \frac{\sum_i \widehat{\pi}_i d(\pi_i < t_\pi)}{D} \tag{1.10}$$

As suggested in Newton et al. (2004), an estimate of the expected FDR conditional on data can be a suitable way to determine a decision function. Indeed, one can declare a null hypothesis as rejected if $\pi_i < t_\pi*$, where $t_\pi*$ is fixed to achieve a certain pre-set estimated FDR, say $E(\widehat{FDR|data}) \geq c$. Moreover, the authors observed the dual role of $\widehat{\pi}_i$ in decision rules like $d_i(\widehat{\pi}_i, t_\pi)$. It not only determines the decision on H_{0i} but also reports both the probability of a false discovery as $\widehat{\pi}_i$, if $d_i = 1$, and the probability a false non-discovery as $1 - \widehat{\pi}_i$, if $d_i = 0$.

We will talk in more detail of these concepts in sections 3.2 and 3.2.1 when introducing and discussing the model proposed to address the case of study under exam.

1.3.1 Posterior probability adjustment for multiple testing

Estimating the posterior probability that the null hypothesis is true underlies considering the null hypothesis H_{0i} as a random variable rather than an unknown fixed parameter. Berry and Hochberg (1999) observed "Posterior inference adjusts for multiplicities, and no further adjustment is required". The statement is true provided the assuming a probabilistic model for the null hypotheses. First, the probability model needs to include a positive prior probability for the event "null hypothesis is true". Second, the model needs to include a hyperparameter that defines the prior probability mass for all null hypotheses themselves. Moreover they comment that "finding posterior distribution of parameters is only part of the Bayesian solution. The reminder involves decision analysis". Paper of Muller et al. (2006) discusses such a decision theoretical perspective and derive optimal decision rules based on the π_i's under several loss functions dependent on FDR and FNR (the False Non Discovery Rate). The optimal rule, determined by minimizing expected FDR conditional on data, is of the form $d_i = I(\pi_i < t_\pi)$, where t_π can be analytically determined under several loss functions.

To sum up, a full Bayesian hierarchical modeling is required to achieve what Berry and Hochberg called a posterior probabilities adjustment. The underlying idea of a fully Bayesian approach is evaluating the ith test by means of π_i, i.e. the ith posterior probability, but exploiting a Bayesian shrinkage estimation such that all observations (not only ith observation) contribute to estimate π_i. Thus, the posterior distribution of π_i will depend on all the observed data, not only on the ith observed summary statistics.

Chapter 2

Multiple testing on large datasets of Standardized Mortality Ratios

Multiple testing in epidemiological applications is not always considered a primary issue. Some authors deny the adoption of procedures to account for $FWER$ (Rothman and Greenland, 1998), others advocate the control of it in epidemiological surveillance applications (Frisén, 2003; Kulldorff, 2001; Elliott et al., 2000), though it has been noticed the cost in sensitivity of adopting the control of $FWER$ in on-line monitoring (Rolka et al., 2007). It seems clear that multiple testing issues have to be considered in relation to a particular application. An important issue to carefully evaluate is the choice of the particular global error measure (see table 1.1 in chapter 1) to control/estimate in each particular case study. In this work we do not want to discuss multiple testing control from a theoretical point of view and claim it is necessary in the example we will introduce in this chapter, but we want to show that it can be viewed as a possible way to conduct a descriptive analysis of geographical epidemiology starting from the collection of many disease indicators. We shall describe a common spatial epidemiological example, its main features, its objectives and the statistical issues which arise. Then, in chapter 3, we shall attempt to build a multiple testing procedure for it by means of a Bayesian hierarchical model that allows for estimating the FDR. We will give reasons why this can be thought of as an interesting alternative to address a descriptive geographical epidemiological analysis, mostly in cases where data are over-dispersed and spatially correlated.

Firstly we introduce the epidemiological case under study and give some examples of its possible objectives in practice. Briefly, a descriptive analysis of Standardized Mortality or Morbidity Ratios (SMR), collected in many areas, is undertaken to identify unusually high risks. The aim is to screen the health status of area-specific populations, to identify priority for public health interventions or to suggest further analytical studies. For instance, an epidemiologist could be asked to look

at the difference in rates among the areas and attempt a ranking of higher-risk areas in order to allocate resources for public health administrative objectives (Greenland and Robins, 1991). Another example is the mapping of risk indicators relative to some disease of interest; here the scope is to describe the spatial distribution of the disease and to get clues about a possible association between the disease itself and the exposure to environmental risk factors. Moreover, a preliminary analysis can be done to detect clusters of one or more diseases under evaluation. For instance, one could test for the randomness of any pattern that can be found across areas, or screen for evidence of an individual disease hot spot (without any preconception about its likely location); the former are called tests for clustering and the latter tests for the detection of clusters (Besag and Newell, 1991). Thus, we see the range of preliminary statistical tools can be large and cannot be discussed here, just consider that, in general, methods use many different statistical inferential approaches according to their specific objectives and data features; see (Lawson et al., 1999) for a review of statistical methodologies available for addressing geographical analysis and some interesting remarks about their appropriateness in guiding public health policy decisions.

Among the examples above mentioned we will only discuss Bayesian disease mapping models, i.e. methods that give smoothed point estimates of risks in each area of the map considered. We would like to point out that both disease mapping models and the methodology proposed here to perform a multiple testing procedure controlling the FDR are only suitable for an analysis that may form the basis for subsequent epidemiologic investigations but will rarely be an end in themselves. In this work, the kind of preliminary analysis we will pursue is about testing each area-specific risk for a disease of interest (or more diseases that have the effect of augmenting the number of tests) being aware of the proportion of unusual high risks that more likely may have originated by chance.

As regards the importance of the scale to which the events disease can be recorded, we can distinguish data collected at count (areal data) or point (case-event data) level, the former being our focus. For areal data analysis what is very relevant is the level of aggregation, i.e. if we have small or big spatial regions in the map under study. Often counts aggregated in small areas produces a zero count, a situation denoting data sparseness. An area is called small when a small count of disease events is expected inside it (sometimes the expected count is even lower than 1); this can either be due to the rarity of the disease under examination or because of the small number of people living there. The presence of small areas is one of the two main challenging issues we want to focus on, the second being the presence of spatial correlation between the risks.

In section 2.1 we give reasons for addressing an epidemiological descriptive analysis by means of a a multiple testing procedure on Standardized Mortality Ratios collected in many regions. We do not loose generality if we have many disease indicators collected in many areas. In section 2.3

we introduce the multiple hypothesis testing framework

To sum up, in this chapter we pose the ground for introducing a method to estimate the False Discovery Rate and controlling the multiple testing error, arguments discussed in chapter 3. We also consider p-value based control methods and claim their inappropriateness for several reasons. Hence, we propose to estimate the FDR by using a methodology developed through a well known disease mapping model (Besag et al., 1991) that is flexible as regards problems due to small areas and spatial correlation. We will also note that, in principle, the proposed model is able to address both the FDR estimation relative to any possible set of rejected hypotheses and the point estimation of true relative risk values.

2.1 The rationale of the work

Epidemiological studies aiming to identify unusually higher risks for one or more diseases over a map of geographic areas are denoted as descriptive studies even if the methods involved in such analysis often stress complex modelling assumptions. The general aim is to screen the health status of each area population by means of suitable disease indicators to identify risk increments for the examined diseases. Analyses are usually carried out on a predefined number of areas at a national, regional, or municipality level. The case we are focusing on is when a number of indicators are available collected in many regions, producing a large dataset. We also consider the presence of small areas over the map. Starting with the knowledge of such large datasets we want to highlight as many anomalies as possible, controlling some sort of measure which inform us about the anomalies that are imputable to only random error. Moreover, we would like to be able to evaluate the magnitude of risks in the areas declared as possibly at high-risk.

A well suited motivating example can be found in Catelan and Biggeri (2008), where authors pursue a statistical approach to rank multiple priorities in a case of study of environmental epidemiology. They mention the control of the positive-False Discovery Rate like in Storey (2003) as a possible approach to such a case study. Our work is in the same direction as regards the epidemiological rationale, that is finding discoveries, or in other words, providing clues of which indicators are susceptible to represent high-risk situations, these being the priorities of investigation in a perspective of a preliminary and explorative statistical tool for epidemiologists. On this note someone may argue that controlling the False Discovery Rate determines a decision-oriented approach which, traditionally, is not at all representative of a standard descriptive analysis. However, we think in such a case study a testing framework can still be though of as only addressing a preliminary statistical analysis since "descriptive" in this epidemiologic field does not strictly mean an analysis carried out by only summarizing empiric observations. Furthermore, the "ecology fallacy"

that typically affects such kind of aggregate indicators, limits conclusions that can be worked out by them, and makes such analyses advisable just at a preliminary stage (Lawson et al., 1999). Indeed, the further investigations aiming to confirm hypotheses generated by a descriptive analysis are, in the field of epidemiology, expensive observational studies (cohort studies, case control studies) conducted at an individual level. Thus, a decision based approach based on testing and control the FDR on observed areal disease indicators cannot be considered a statistical tool finalized to confirm the presence of environmental risk factors in those areas highlighted as at high-risk. It can, at most, generate hypothesis to further investigate through observational individual level studies.

2.1.1 The small areas issue

The recent availability of geographical indexed health and population data, together with advances in geographic information systems, has encouraged the epidemiological analysis on a small geographic scale. A lot of issues arise with data collected in small areas in epidemiological applications of spatial statistics (Elliott et al., 2000). Some motivations are built around the increasing interpretability of small-scale studies, as they are less affected, in principle, by the ecology bias due to aggregating counts in areas that are heterogenous about the exposure to environmental factors (within area clusters presence is less probable if counts are aggregated at a small-scale). Conversely, small-scale studies require more sophisticated statistical techniques because the data are usually sparse with low (even zero) counts of events in most of the regions; a situation of data sparseness may still arise in large-scale studies when the disease is very rare. Furthermore, there is often evidence of over-dispersion of the counts with respect to the typically assumed Poisson model (see section 2.3.3) as well as spatial patterns indicating dependence between area-specific risks, both reasons making it suitable to approach the analysis by following the Bayesian hierarchical model paradigm. More details about the implications of the over-dispersion in the case under study can be found in section 2.3.3

2.1.2 The inferential approach

There are two main reasons why we undertake a Bayesian paradigm. First, data frequently encountered in practice are in the form of large datasets of disease indicators collected at a small-scale, with underlying disease relative risks being spatially correlated across areas, both issues representing statistical challenge well addressed by means of full Bayesian hierarchical models. The second reason arises from the inferential approach we want to undertake for addressing an analysis of such a large dataset, that is a multiple testing procedure controlling a particular multiple testing error, i.e. the FDR (1.1). To this end Bayesian solutions have been suggested by some authors

in microarray data analysis for the selection of differentially expressed genes (Newton et al., 2004; Broet et al., 2004). Common features between the microarray context and our case study lies in the fact that a strict control is not recommended, the identification of as many as possible anomalies being the main interest. The main change of perspective lies in considering the null hypothesis as a random variable rather than a fixed unknown parameter. As a result, for the Bayesian statistician, the inferential target will be the posterior probability π_i of each area-specific null hypothesis (1.8) rather than p-values.

As said, our approach is decision-oriented, that is we want to make inference on a lot of null hypotheses since we are mostly interested in a method to actually provide rules for deciding which areas can be claimed high-risk areas, i.e. finding a rule for rejecting the null hypothesis, or for selecting discoveries in Benjamini Hochberg terminology. The reason why we propose a model for estimating the FDR is that controlling the FDR in our case would mean measuring the error we incur in selecting high-risk areas, so that, for any given set of rejected hypotheses we obtain an inferential tool to estimate the FDR. However, to be able to determine a selection rule we still need to change perspective: we need a method which, fixing a desired FDR level, makes us able to select as many high-risk areas as possible, or discoveries, being sure that the expected FDR would not be greater than what is a priori pre-specified. We saw in chapter 1 that the Benjamini-Hochberg procedure assures this under independence of p-values, whereas Storey builds a more powerful method by means of more complex assumptions. We can say in advance that such methods are inappropriate when data show spatial correlation and over-dispersion (see section 2.3).

In chapter 3 we will discuss a model for estimating FDR overcoming in part the mentioned difficulties. We will also discuss a way to determine selection rules based on the knowledge of the estimated FDR for any set of areas declared at high-risk. As long as the model is good at estimating the FDR, we can be reasonably certain in claiming a given number of discoveries at the pre-chosen level of FDR. We can think of such a pre-chosen FDR level as a nominal value that the model aims to predict. Thus, the reliability of such a kind of rule will be dependent on the ability of the model to accurately estimate the FDR, especially avoiding under-estimation in order to achieve a conservative control (similarly to the Storey's perspective (1.4)). In chapter 4 we will discuss about the spatial contexts, frequently met in practice, where the model can achieve conservative FDR estimation. Moreover in chapter 4 we shall regard the FDR estimation capability as an interesting way to determine a "non-arbitrary" rule for selecting high-risk areas. However, it is worth being aware of the lack of sensitivity or specificity that such rules may yield for some FDR (nominal) values; see section 4.4.

We think that the method we will later introduce offers a first attempt to measure the error in

making statements about many hypotheses of absence of risk evaluated in small areas. We recall that the idea of selecting possible high-risk areas is of some sort of interest in epidemiology and some authors have suggested rules focused to this aim. For instance, Richardson et al. (2004) proposed to base a rule on the posterior probability that the relative risk is greater than 1. After a large simulation study a value of 0.8 came out as more appropriate in most spatial scenarios evaluated. Evaluating the posterior density placed in the right tail beyond 1 can be useful information, even if the goodness of such a rule is conditional on what value is chosen for the threshold (0.8 or another value?). Thus, unfortunately such a rule needs an arbitrary choice by the practitioner; a choice whose effect in terms of error produced in declaring high-risk areas cannot be known (because the object which the rule is based on, i.e. the posterior density, is an estimate of the true relative risk value and there is no point in controlling a measure connected to the number of *Type I* errors). With the rules suggested in chapter 3 a threshold for the FDR needs to be chosen, hence making the practitioner aware of the number of errors he could at most incur. So, the only arbitrariness introduced is the choice of the FDR level, analogous in some sense to choose the size α for a test built with the Neyman-Pearson lemma. However, the goodness of the rule suggested still depends on the accuracy of the FDR estimation which the model proposed is demanded.

2.2 Standardized Mortality or Morbidity Ratios

It is usual to assess the disease risk in a map of contiguous regions by collecting the observed counts and calculating the expected counts. The ratio of observed to expected counts within tracts is called Standardized Mortality/Morbidity Ratios (SMR) and this ratio is an estimate of "relative risk" within each tract (i.e. the ratio describes the relative risk of being in the disease group rather than the background group). Such indicators are easy to compute and often used in geographical epidemiology studies and also in contexts not involving spatial issues like occupational epidemiology (Tsai et al., 1986). Part of the section 2.3 focuses on the issues which arise employing such indicators for testing null hypotheses in such a case study. We shall discuss characteristics of a multiple testing setting based on such indicators, then we will recognize the inappropriateness of frequentist p-value based methods for several reasons and claim the usefulness of a Bayesian hierarchical approach.

We now define the likelihood model from which $SMRs$ can be worked out as maximum likelihood estimators. We assume the disease occurrence is available in the form of count of cases over a map of spatial regions. Small areas are those where we expect a small number of cases because of the size of the area itself or the rarity of the disease. Within a map of N areas (census tracts, postal districts, municipalities), let y_i, e_i and r_i denote respectively the observed count, the

expected count and the unknown relative risk in area i. It is usual to stress an i.i.d. Poisson model for observed counts:

$$Y_i \sim Poisson(e_i \cdot r_i) \tag{2.1}$$

Thus, the likelihood of the relative risks r_i is:

$$[y_i|e_i] = \exp(-e_i \cdot r_i) \cdot \frac{(e_i \cdot r_i)^{y_i}}{y_i!} \tag{2.2}$$

Here and in the following, the notation $[a|b]$ generically denotes the conditional distribution of a given b. Similarly [a] will denote the marginal distribution of a.

The maximum likelihood estimator for r_i is actually the Standardized Mortality Ratio observed in area i:

$$\widehat{r_i} = SMR_i = \frac{y_i}{e_i} \tag{2.3}$$

The term e_i, that informs us about the area size, is assumed as known although it cannot actually be observed, but it depends on some underlying assumptions about the population at risk underlying the map. Operatively, it can be worked out after having stratified the population at risk by age groups (or age-sex groups) and assuming a multiplicative model for such age group risks. Stratifying is useful to make allowance for possible confounders. Since we want the SMR_i to be an indicator of the association between the disease in question and the environmental exposure in area i, we would like to get rid of all other possible variables that could modify ("confound") the actual environmental exposure of people in area i. For instance, older people may be thought to be more exposed than younger people to a given disease, or, some disease may be more dangerous for males than for females. We want the latter variables, age and sex, not to affect the SMR_i as we need such indicators to inform us on only the risk associated with living in area i, regardless of if the residents in area i are exposed to any other risk factors (age and sex in this case) not under study. In other words we might know that age and sex are important etiological factors for the disease of interest but in such an analysis we only want to focus on environmental factors. With this aim we can proceed by applying two indirect standardization methods: 1) a standardization using internal reference rates; 2) a standardization using reference rates of a standard external population. In the latter case we have:

$$e_i = \sum_j P_{ij} \cdot q_j$$

where q_j is the disease rate in age group j for the reference population and P_{ij} are the observed person-years at risk in area i for age group j (the number of persons in age group j who live in area i times the number of years over which we collect the count of disease y_i). Here, it is implicitly

assumed that the risk associated in living in area i (r_i) acts proportionally on the baseline risk for each strata (q_j), hence:

$$q_{ij} = r_i \cdot q_j \tag{2.4}$$

Without introducing the latter assumption we propose propose a model for each age group observed count (Pascutto et al., 2000):

$$Y_{ij} \sim Poisson(P_{ij} \cdot q_{ij})$$

However, the summation over the jth age groups in the (2.4) plus the assumption (2.4) lead directly to model 2.1 where the count in area i is distributed as Poisson with mean $(e_i \cdot r_i)$. The multiplicative model 2.4 is suitable as it allows for easily computing each area i expected count (e_i) that is the count of disease events we would expect if the disease rate in area i were equal to that of the standard population. Underlying e_i there is already an idea of null hypothesis; see section 2.3. Model with a combination of an additive and a multiplicative effect has also been proposed in literature (Best et al., 2000).

In the internal standardization case we use as reference rates those of the population of the whole map, hence:

$$q_j = \sum_i \frac{y_{ij}}{P_{ij}}$$

Thus, this method centers the data as $\sum_i e_i = \sum_i y_i$ and the overall mean disease rate is equal to 1. It is the most used standardization method because it requires only the observed data. The external standardization, instead, builds expected counts consistent with the mean disease rate of another population assumed as a reference; i.e. a population supposed not to be exposed to the same environmental risk factors under study. Using one instead of the other standardization makes a difference in the following sense: with internal standardization we will obtain SMRs greater than 1 (constant mean rate), but also lower than 1, so adopting it we can describe the internal variability of disease rates. With external standardization, data are not constrained to the observed/expected equivalence over summation ($\sum_i e_i = \sum_i y_i$) hence we can still obtain all disease rates greater or lower than 1, because 1 in this case is not the constant mean rate.

There are two other assumptions underlying model 2.1. The first concerns the Poisson probability distribution: in each area i, individual risk levels are independent of each other, that is the susceptibility to the disease is the same for all people living in that area. The second regards the independence and identity assumption (i.i.d.): counts of disease have no spatial correlation. In general, deviations from such assumptions yield counts more variable than what is expected under the Poisson model. This situation is commonly referred as over-dispersion, that is when the empirical variance of the data is larger than the variance specified by the model assumed to describe

the data. Typically it is assumed that the Poisson model is appropriate for rare and non infectious disease, though care should still be take about the level of aggregation of the disease events in counts, and the differences yielded by considering a regular grid or a real map with diverse area shapes; see further considerations on Poisson lack of fitting in section 2.3.3.

2.3 The multiple testing framework

Since we want to highlight unusually high relative risks we need many one tailed tests of hypothesis where the alternative hypothesis is that of higher relative risk value. Hence the two competing hypotheses are of the form:

$$H_{0i} : r_i = 1 \tag{2.5}$$

$$H_{1i} : r_i > 1 \tag{2.6}$$

As we shall see in section 2.3.1 a p-value can be computed under the model where Y_i is distributed as the Poisson of mean e_i by calculating the c.d.f. of such Poisson distribution.

It is worth mentioning now that when we set up the model for estimating the FDR (see chapter 3) we will consider the simple null hypothesis $r = 1$ which makes the model specification easier. Anyway, we want to say in advance that a complication will be met: a small posterior probability π_i can arise either in case where area i risk is lower or greater than 1. Hence, the practitioner, after having computed posterior probabilities for all areas, will have to work out the FDR estimation by only considering the set of the π_i's relative to areas eligible as possible discoveries (or possible high-risk areas). We will consider as eligible for becoming discoveries the areas where the observed is greater than the expected count.

If the inferential aim is to conduct an hypothesis test in each of the N areas the multiple testing problem ought to be addressed. In fact, doing a lot of tests we could incur wrong rejections making necessary control of a global error measure. We have decided to control the FDR aiming to provide information about how many False Discoveries we can expect, where a set of discoveries is defined as the set of areas that, by means of the decision rule, we can declare as being at high-risk.

As regards the choice of which multiple testing error control in such case we agree with authors that advocate the False Discovery Rate as a more appropriate measure than the $FWER$ in all cases where we need to find as many effects as possible (clues, anomalies etc.) in the dataset. As already mentioned in this chapter, this is consistent with a descriptive analysis aimed at screening disease indicators in an exploratory fashion rather than in a confirmatory study perspective, exactly like our case. In fact, the probability of making at least one false discovery ($FWER$) is not appealing here,

since we could reasonably let ourselves make even more than one error in declaring rejections; in other words, the decision to conduct more epidemiological investigations (on the whole map under study) need not be erroneous even if more than one null hypothesis is falsely rejected. Indeed, the control of $FWER$ by a Bonferroni adjustment would yield a lot of false non-discoveries as it is very conservative with respect to the null hypothesis. Moreover, the decision to keep on investigating with further high-cost epidemiological studies can not only be connected to this kind of geographical analysis because they are not free of troubles. Such SMRs analyses stand at the lowest level of proof about confirming etiological effects due to exposure to environmental risk factors because of the "ecology fallacy" and many other possible confounders unobserved at area level. The analysis is only required to throw light on possible anomalies. On this note, providing a method to estimate the FDR given a number of discoveries is a possible way to proceed. If, moreover, the method can estimate both false discoveries and relative risk values (i.e. the magnitude of the discoveries) it may probably be viewed as a more informative way than merely plotting relative risk values on a map or ranking high-risk areas. The model proposed in chapter 3 is indeed aimed at both estimating FDR and relative risks.

As an example, Biggeri et al. (2007) applied Storey $pFDR$ estimators to achieve q-values relative to disease indicators tested in some areas of Sardinia and observed the usefulness of the False Discovery Rate to asses the global degree of risk of a given area where more than one indicator was available (several causes of disease under examination). For each area they were able to select some discoveries at the level of FDR (indeed the $pFDR$) indicated by the q-value correspondent to each p-value. The limitation here is that a q-value (i.e. an estimation of the proportion of false discoveries) can be computed for only ordered sets of p-values (also denoted as p-value monotonic sets). We believe the FDR is a useful measure also for our case study, that is slightly different as we aim to make test on $SMRs$ collected in small and contiguous areas and also addressing case where risks show positive spatial correlation. As we will see in section 2.3.2 in our mind a small area is when the expected count e_i is lower than 5.

In the following we shall approach the methodology applied for estimating the FDR gradually. Firstly we will mention related arguments like the mapping issue, the interpretation of the over-dispersion as regards the null and alternative hypothesis and ways for computing a p-value for testing an SMR.

2.3.1 Mapping significance or mapping Relative risks

Producing maps of indicators telling us about the disease under examination is a primary aim of any descriptive analysis based on disease counts collected in many areas. A lot of authors have

noted the difficulty in interpreting maps of $SMRs$ (Cressie, 1993; Banerjee et al., 2004; Gelman and Price, 1999; Mollié, 1996; Elliott et al., 2000; Pascutto et al., 2000; Schlattmann et al., 1993) when the population underlying the areas is heterogeneous and mostly when some counts arise by counting events in poorly sampled areas. Poorly sampled areas are actually the small areas, those where the expected count is very small (even lower than 1) hence yielding a large SMR. Because of small areas the map will show a discontinuity in estimated rates since low expected count areas could yield a huge SMR for the occurrence of only a few cases of disease. On the other hand such estimates will be very inaccurate since the standard error of SRM_i depends on the population living in area i, precisely:

$$\widehat{sd}(SMR_i) = \frac{\sqrt{y_i}}{e_i} \tag{2.7}$$

where e_i, as known, gives a clue about the area size. To sum up, mapping the $SMRs$ (sometimes called crude rates) will actually highlight high risks only in small areas. Disease mapping models address such problems allowing us to work out adjusted estimates of the true relative risk values, smoothed with respect to the $SMRs$.

We want to attempt to set a hypothesis test for each area (region) of the map. In literature we can find a few examples that underly the idea of testing hypotheses on the collected $SMRs$ rather than pursuing the inferential goal of estimating the relative risk values. As an example, Cressie (1993), quoting an earlier work of Choynowski (1959), considers a map of p-values of the form:

$$p_i = \begin{cases} 1 - P\left(Y_i \leq y_i \mid H_{0i} : Y_i \sim Poisson(e_i)\right) & e_i \leq y_i \\ P\left(Y_i < y_i \mid H_{0i} : Y_i \sim Poisson(e_i)\right) & e_i \geq y_i \end{cases}$$

where recall y_i and e_i are both known values; the former being empirically observed and the latter being computed by introducing assumptions discussed in section 2.3. By Analyzing the North Carolina Sudden Infant Death Syndrome (SIDS) dataset, where for 100 counties counts of numbers of live births and numbers of sudden infant deaths are available, Cressie, as regards the p-values computed as above, observed as "an extreme value ... may be more due to its lack of fitting to Poisson model than to its deviation from the constant rate assumption". Recall that the constant rate assumption in our setting corresponds to the null hypothesis (2.5).

In Shlattmann (1993) there is the idea of assigning area to high risk groups. The method aims to estimate the heterogeneity in relative risks, another issue that is typically pursued preliminarily to the construction of maps of disease. This is when we can reject the hypothesis that $y_i \sim Poisson(e_i \cdot r)$ in favor of the alternative $y_i \sim Poisson(e_i \cdot r_i)$, i.e. that relative risks are not all constant in the map. Shlattmann moreover discusses the multiple testing issue recognizing

its importance and noting that a Bonferroni adjustment on Poisson p-values, besides leading to a dramatic loss of power, is unable to provide a consistent estimate of the proportion of true null hypothesis either. However, the mixture modeling approach for disease mapping proposed by Schlattmann may potentially represent a useful ground for inclusion of FDR estimation.

2.3.2 *P*-values computation in small areas

Frequentist methods to control/estimate the FDR are based on the knowledge of only p-values. In order to assess the possibility to apply such methods we try to find a suitable p-value for evaluating the null hypothesis of absence of risk. Following the hypothesis testing setting in section 2.3 we could consider each ith area-specific p-value of the form:

$$p_i = 1 - P\left(Y_i \leq y_i \mid H_{0i} : r_i = 1\right) \tag{2.8}$$

Recall our interest is in an unilateral alternative hypothesis: a risk higher than what is expected. Thus, in principle, extremely small p-values should occur in areas where the observed count is unusually higher than the expected one. However, as Cressie (1993) noted evaluating the null hypothesis with such kinds of p-value can lead to troubles when the disease outcome is rare or areas are small. Moreover Mollié (1996) have correctly noted that, while mapping $SMRs$ highlights higher risk just in low-populated areas, significance maps can highlight unusual risks just in high-populated areas, i. e. where the standard error of the SMR is small.

There are two main problems: first, a p-value calculated in a small expected count area, cannot guarantee the same empiric evidence of a p-value computed in a bigger expected count area; second, it can be shown via simulation that, for the same level of risk greater than 1 (as ex. $r = 1.5$, i. e. a 50 % of risk increment compared with that expected), a p-value calculated in small expected count areas is less extreme (i.e. more conservative) than one calculated in bigger expected count areas; see Figure 2.2. To sum up, we cannot trust each p-value to the same degree (first point), and, even worse, tests evaluated with such p-value are powerless in poorly sampled areas (second point). See Figure 2.1 which shows a plot of the power against the expected count level concerning several values of the alternative hypothesis.

Moreover, note the p-value computation involves the cumulative distribution function of the Poisson random variable Y_i evaluated in the right tail beyond the observed count y_i. Consider the

following two ways to obtain a p-value:

$$p_i = 1 - P\left(Y_i \leq y_i \mid e_i\right) = 1 - \sum_{k=0}^{y_i} \frac{\exp(-e_i) \cdot e_i^k}{k!} \tag{2.9}$$

$$p_i = 1 - P\left(Y_i < y_i \mid e_i\right) = 1 - \sum_{k=0}^{y_i - 1} \frac{\exp(-e_i) \cdot e_i^k}{k!} \tag{2.10}$$

It is clear that whether or not including the observed value y_i in the summation makes a difference, but such a difference is more emphasized in case where e_i (the mean parameter of the Poisson variable Y_i) is a small value, i.e. in small areas. The small area p-values conservativeness is due to the discrete distribution of the test statistics y_i, that is indeed a count. Generally, with a discrete distribution it is not possible to construct confidence intervals with specified coverage (the probability that the confidence interval contains the parameter of interest). Thus, one typically uses confidence intervals with the nominal coverage as the lower bound for the actual coverage. This will result in conservative p-values and conservative confidence intervals, that is to say that the significance level of the test is less and the coverage probability of the confidence interval is greater than nominal. This conservativeness is stronger in small areas, where the expected value of the Poisson distribution is small; see results obtained via simulation by Kulkami (1998).

The above described is just one of the possible ways to obtain a p-value for testing a count having available an expected count as the true value under the null model. Some proposals of p-value computation for testing an SMR have involved normal approximation for the $\log(SMR)$ (Armitage, 1971; Banerjee et al., 2004) that, however, improves as long as the number of observed deaths gets larger. Other authors have suggested methods exploiting the relation between χ^2 and Poisson distribution (Ulm, 1990) to calculate an exact confidence interval and find a p-value by means of only a table of the χ^2 distribution. Also non computer intensive algorithms that refine the coverage, so achieving less conservative p-values, are available (Kulkami et al., 1998).

2.3.3 The Over-dispersion issue

Let us suppose to set up a multiple testing procedure by evaluating each area-specific test with a p-value expressed as (2.8). In the next section we will focus on problems relative to applying frequentist p-value based methods. Here we consider the over-dispersion case in relation to p-values.

In previous sections we said that it is likely to find extreme p-values in areas with a large population, conversely to mapping the $SMRs$ that would highlight just small areas because of small denominators. We also pointed out, quoting Cressie (1993), that extreme p-values may be more due to a lack of fitting of the Poisson model than to an actual deviation from the null hypothesis that the mean of the Poisson is e_i. Lack of fit of the Poisson model can be equivalently denoted as presence of an "extra-Poisson variability" or in general as "over-dispersion". Such phenomena can

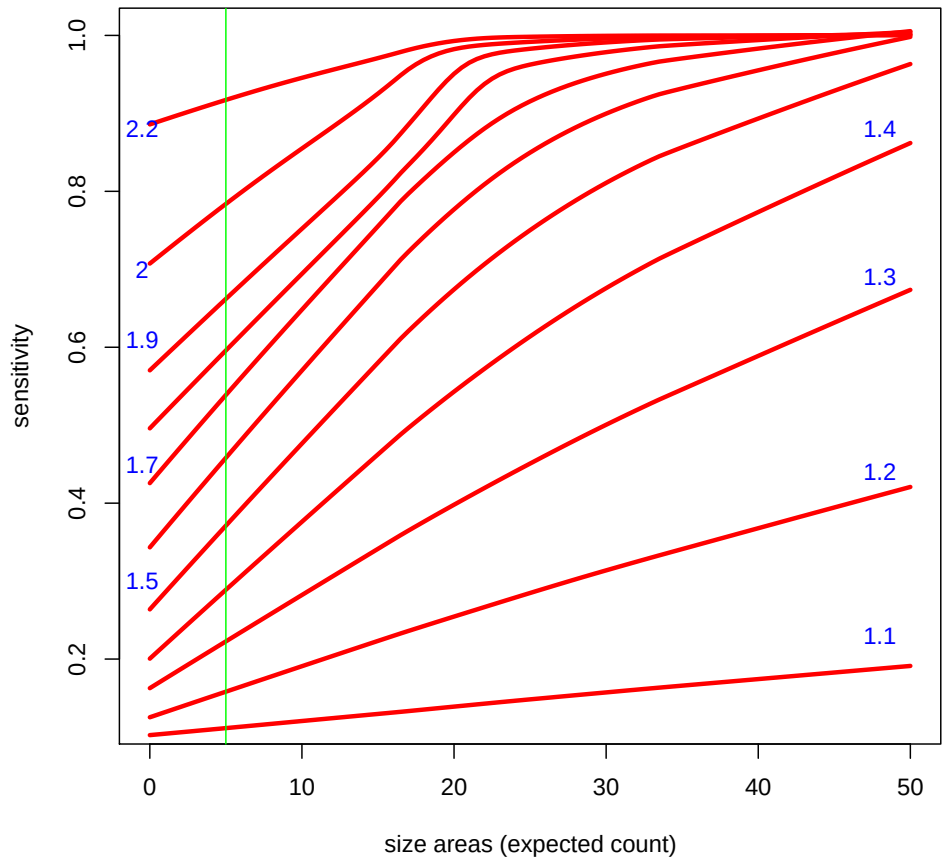


Figure 2.1: Power against size area. For each value of expected count (size area) in the horizontal axis, we see plotted the sensitivity of the test that reject the null hypothesis (2.5) when the p-value calculated with formula (2.8) is lower than 0.05. We also see that the sensitivity is generally lower in small areas case. Each line corresponds to several values (coloured blue) of the alternative hypothesis ($r > 1$) under which the power is calculated as the proportion of the times that the null hypothesis is correctly rejected. The green line is in correspondence with an expected count of 5, a limit under which we arbitrarily consider an area to be small.

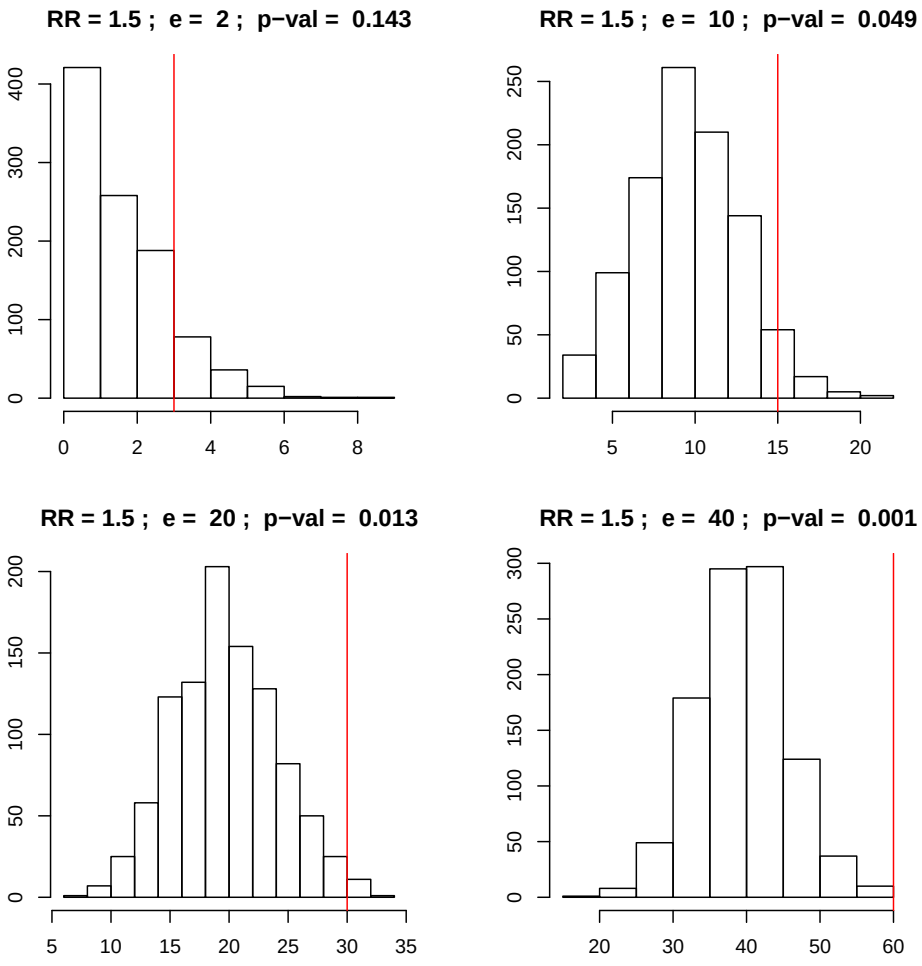


Figure 2.2: Histograms showing p-values obtained with formula (2.8) in four possible areas differing for their expected count. The expected counts are 2, 10, 20, 40 and the observed counts (red line) are respectively 3.5, 15, 30, 60 consistent with a relative risk constantly equal to 1.5. Thus, given the same value for the relative risk, bigger areas show more extreme p-values, hence yielding a more sensitive test.

be frequently encountered in the case of counts collected in small areas (Mollié, 1996; Haining et al., 2008). Generally, for rare disease and for small areas, variation in the observed number of events exceeds that expected from Poisson inference. In a given area, variation in the number of events is due partly to Poisson sampling, but also due to extra-Poisson variation arising from variability in the disease rate within the area, which result from heterogeneity in individual risk levels within the area.

Heterogeneity of the individual risks can be due to spatial interaction effects at area level. Consider the case of infectious disease modelling where one individual having the disease raises the risk of infection for individuals in the same spatial unit. The consequence is a tendency for cases to cluster so that some areas have large counts (where the disease has started and spread) and some others have small, possibly zero, counts (where the infection has not yet arrived). The same effect can be created by unobserved environmental factors that operate at area level, determining non independent risk levels within the area. The presence of a pollution source, for instance, my have a non homogeneous effect for the whole inner-area population determining non independent individual risk levels. Moreover, though the classic standardization operation aims to eliminate the effect of confounding factors like age and sex, a variety of other unmeasured factors can influence the individual response. Thus, the inner-area heterogeneity can be due to a number of unobserved variables, such as lifestyle and genetic inheritance. To sum up, individual risk heterogeneity is to be expected when dealing with aggregates, especially of non-experimental subjects (like our case where we simply collect death events), and is a source of over-dispersion. Finally, if the scale of the spatial unit used to record the data is such that one of the above factors (environmental, social, or genetic), or simply the contagiousness of the disease, could also operate between the spatial units, it may induce positive spatial correlation in the counts too. The term positive (negative) spatial correlation refers to the property of attribute measured at nearby or adjacent geographical locations having similar (dissimilar) values. Thus, we can see as the invalidity of assumptions in the model (2.1), i.e. the i.i.d assumption and the mean-variance equality, is actually due to the lack of independence between individual underlying risks of people that can operate within areas or between areas.

To understand whether or not testing $SMRs$ for screening health population status with p-values (2.8) is appropriate we should ask ourselves two things: if $SMRs$ can be appropriate for evaluating the health status of a region, and if an extreme p-value signals unusual risks. The former question could be answered, as already discussed, saying that, though we can standardize for some measurable confounders, we could never assume an $SMR > 1$ due to exposure to environment risk factors alone (any standardization carries the risk of over-simplification); thus, SMRs are

appropriate in preliminary studies where the generation of hypotheses is the prime interest, whereas to confirm them more investigation is necessary. About the second point we believe a small p-value actually signals unusual risks even if we have to admit the loss in sensitivity caused by p-values (2.8) in small areas.

Indeed, though we cannot say whether an extreme p-value for area i is due to $r_i > 1$ (higher risk in area i) instead of $V[Y_i] > E[Y_i]$ (dependent individual risks for people living in area i) we above argued that both situations are unusual from an epidemiological point of view and the practitioner would probably be interested in highlighting both. As an example, if a small p-value was due to the lack of fit of the Poisson model, i.e. was due to over-dispersion and not to a relative risk greater than one, it could even mean there is a within area cluster in area i, hence identifying a case that deserve attention by epidemiologists. Therefore, the difficulty in disentangling what can be the true alternative model if a deviation from the null model has been observed is not a big problem. However, in the next section we will argue that applying a control/estimation of the FDR with p-values based methods is not appropriate in small areas and spatial correlation cases and in section 2.4.1 claim a Bayesian paradigm is helpful.

2.4 Traditional p-value based procedures for SMR multiple testing

In the first chapter we briefly introduced p-value based control methods. We considered the Storey method to estimate $pFDR$ and the sequential procedure for p-value selection by Benjamini and Hochberg. We can meet three main problems in considering a p-value based method as appropriate, all of them being connected to the over-dispersion issue.

First, a test evaluated by calculating a p-value is powerless in small areas: consider for example the p-value (2.8), then given the same true relative risk value, say 1.5, p-values calculated in poorly populated areas will be less extreme than p-values calculated in large areas. Thus, p-values in small areas will be conservative w.r.t. the null hypothesis. This also raises the second point, that is whether or not to assume all p-values as identically distributed under the null hypothesis. Recall that Storey (2003) (see section 1.2) correctly assumes that under H_0 p-values ought to be uniformly distributed between 0 and 1. Since each p-value of the form (2.8) depends on the expected count e_i, in principle, the uniform distribution would be verified if all expected counts were equal. Changing perspective, we could realize that this is not strictly necessary, but what is required is for the Poisson model to be the true model under H_0. As we saw in section 2.3.3 this is achieved when there is no over-dispersion, and this could probably be the case when there are no small areas. Putting this

in another way, Storey's assumption of p-values uniformly distributed under H_0 may probably be tenable if all of them were computed in areas with a large population; i.e. where the sampling error variability is small for each SMR and p-values are equally informative. The latter was also checked from an empirical point of view, by a small simulation study here not reported. We noticed a bigger mass of probability around 1 (i.e. far from values representing extreme p-values) in the histogram of some p-values (computed by formula (2.8)) generated by having simulated counts under the null hypothesis ($Y_i \sim Poisson(e_i)$) using an heterogeneous expected count pattern $\boldsymbol{e} = (e_1, ..., e_N)$ containing even small values. Also Gòmez-Rubio et al. (2005) noticed the same phenomena in the empirical p-values distribution relative to data collected in small areas.

The third and final consideration is that p-values cannot be assumed as independent, an assumption on which most Frequentist methods rely on. The non independent tests issue is related to the lack of fit of the i.i.d. Poisson model assumption that occurs (as argued in section 2.3.3) when both environmental and non environmental factors determine a positive spatial correlation between area-specific risks.

Therefore, all reasons that lead to over-dispersion in counts are what does not allow for a multiple testing procedure based only on p-value knowledge. Note that the lack of fit of the i.i.d. Poisson model was what inspired many authors to develop disease mapping models mostly following the Bayesian paradigm. We hence decided to exploit the same methodology implemented in disease mapping methods so attempting to include the estimation of FDR in a classical fully Bayesian disease mapping model, namely the Besag York Mollièe model (Besag et al., 1991).

2.4.1 Additional remarks on multiple testing issue and over-dispersion

As regards the relation between the need for a multiple testing control and the lack of reliability of p-values it is worth recalling that some authors deny, in principle, the adoption of any multiple testing control (Rothman, 1990), while others consider the Bayesian hierarchical modelling paradigm fruitful in cases when over-dispersion arises and more flexible ways to allow for random effects capturing the unobserved variability are needed (Greenland and Robins, 1991). In fact, as noticed in our case study, in some context traditional frequentist methods are inappropriate since the set of estimates (or p-values) under the null model are affected by heterogenous standard errors (i.e p-values cannot be assumed as i.i.d. from a $Uniform(0, 1)$ distribution).

As a further reflection we would like to mention the work by Greenland and Robins (1991) who discussed the usefulness of an empirical Bayes adjustment in some application involving $SMRs$ for administrative resource allocation. Here the objective was basically to allocate resource proportionally to a set of risk estimates relative to different spatial regions. We have to say that authors

do not treat the problem of doing tests on observed $SMRs$ as we want to do but the work is interesting for us because, as we have already claimed, problems affecting small area $SMRs$ are the same that cause troubles on evaluating the null hypothesis (2.5) in small areas themselves. Authors explicitly examine the small areas issue, hence the sampling error variability over the map that generates non identically distributed $SMRs$. He mentions two motivations for pursuing an adjustment for the ML estimates (the $SMRs$). First, when "chance (sampling error) not only can cause the unusual findings in principle, but it does cause many or most such findings". This is exactly our case; let us analyze this point both regarding the goal of estimating the true relative risks or testing a multiplicity of null hypotheses. In the former case, some small area $SMRs$ could be identified as high-risk areas only because of the small sample size. In the latter, p-values can detect extremes only when the sample size is enough for the observed statistics (SMR or observed count) to reject the null hypothesis. In both cases the sampling variability is mainly responsible for unusual findings. Note that what we denote as sampling variability is actually the variability of the population underlying the map, hence what causes unusual findings is actually the extra-Poisson variability and not the variability of the true $SMRs$. Therefore, authors deny the adoption of multiple comparison adjustments when unusual findings are those caused by variability of the true $SMRs$, instead they advocate an adjustment when such unusual findings are "caused" by an extra-variability that the model describing true SMRs variation cannot account for.

A second motivation mentioned by the authors for making it necessary to control the multiple comparisons issue was, when "no one would want to earmark for further investigation something caused by chance". This is our case as well: we want to pursue the control/estimation of the FDR determined by Poisson variability as long as we want to control/estimate the FDR caused by the extra-Poisson variability. Authors do not mention any frequentist procedures for adjusting the SMR estimates just because such cases are only "adjustable" with the help of the Bayesian paradigm. Frequentist p-value based procedures fail because:

- p-values do not have the same empirical strength;
- p-values are conservative w.r.t. the null hypothesis in small areas;
- p-values can be spatially correlated.

The author looks at the empirical Bayes estimation as a direct way to adjust the error due to sampling variability, for both the Poisson and the extra-Poisson sources. A shrinkage estimation partially overcomes the problem of testing many observed $SMRs$ which are not identically distributed under the null hypothesis. Empirical Bayes estimators result from a compromise between what is empirically observed and what is subjectively assumed a priori about the true $SMRs$

distribution. The simplest case is the one of exchangeable assumed relative risks r_i given some hyper-parameters; for instance a poisson gamma model where each area-specific relative risk is gamma distributed with fixed hyper-parameters estimated directly by the data (assuming counts distributed as a negative binomial). Since relative risk values over the map have, a priori, a common distribution, the shrinkage effect leads to posterior estimates of the relative risks that are smoothed towards a global mean.

The argument the author uses to justify the appropriateness of the shrinkage estimation is simple. The distribution of the observed $SMRs$ in a typical case of data sparseness (small area case) is the sum of two distributions: the distribution of the true SMRs (assumed a priori both by Frequentists and Bayesians even if for frequentist it is a degenerate distribution because each SMR is unknown but fixed), and the distribution of sampling errors. This framework raises a two stage sampling model since, firstly, nature "samples" the true $SMRs$ from a true-$SMRs$ distribution; secondly, the statistician takes a sample to estimate these true $SMRs$. This is also a model for the over-dispersion since we are assuming the variance of the observed distribution is the sum of the variance of the true distribution (the model) and the average variance of the sampling error. Under such assumptions if the statistician observes extreme values he will be more inclined to think this is due to the sampling error rather than due to an extreme value under the true SMRs distribution. Note, this is exactly the same point argued by Cressie and others authors in interpreting maps of $SMRs$. Authors discuss this by making an interesting analogy with the well known "regression to the mean" phenomenon occurring in taking measures on people whose true values have a bell-shaped distribution. When subjects are sampled from a population in which true values have such a symmetric distribution, but the values are measured with error, extreme measured values are likely to be the product of extreme errors. So, if a subject has an extreme value on the first measurement, it will probably regress toward the population mean upon the following observations. We can view the true and the observed $SMRs$ distribution as representing a distribution before and after misclassification due to sampling error.

Moreover, shrinkage estimation can produces estimates (the posterior means of the relative risk, that is the adjusted SMR) that are optimal under a squared error loss function (Carlin and Louis, 2000). The above considerations are the reasons why the authors mention a multiple comparisons problem as an opportunity, rather than a problem, for the practitioner close to the Bayesian perspective, because he can improve estimates through "judicious use of any prior information (in the form of model assumptions) about the ensemble of parameters being estimated". In some sense shrinkage estimation allows an adjustment for multiple inference in cases where the null model does not fit the data properly, when data are over-dispersed. From the multiple testing setting point of

view established in section 2.3 we need the shrinkage estimation to control false discoveries caused by extra-Poisson variability other than Poisson variability. This is another way to understand the main reasons why we need the Bayesian paradigm. The model we shall propose in chapter 3 will attempt to achieve shrinkage estimation of the posterior probability of the null hypothesis smoothing such values toward both a global and a local (neighborhood) mean. A Bernoullian random variable indicating the the null hypothesis is true has to be introduced in the hierarchical model and its posterior mean (i. e. the posterior probability of the null hypothesis (1.8)) will became the target of inference for making decisions.

Moreover, in principle, making a decision should consider more than the likelihood model assumed for describing the data. Decisions should be based not only on observed test statistics, or on point estimates, but weighted in terms of pre-defined loss functions. Examples of such a decision theoretical approach are rare in spatial epidemiological applications. An alternative idea regarding the issue of determining decision rules for selecting high-risk areas connected to FDR estimation will be discussed in section 3.3.

Chapter 3

A Bayesian Hierarchical model for False Discovery Rate estimation

3.1 Bayesian disease mapping

The methodology proposed in this chapter and illustrated in section 3.2.1 to overcome the inappropriateness of p-value based methods has the same foundation used in Bayesian hierarchical disease mapping models. For this reason we give a brief review of such methods.

Extra-Poisson variation is due to heterogeneity of individual risks within each area hence it can be accommodated by allowing each relative risk r_i to vary within the area i itself. Bayesian methods can be used for this, giving smoothed posterior estimates of relative risks. Moreover they allow for the introduction of spatial random effects, making it possible to address cases where positive spatial correlation is expected.

Bayesian models in this context combine two types of information: the information provided for each area by the observed counts which are usually assumed as independent and identically distributed as the Poisson (2.2) so to working out the $SMRs$ as ML estimates, and prior information on the relative risks $\boldsymbol{r}$ specifying their variability in the map by means of the prior distribution $[\boldsymbol{r}]$. Here, the term $\boldsymbol{r}$ in bold means it is a parameter vector with N components each one indicating the relative risk in each single area, $\boldsymbol{r} = (r_1, ..., r_N)$, whereas the notation $[\cdot]$ as usual indicates a distribution function.

The prior distribution $[\boldsymbol{r}]$ reflects prior belief about variation in relative risks over the whole map and is often parameterized by hyperparameter γ. Introducing a prior distribution for $\boldsymbol{r}$ means moving from an i.i.d assumption on counts (like in model 2.1) to an exchangeability assumption on counts, that is to say the y_i are conditionally independent given $\boldsymbol{r}$, and y_i depends only on r_i. Nevertheless, in both i.i.d and exchangeable case, the likelihood function of the relative risks $\boldsymbol{r}$

conditional on the data $\boldsymbol{y}$, it is given by the product of N independent Poisson distributions:

$$[\boldsymbol{y}|\boldsymbol{r}] = \prod_{i=1}^{N} [y_i|r_i]$$

However, what can change is the way to make inference on the unknown relative risks $\boldsymbol{r}$. Frequentist inference on $\boldsymbol{r}$ is based on the likelihood function $[\boldsymbol{y}|\boldsymbol{r}]$ where $\boldsymbol{r}$ is seen as an unknown fixed parameter. It is found through the maximization of the likelihood function, yielding, in such particular example, SMR_i as the ML estimates of the ith area specific relative risk. Statistical properties of such an estimator are worked out by considering its distribution over the sample space, hence considering hypothetical repeated experiments. Bayesian inference, instead, is based on the posterior distribution of $\boldsymbol{r}$ given the data $\boldsymbol{y}$:

$$[\boldsymbol{r}|\boldsymbol{y}] \propto [\boldsymbol{y}|\boldsymbol{r}][\boldsymbol{r}]$$

where $\boldsymbol{r}$ is treated as a random variable. The term $\boldsymbol{r}$ is often parameterized by an hyperparameter γ in turn distributed with $[\gamma]$, yielding the following conjoint posterior

$$[\boldsymbol{r}, \gamma|\boldsymbol{y}] \propto [\boldsymbol{y}|\boldsymbol{r}][\boldsymbol{r}|\gamma][\gamma].$$

Thus, the marginal posterior distribution for $\boldsymbol{r}$ given the data $\boldsymbol{y}$ is generically expressed as:

$$[\boldsymbol{r}|\boldsymbol{y}] = \int [\boldsymbol{r}, \gamma|\boldsymbol{y}]\, d\gamma. \tag{3.1}$$

In principle, if $[\boldsymbol{r}|\boldsymbol{y}]$ is known, one can compute each moment of the distribution of the risks by integral calculation (in case where it is tractable); for instance, a point estimate of the set of relative risks can be provided by computing the posterior mean

$$E[\boldsymbol{r}|\boldsymbol{y}] = \int \boldsymbol{r} \cdot [\boldsymbol{r}|\boldsymbol{y}]\, dr$$

Unfortunately, in many non trivial cases posterior distributions are not analytically tractable; moments of the distribution are not available in closed form.

To sum up, considering $\boldsymbol{r}$ a multivariate random variable (and specifying its probability distribution as dependent on the hyper-parameter γ) rather than a fixed unknown vectorial parameter distinguishes the Bayesian from the Frequentist paradigm. Moreover, the way in which prior beliefs on hyperparameter γ are specified denotes different approaches developed in the framework of Bayesian statistics. An important distinction is between the empirical Bayes (EB) and the fully Bayesian approaches. The EB approach assumes the hyperparameter γ as known and drawn from an unspecified distribution. The EB idea consists in approximating (3.1) by:

$$[\boldsymbol{r}|\boldsymbol{y}, \widehat{\gamma}] \propto [\boldsymbol{y}|r] \cdot [\boldsymbol{r}|\widehat{\gamma}]$$

where the unknown hyperparameter γ is replaced by an estimate $\widehat{\gamma}$ often worked out as ML estimate from the marginal likelihood of γ:

$$[\boldsymbol{y}|\gamma] = \int [\boldsymbol{y}|\boldsymbol{r}] \cdot [\boldsymbol{r}|\gamma]\, d\boldsymbol{r}$$

A point estimate of the relative risk parameter $\boldsymbol{r}$ can be provided by the posterior mean $E[\boldsymbol{r}|\boldsymbol{y}, \widehat{\gamma}]$. As a result of the plug-in estimation of γ the variability in $\boldsymbol{r}$ is under-estimated because no allowance is made for uncertainty in γ.

The fully Bayesian approach can give a suitable solution to incorporate variability in the hyperparameter γ. A three-stage hierarchical model can be introduced where the hyperprior distribution $[\gamma]$ is also specified. In such complex models, moments from the posterior distribution can be estimated with Markov Chain Monte Carlo (MCMC) algorithms. They allow to draw non-independent samples of each parameter as result of a realization of a Markov chain whose equilibrium distribution is the posterior distribution of interest (Mollié, 1996). The model proposed in section 3.2.1 will be estimated by means of MCMC algorithms available in OpenBugs free software.

Much work has been done about the prior model choice of the second level hierarchy, that is the prior specification of risks $\boldsymbol{r}$, while maintaining the parametrization (2.1) for counts $\boldsymbol{y}$. A Bayesian approach, as many authors observed (Lawson et al., 1999), is appealing because it is more flexible addressing spatial dependence and small area issues. However, other non-Bayesian ways to develop disease mapping models have been explored; for a review see Lawson et al. (2000), Best et al. (2005).

3.1.1 Independent prior

When there is no prior information to assume positive spatially correlated risks we can specify a prior for $\boldsymbol{r}$ which assumes a spatial unstructured heterogeneity by considering exchangeable risks given γ:

$$[\boldsymbol{r}|\gamma] = \prod_{i=1}^{N} [r_i|\gamma].$$

In this case the prior distribution $[r_i|\gamma]$ is the same for each area i. A suitable specification for such prior is given by the Gamma distribution of parameters α and ν, that is the natural conjugate distribution for the Poisson; i.e. $[r_i|\gamma] = Ga(\alpha, \nu) = \alpha^{\nu} \cdot r_i^{\nu-1} \exp(-\alpha \cdot r_i)$ of mean $\frac{\nu}{\alpha}$ and variance $\frac{\nu}{\alpha^2}$. Therefore:

$$[\boldsymbol{r}|\gamma] = [\boldsymbol{r}|\alpha, \nu] = \prod_{i=1}^{N} [r_i|\alpha, \nu] \tag{3.2}$$

With such a conjugate Gamma prior for relative risks $\boldsymbol{r}$, the marginal posterior distribution $[\boldsymbol{r}|\boldsymbol{y}, \gamma]$ is the product of N marginal posterior Gamma distributions, $Ga(y_i + \nu, e_i + \alpha)$, each one

having mean:

$$E[r_i|, y_i, \alpha, \nu] = \frac{y_i + \nu}{e_i + \alpha} = \kappa_i \cdot SMR_i + (1 - \kappa_i) \cdot \frac{\nu}{\alpha} \tag{3.3}$$

where $\kappa_i = \frac{e_i}{e_i+\alpha}$. Thus, we can estimate the posterior relative risk in area i as a weighted average of the SMR_i and the prior mean of the relative risks in the whole map, the weight being inversely related to the variance of the SMR_i. We can see that if an SMR is calculated in a very small area, hence having a big standard error (2.7), this will give a weak contribute to the estimator (3.3). Prior specification of unknown parameters α and ν can be done following the EB or the fully Bayes approach. The former proceeds by working out the ML estimates, $\widehat{\alpha}$ and $\widehat{\nu}$, from the marginal likelihood $[\boldsymbol{y}|\alpha, \nu]$, which is a product of N negative binomial distributions $[y_i|\alpha, \nu]$. Then, an estimate of the ith posterior relative risk, $\widehat{r_i}$, can be provided by calculating $E[r_i|y_i, \widehat{\alpha}, \widehat{\nu}]$, plugging in the ML estimates $\widehat{\alpha}$ and $\widehat{\nu}$. Alternatively we can draw a sample from the posterior distribution $Ga(y_i + \widehat{\nu}, e_i + \widehat{\alpha})$ applying sampling routines available in statistical software packages and computing any Monte Carlo summary statistics of interest in the drawn sample. Differently from the EB approach, the fully Bayes approach proceeds by specifying an hyperprior distribution (even a degenerate distribution, i.e. a scalar value) for hyperparameter γ then working out posterior estimates for risks by sampling from the joint posterior distribution $[\boldsymbol{r}, \gamma|\boldsymbol{y}]$. For complex models, implementing Gibbs sampling algorithms is needed to obtain samples from the posterior distribution of the parameters of interest.

An alternative independent prior specification for relative risks consists in assuming a normal distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2 on the logarithmic transformation of the relative risks, say $\boldsymbol{x} = \log \boldsymbol{r}$. Analogously to the prior specification (3.2) we have:

$$[\boldsymbol{x}|\gamma] = [\boldsymbol{x}|\mu, \sigma^2] = \prod_{i=1}^{N} [x_i|\mu, \sigma^2] \tag{3.4}$$

To allow for area specific covariates such an independent normal prior can be generalized by setting $\mu = Z\beta$, where Z is a matrix of p known covariates and $\beta = (\beta_1, ..., \beta_p)$ is a vector of covariate effects. As a result of this parametrization, the prior mean of the log relative risks $\boldsymbol{x}$ is not constant across areas, the posterior distribution being the product of N independent but not identical normal distributions:

$$[\boldsymbol{x}|\gamma] = [\boldsymbol{x}|\mu, \sigma^2] = \prod_{i=1}^{N} [x_i|\mu_i, \sigma^2]$$

where $\mu_i = (Z\beta)_i$. Normal priors allow more easily than the conjugate gamma prior for dependence between components of vector $\boldsymbol{r}$. This is useful when positive spatial correlation between risk is a priori expected.

3.1.2 Spatially structured prior

Assuming a spatially structured prior distribution for $\boldsymbol{r}$ means to take into account that geographically closed areas tend to have similar relative risks. To express in mathematical terms a local spatial variation of risks, nearest neighbour Markov Random Fields (MRF) can be useful:

$$[x_i|x_j, j \neq i] = [x_i|x_j, j \in \delta_i] \tag{3.5}$$

Assuming the conditionally specification (3.5) for parameters $\boldsymbol{x} = \log \boldsymbol{r}$ means assuming that the conditional distribution of the log relative risk in area i, given values for the log relative risks in all other areas $j \neq i$, depends only on the log relative risks in the neighbouring areas (denoted as δ_i) of area i. The joint distribution of the log relative risks can be determined, up to a normalizing constant, from the knowledge of each conditional distribution (3.5) by applying *Brook's Lemma* (Besag, 1974). Moreover, the *Hammersley-Clifford Theorem* shows that if we have a MRF, i.e. if a set of full conditionals defines a unique joint distribution, then this joint distribution is a Gibbs distribution. Informally, $[x_1, ..., x_N]$, is a Gibbs distribution if it is a function of the x_i only through a function of those x_j which belong to the set of the neighbouring areas of area i ($j \in \delta_i$). Specifying the prior model for $\boldsymbol{r}$ by a set of full conditional distributions such that the joint distribution is uniquely determined as a Gibbs distribution allows to make posterior inference by implementing Gibbs sampler algorithm. It is thus possible to simulate realizations from the joint posterior distribution of the log relative risks by simulating from each full conditional separately, still being sure that there is a unique equilibrium distribution for this sampler; see Banerjee et al. (2004) and references therein for more theoretical details.

A very useful prior specification for $\boldsymbol{x} = log(\boldsymbol{r})$ is the intrinsic Autoregressive model (IAR or intrinsic CAR).

$$[x_i|x_{-i}, \sigma^2] = Normal(\bar{x}_i, \frac{\sigma^2}{w_{i+}}) \tag{3.6}$$

where $\bar{x}_i = \sum_{j \in \delta_i} \frac{x_i}{w_{i+}}$ denotes the mean of the x_j in areas adjacent to area i, and x_{-i} denotes the log relative risks in all the areas $j \neq i$.

Therefore, the conditional prior distribution of x_i, given all the other log relative risks in the map, is assumed normal with mean the average of the x_j in the neighbouring areas and variance inversely proportional to the number of neighbouring areas (denoted as w_{i+}). This model differs from the proper conditional autoregressive model (CAR) where the conditional variance for x_i given all the other log relative risks is constant. CAR is suitable for regular maps, whereas IAR is more appropriate for irregular maps, i.e. where the number of neighbors varies. Model (3.6) identify the following joint prior distribution for $\boldsymbol{x}$ given the hyperparameter γ (here $\gamma = \sigma^2$ since it is the

variance of a normal distribution):

$$[\boldsymbol{x}|\gamma] = [\boldsymbol{x}|\sigma^2] \propto \frac{1}{\sigma^n} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j<i} w_{ij}(x_i - x_j)^2\} \tag{3.7}$$

This is a Gaussian MRF where the mean is zero and its precision matrix has diagonal elements $\frac{w_{i+}}{\sigma^2}$ and off-diagonal elements $\frac{-w_{ij}}{\sigma^2}$, w_{ij} are pre-chosen non-negative weights, with (in the simplest case) $w_{ij} = 1$ if i and j are neighbouring areas, $w_{ij} = 0$ for the remaining areas and $w_{i+} = \sum_{j=1}^{N} w_{ij}$. A prior of the form (3.7) is a pairwise difference distribution and it is not proper, i.e. its integral is not finite, hence the mean and other moments of such a distribution cannot be determined. The impropriety is also evident since we can add any constant to all of the log relative risks x_i and (3.7) is unaffected. Constraining the set of the x_i's to sum to zero can solve the problem. Thus, such a model can never be taken as a model for describing data, since data could not arise under an improper density function and yet we could not impose a centering constraint on random realizations. Model (3.7) can however be assumed as a prior for parameters in the model that play the role of random effects. In our case, we chose this model for specifying the distribution of area-specific log relative risks $\boldsymbol{x} = (x_1, ..., x_N)$. This choice is appropriate and yields a posterior distribution $[\boldsymbol{x}|\boldsymbol{y}]$ that is proper (Mollié, 1996). However, for the identification of posterior log relative risks the impropriety of (3.7) cause troubles; such parameters can be identified only up to an additive constant. Thus, it is convenient to introduce an intercept α such that $\boldsymbol{x} = \alpha + \boldsymbol{u}$ can be identified by imposing the constraint $\sum_{i=1}^{N} u_i = 0$. Indeed, constraining the random effects to sum to zero and specifying a separate intercept term with a uniform prior on the whole real line is equivalent to the unconstrained parameterisation with no separate intercept (Besag and Kooperberg, 1995). Note, in implementing Gibbs sampler this constraint can be imposed numerically by recentring each sampled $\boldsymbol{u}$ vector around its own mean in each MCMC iteration. OpenBugs free software can automatically impose such a sum-to-zero constraint. To sum up, a specification of the spatially structured prior above as a CAR (or IAR) is usually proposed and it is fruitful since model (3.7) is also a Gibbs distribution, precisely a distribution for x_i which depends only on neighbouring areas log relative risks. Working with the N full conditionals (3.6) is better than seeking to write down the joint distribution for several reasons. First, the possibly large number of areal units, second, and most important in practice, it has the advantage of developing MCMC computation by implementing Gibbs sampler algorithm to sample realizations of each separate log-relative risk from its full conditional distribution. Furthermore, a local specification where the risk of area i is dependent on risk of its neighbors is a natural prior belief in many applications.

3.1.3 The Besag York Molliè (*BYM*) model

Besag et al. (1991) proposed a prior distribution for the log relative risks that ranges from prior independence to prior local dependence, called a convolution Gaussian prior. In this model log relative risks are the sum of two independent components:

$$\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$$

where $\boldsymbol{v} = (v_1, ..., v_N)$ is an independent normal variable with zero mean vector and variance ${\lambda_v}^2$ as in prior specification (3.4), whereas $\boldsymbol{u} = (u_1, ..., u_N)$ is assumed to be a IAR model as in specification (3.7) with each conditional variance equal to $\frac{\lambda_u^2}{w_{i+}}$. Thus, random effects $\boldsymbol{v}$ capture extra-Poisson variability in the log relative risks that vary globally (*heterogeneity*), i.e. over the entire study region, while random effects $\boldsymbol{u}$ model extra-Poisson heterogeneity in the log relative risks that varies locally (*clustering*), so that nearby regions will have more similar rates. Each x_i conditional variance is the sum of the conditional variance of the term u_i and of the marginal variance of the term v_i:

$$V[x_i|x_{-i}, \gamma] = V[x_i|x_j, j \in \delta_i, \lambda_u^2, \lambda_v^2] = \frac{\lambda_u^2}{w_{i+}} + \lambda_v^2$$

Prior choice for variance terms λ_u^2 and λ_v^2, or alternatively for precisions $\tau_u = \frac{1}{\lambda_u^2}$ and $\tau_v = \frac{1}{\lambda_v^2}$ is a critical point discussed by many authors; see Banerjee et al. (2004), Mollié (1996) and references therein. The value of prior precisions τ_u and τ_v will control the amount of extra-Poisson variability allocated to the two components of the log relative risk, the "clustering" term u_i and the "heterogeneity" term v_i. Usually a non informative prior is desired, but for computation reasons (for example if OpenBugs is used for drawing posterior inference) it still need to be a proper distribution. Moreover, if the prior beliefs are expressed as fixed values, they obviously cannot be chosen as arbitrarily small, as $x_i = u_i + v_i$ would be unidentifiable: a small prior precision, i.e. large prior variance, will cause small convergence of MCMC algorithms. On this note, see Eberley and Carlin (2000) where an investigation of the convergence for posteriors $[x_i|\boldsymbol{y}]$, $[u_i|\boldsymbol{y}]$ and $[v_i|\boldsymbol{y}]$ in the Besag York Mollié model is performed by using fixed τ_u and τ_v values. In the case we decide to specify third-stage prior distributions on τ_u and τ_v, instead of fixed values, such prior cannot be arbitrarily vague for the same identifiability problem. A typical choice in this case is the conjugate Gamma family distribution; Kelsall and Wakefield (1999), for instance, suggested a $Gamma(0.5, 0.0005)$ for the precision parameter τ_u of the spatial random effects (recall $\boldsymbol{u} \sim CAR(\tau_u)$). Gelman (2005) noted the strong sensitivity of inferences to the low values of the Gamma parameters when the standard deviation is estimated near zero; in our case study it would occur when log relative risks are quite homogeneous. For a non informative specification Gelman

recommends a uniform distribution with finite range (for example the interval (0,100)) on the prior standard deviation of the random effects, $\frac{1}{\sqrt{\tau_u}}$ and $\frac{1}{\sqrt{\tau_v}}$.

However, it has to be recognized the difficulty of finding a way to specify a "fair" prior for such two precisions, i.e. equal prior emphasis on clustering and heterogeneity terms. To this aim, specifying prior Gamma distributions with equal parameters for both τ_u and τ_v is not correct since, in the former case, the precision is specified conditionally (before playing the role of the conditional prior precision τ_u has to be multiplied by the number of neighbour ω_{i+} that varies among the regions), whereas in the latter τ_v is specified marginally (see Banerjee et al, 2004). To address this critical issue a proposal was advanced in a work by Bernardinelli et al. (1995); the authors noted that the prior marginal standard deviation of v_i is approximately proportional to the prior conditional standard deviation of u_i:

$$sd(v_i) = \frac{1}{\tau_v} \approx \frac{1}{0.7 \cdot \sqrt{\bar{m} \cdot \tau_u}} \approx sd(u_i) \tag{3.8}$$

In conclusion, if a non informative prior is sought, checking sensitivity for different prior specifications of parameters τ_u and τ_v in any real case study is always recommended.

As regards inference for such a model, posterior log relative risks can be estimated using MCMC computation by sampling at each iteration (of the Gibbs sampler algorithm) a realization from the posterior distribution of both $\boldsymbol{u}$ and $\boldsymbol{v}$ and then sum them to obtain a realization from the log relative risk posterior distribution $[\boldsymbol{x}|\boldsymbol{y}]$. As said in the previous section, since the u_i's are specified conditionally, a sum-to-zero constraint is needed for their identifiability, hence an intercept α with a diffuse prior on the real line is usually introduced. In practice Openbugs free software is an useful tool since it implements the required algorithms for the BYM model parameters to be estimated, especially the algorithms for sampling from the full conditionals relative to parameters a priori specified as an intrinsic CAR model.

Besag York and Mollié model (hereafter BYM) has been implemented in many practical small areas disease mapping exercises for its flexibility to capture variability caused by many possible unobserved factors through the *heterogeneity* and the *clustering* terms. Some authors have studied sensitivity of the model to several hyperprior specifications; useful comments and further bibliographic references on this issue can be found in Banerjee et al. (2004) and in Molliè (1996). Others compared BYM with a range of spatial models for relative risks estimation with respect to goodness of fit to simulated data derived by a range of models (Lawson et al., 2000), or with respect to the amount of smoothing of the risk actually performed (Richardson et al., 2004). As regards the goodness of fit issue, the BYM model was found by Lawson as being the most robust model (as well as the Gamma-Poisson exchangeable model here described in (3.2)) across a range of diverse models; instead, for instance, mixture models as in Schlattmann et al. (1993) and non-parametric

smoothing methods (not described in this work) performed poorly in this sense.

As regards the degree of smoothing of Bayesian models it is worth noting that in general Bayes procedures offer a tradeoff between bias and variance reduction of the relative risk estimates. Particularly in maps where the sample size is small (expected counts are small), they provide a set of point estimates with good properties in terms of minimizing squared error loss (Carlin et al., 2000). This variance reduction is attained through borrowing information due to the adopted hierarchical structure, leading to estimates shrunk towards a global mean, or in general towards a value related to the distribution of all the units included in the hierarchical structure. Thus, the effect of shrinkage depends on the prior distribution chosen for $\boldsymbol{r}$ and it is conditional on such a prior belief being close to the unknown true model for $\boldsymbol{r}$. If we have appropriate prior information and we can express it in parametric form in the hierarchical prior structure, we are able to depict at best the true pattern of relative risk values.

3.2 FDR estimation through posterior probabilities

Storey's result (1.6), exposed in chapter 1 can be synthesized as:

$$pFDR = P(H_0 \text{ is true}| \text{ reject } H_0),$$

where $pFDR$ is in principle equivalent to FDR; see Storey (2003) for a full discussion of their differences. The above result derives from assuming counts $\boldsymbol{y}$, or a function of counts, or even p-values, distributed as a two components mixture of the null and alternative hypothesis distributions. Starting from such an assumption, authors in the field of microarray analysis (Newton et al., 2004; Broet et al., 2004) proposed Bayesian methods where an estimate for FDR is provided by means of a posterior probability conditional on the data:

$$FDR_{Bayes} = P(H_0 \text{ is true}| \text{ reject } H_0, \text{ data}). \tag{3.9}$$

We can use the above posterior probability to make inference about FDR since the posterior probability of the null hypothesis given the observed data provides a posterior estimate of the *type I* error probability, that is the FDR in the trivial case where we test only one null hypothesis.

More interestingly, in a set of N tested null hypotheses, we compute an estimate of each posterior probabilities π_i of the form (1.8) where as usual $i = 1, ..., N$ indexes the regions in the map under study. Then, an estimate of the FDR, which we incur rejecting a given set of null hypotheses, is provided by the average of all the estimated posterior probabilities referred to the areas where we reject the null hypothesis. In other words, given any set of areas declared at high-risk (areas where H_0 is rejected) we can estimate the proportion of "false alarms" by averaging all declared

high-risk areas $\widehat{\pi}_i$ (i.e. the estimated posterior probability that the area i is at null-risk). To decide which areas should be considered at high-risk we need to fix a decision rule; in our case we want to focus on rules that are a function of posterior probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ and a threshold t_π (a value between 0 and 1) that serves as a cut-off for the π_i's. This proposal differs from the traditional frequentist methods for controlling/estimating FDR, that determines rule for rejecting null hypotheses finding a cut-off value for p-values.

We will talk formally of FDR estimation and decision rules for selecting high-risk areas after discussing the proposed model for estimating the π_i's. It is worth, however, clarifying in advance the relation between posterior probabilities π_i's and FDR. Posterior probabilities that H_{0i} is true will be calculated for each area i, and according to the Bayesian perspective they are conditional on the observed data, H_0 playing the role of a Bernoulli random variable. The FDR is conditional on a set of posterior probabilities π_i's since what requires for being calculated is just a set of individual area-specific π_i. More precisely, it can be calculated for any set of π_i's, this posing a strong difference from Storey's $pFDR$ estimation method that can be calculated for only ordered sets of p-values. Thus, the FDR estimator does not strictly need to depend on pre-determined decision rules for selecting a set of high-risk areas with their corresponding π_i's (note as formula (3.9) may be misleading in this sense since it is conditional on having rejected H_0). However, as we will see in section 3.3 determining selection rules that directly depend on FDR estimation can be interesting for epidemiologists: choosing the threshold t_π by predicting the FDR that would arise from this choice is a useful and non-arbitrary way to proceed in the selection of high-risk areas. It, in fact, allows for being aware of the error made when rejecting, that is the idea underlying any multiple testing control method.

3.2.1 Our model proposal: *BYM mix*

As said in chapter 1, Storey's method changes perspective in controlling the proportion of false discoveries. It pursues the estimation of FDR given a threshold for the set of p-value (i.e. given a rejection region), whereas the Benjamini-Hochberg sequential procedure seek to find the thresholds for the p-values in order to control a pre-specified value for the expected FDR. The Storey's method improvement results in gaining more power with respect to Benjamini-Hochberg, due to the attempt to estimate the overall probability of the null hypothesis which all test statistics (or all p-values) contribute to (see section 1.2).

The model proposed seeks the estimation of the expected FDR conditional on data. However, when dealing with small areas, a complication arises since Storey's assumption is not tenable: p-values are conservative in small areas and do not guarantee the same power all over the map as

they are not identically distributed; moreover they are not even independent if spatially correlated risks are likely to occur (see section 2.4).

The model we propose and describe below has a probability assumption on null hypotheses $\{H_{01}, ..., H_{0N}\}$ and introduces random effects such that the posterior distribution of each H_{0i} depends not only on the observed y_i but on all observed counts, especially on those in its contiguous areas δ_i. In other words, by estimating the probability that the null hypothesis is true in area i, we have a tool for performing a Bayesian test procedure that evaluates the test in area i by means of all the summary statistics in the map.

More formally we assume each null hypothesis H_{0i} in area i a Bernoulli prior distribution whose hyperparameter is distributed with the uninformative Uniform$(0, 1)$. Posterior distribution of H_{0i} will directly depend on data y_i and also, through spatial prior random effects, on data in the neighbouring areas (δ_i). We need to assume a two components mixture on each area observed count distribution, where the distribution under the alternative is a Poisson with mean equal to the sum of random effects capturing spatially structured and unstructured extra-Poisson variability. In this way we want to substitute the untenable assumption of an i.i.d. mixture on counts with the assumption of exchangeable counts, i.e. counts independent conditionally on the value of given random effects. To this aim we exploit the hierarchical structure of the *BYM* model described in section 3.1.3, together with the introduction of the parameters needed to specify the two components mixture on each log relative risk x_i. The first level of the hierarchy specifies the likelihood for counts $\boldsymbol{y}$:

$$[y_i|r_i] = \text{Poisson}(e_i \cdot r_i),$$

the second level specifies the prior distribution for $x_i = \log r_i$,

$$\begin{aligned}
&x_i = H_{0i} \cdot \mu_{0i} + (1 - H_{0i}) \cdot (\alpha + v_i + u_i) \\
&\mu_{0i} = 0 \\
&[H_{0i}|\phi_i] = \text{Bernoulli}(\phi_i) \\
&[\alpha] = \text{Uniform}(-\infty, +\infty) \\
&[v_i|\lambda_v^2] = \text{Normal}(0, \lambda_v^2) \\
&[u_i|\lambda_u^2] = \text{Normal}\left(\overline{u}_i, \tfrac{\lambda_u^2}{\omega_i +}\right)
\end{aligned}$$

where μ_{0i} is a constant equal to 0 to meaning that if the null hypothesis is true ($H_{0i} = 1$) the relative risk r_i is equal to one. This is consistent with the null hypothesis of absence of risk defined

in formula (2.5). Instead, when the alternative hypothesis is true ($H_{0i} = 0$) the log relative risk is assumed as the sum of α (a baseline risk mean), u_i (a spatially structured random effect) and v_i (a spatially unstructured random effect). Note this is not consistent with the definition of the alternative hypothesis (2.6) because it identifies a bilateral alternative. Since under the alternative hypothesis the *BYM mix* model can describe both relative risks greater and lower than 1, we obtain that a small value for π_i can denote either i as an higher-risk area or i as a lower-risk area. For this reason, a partition of the posterior probabilities π_i's between eligible and non-eligible areas to be potentially declared at high-risk is needed. A way to proceed might be selecting as eligible to be potentially declared at high-risk those areas that present an observed count greater than the expected. Therefore, for a practitioner that aims to test the alternative hypothesis of a relative risk greater than 1, as stated in (2.6), the set of π_i's to consider is that which realizes $y_i \geq e_i$.

About the hyperprior specification we pursue a fully Bayesian approach specifying an hyperprior distribution for λ_u, λ_v and $\boldsymbol{\phi} = (\phi_1, ..., \phi_N)$:

$$[\phi_i|0,1] = \text{Unif}(0,1)$$
$$[\lambda_v^2|a_v, b_v] = \text{InvGamma}(a_v, b_v)$$
$$[\lambda_u^2|a_u, b_u] = \text{InvGamma}(a_u, b_u)$$

For each hyperparameter ϕ_i, that is the prior mean for the null hypothesis H_{0i} in area i, we assume an uninformative Uniform prior on interval $(0, 1)$ is an appropriate choice since it describes a vague prior belief. As regards the choice for the prior distribution of variance parameters of $\boldsymbol{v}$ and $\boldsymbol{u}$ random effects we follow usual specification adopted for the *BYM* model (see discussion in section 3.1.3). Usually it is specified the conjugate Inverse Gamma, even if Gelman (2005) suggested the use of a uniform prior for the standard deviation $\sqrt{\lambda_v^2}$ and $\sqrt{\lambda_u^2}$. Checking sensitivity of posterior estimates for different choices of the hyperprior is advisable in any real datasets. When applying the *BYM mix* model to simulated datasets (see chapter 4) we always used a uniform prior on the range (0,100) on the standard deviation, thinking of this range as a conservative choice after checking it did not cause troubles of low convergence of the Gibbs sampler. If the Gamma family is chosen, Molliè (1996) comments about how to get a prior guess of the parameters of the InverseGamma, a_v, b_v, b_u and b_u, by considering the empirical variance of $\log(SMR)$'s of the observed data.

Our target is calculating an estimate of each posterior probability $\pi_i = P(H_{0i} = 1|data) = E(H_{0i}|data)$ for each area i log relative risk. The π_i's are estimated within the MCMC algorithm by the number of times when $H_{0i} = 1$ divided by the length of the simulation run; such estimator

is called $\widehat{\pi}_i$. Below we report the Bugs code for estimating $\pi_i = P(H_{0i} = 1|data)$ for each area i via MCMC simulation.

```
model {
for (i in 1:N){
y[i] ~ dpois(mu[i])
mu[i] <- e[i] * r[i]
log(r[i]) <- x[i]
x[i] <-H0[i] * mu.0[i] + (1 - H0[i]) * (alpha +u[i] + v[i])
mu.0[i] <- 0
H0[i] ~ dbern(phi[i])
phi[i] ~ dunif(0,1)
v[i] ~ dnorm(0, tau.v)
SMR.adj[i] <- exp((1 - H0[i]) * (alpha + u[i]+ v[i]))
}
u[1:N] ~ car.normal(adj[], weights[], num[], tau.u)
alpha ~ dflat()
for (k in 1:sumNumNeigh) { weights[k] <- 1 }
tau.v <- 1 / (sd.v * sd.v)
sd.v ~ dunif(0, 100)
tau.u <- 1 /(sd.u * sd.u)
sd.u ~ dunif(0,100)
}
```

"Flat" (in line 14) is the name Bugs uses for an improper prior on the real line. Posterior probability of H_{0i}, π_i is calculated as a Monte Carlo mean of node the H0[i] by sampling values at each MCMC iteration from its posterior and finally calculating the empirical mean. Note that, a Monte Carlo mean also for $SMR.adj[i]$ (line 11) provides an estimate of the relative risk in area i, i.e. a smoothed relative risk analogous to the one provided by the classic BYM model.

3.2.2 Full conditional distributions

The parameters of interest of the *BYM mix* model can be suitably estimated by OpenBugs free software. However we derived the full conditional distributions of *BYM mix* model for better understanding the relationships between parameters. As we said, we assume the ith log relative risk as

$$x_i = \log r_i = H_{0i}\mu_{0i} + (1 - H_{0i})(\alpha + v_i + u_i)$$

Given such assumption, the marginal likelihood of parameters $(\alpha, H_{0i}, u_i, v_i)$ is:

$$[y_i|\alpha, u_i, v_i, H_{0i}] = \frac{1}{y_i!}\exp\{-e_i\exp[(1-H_{0i})(\alpha+u_i+v_i)]\}\{e_i\exp[(1-H_{0i})(\alpha+u_i+v_i)]\}^{y_i}$$

$$= \frac{1}{y_i!}\exp\{-e_i\exp[(1-H_{0i})(\alpha+u_i+v_i)] + y_i\log(e_i) + y_i(1-H_{0i})(\alpha+u_i+v_i)\}$$

and recall the prior distributions are:

$H_{0i} \sim \text{Ber}(\phi_i)$
$[H_{0i}|\phi_i] = \phi_i^{H_{0i}}(1-\phi_i)^{1-H_{0i}}$;

$\alpha \sim \text{Unif}(-\infty, +\infty)$
$[\alpha] = const$;

$v_i \sim \text{Normal}(0, \lambda_v^2)$
$[v_i|\lambda_v^2] \propto \frac{1}{\sqrt{\lambda_v^2}}\exp\left(-\frac{v_i^2}{2\lambda_v^2}\right)$;

$u_i \sim \text{Normal}\left(\overline{u}_i, \frac{\lambda_u^2}{\omega_{i+}}\right)$
$[u_i|u_{-i}, \lambda_u^2] \propto \sqrt{\frac{\omega_{i+}}{\lambda_u^2}}\exp\left(-\frac{\omega_{i+}(u_{i+}-\overline{u}_i)^2}{2\lambda_u^2}\right)$
where $\overline{u}_i = \frac{1}{\omega_{i+}}\sum_{j\neq i} w_{ij}u_j$

$\phi_i \sim \text{Unif}(0, 1)$
$[\phi_i|0, 1] = 1$

$\lambda_v^2 \sim \text{InvGamma}(a_v, b_v)$
$[\lambda_v^2|a_v, b_v] \propto \lambda_v^{-2(a_v+1)}\exp\left(-\frac{b_v}{\lambda_v^2}\right)$

$\lambda_u^2 \sim \text{InvGamma}(a_u, b_u)$
$[\lambda_u^2|a_u, b_u] \propto \lambda_u^{-2(a_u+1)}\exp\left(-\frac{b_u}{\lambda_u^2}\right)$

Looking at the joint posterior distribution, we need to consider parameter vectors of N elements, one for each area. The term α could be not considered here since its constant density does not influence the likelihood. The joint posterior distribution is:

$$[\alpha, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r}, \lambda_u^2, \lambda_v^2, \boldsymbol{\phi}|\boldsymbol{y}] \propto \prod_{i=1}^{n} [y_i|\alpha, u_i, v_i, H_{0i}][\boldsymbol{u}|\lambda_u^2][\boldsymbol{v}|\lambda_v^2][\boldsymbol{H_0}|\boldsymbol{\phi}][\alpha][\lambda_u^2][\lambda_v^2][\boldsymbol{\phi}], \tag{3.10}$$

where, the joint likelihood is:

$[\boldsymbol{y}|\alpha, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r}] \propto$

$\propto \prod_{i=1}^{n} \frac{1}{y_i!} \exp\{-e_i \exp[(1 - H_{0i})(\alpha + u_i + v_i)] + y_i \log(e_i) + y_i[(1 - H_{0i})(\alpha + u_i + v_i)]\}$

We derive full conditionals distribution for each of the N elements of vectors $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{r}$, $\boldsymbol{\phi}$ and α. Full conditionals of λ_u^2 and λ_v^2 are the same as the Besag York Mollié model; see Mollié (1996). To derive the full conditional of a parameter, for example u_1, we need to pick out the terms in the joint posterior distribution (3.10) which involve u_1 alone. For other parameters the procedure is the same.

$[u_i|\text{all...}] \propto$

$\propto [u_i|\alpha, u_{-i}, v_i, H_{0i}, \lambda_u^2, y_i] \propto \exp\left\{y_i(1 - H_{0i})u_i - e_i \exp[(1 - H_{0i})(\alpha + u_i + v_i)] - \frac{\omega_{i+}(u_i - \overline{u})^2}{2\lambda_u^2}\right\}$

$[v_i|\alpha, u_i, H_{0i}, \lambda_v^2, y_i] \propto \exp\{y_i(1 - H_{0i})v_i - e_i \exp[(1 - H_{0i})(\alpha + u_i + v_i)] - \frac{v_i^2}{2\lambda_v^2}\}$

$[\alpha|\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{H_0}, \boldsymbol{y}] \propto \prod_{i=1}^{n} \exp\{-e_i \exp[(1 - H_{0i})(\alpha + u_i + v_i)] + y_i(1 - H_{0i})\alpha\}$

$[H_{0i}|u_i, v_i, \phi_i, y_i] \propto \exp\{y_i(1 - H_{0i})(\alpha + u_i + v_i) - e_i \exp[(1 - H_{0i})(\alpha + u_i + v_i)] + H_{0i} \log \phi_i + (1 - H_{0i}) \log(1 - \phi_i)\}$

$[\phi_i|, H_{0i}, y_i] \propto \phi_i^{H_{0i}}(1 - \phi_i)^{1 - H_{0i}}$

Note that $\boldsymbol{u}$,$\boldsymbol{v}$,$\boldsymbol{r}$,α were all a priori independent but at posterior (see full conditional expressions) they are dependent through the likelihood $[y_i|\alpha, u_i, v_i, H_{0i}]$ that involves all such parameters. For instance, the full conditional expression of H_{0i} depends on α, u_i and v_i values.

An interesting point is to look at the "Bayesian adjustment" for multiple testing as intended by Berry and Hochberg (1999). To show how it works in this model it is sufficient noting that the full conditional of each null hypothesis H_{0i} depends on α, i.e. the overall mean of log relative risks, that is a value which all counts $\boldsymbol{y}$ contribute to. The full conditional of each H_{0i} moreover depends

on all counts through dependence on terms λ_v^2 and λ_u^2 which all observations contribute to. But, and most important, we have the H_{0i} full conditional dependence on counts in neighboring areas through u_i dependence on $\bar{u}$ (that is the mean of random terms u_j where $j \in \delta_i$, the neighborhood of area i). Therefore, counts all together give information about each H_{0i}, the neighbouring counts giving a greater contribution.

3.3 $\widehat{FDR}$ based decision rules

An estimate of the posterior probability $\pi_i = E(H_{0i}|data)$ is worked out as a Monte Carlo mean, that is an empirical mean on a sample of realizations from the posterior distribution of H_{0i}:

$$\widehat{\pi}_i = \frac{\sum_j^M H_{0ij}}{M} \tag{3.11}$$

where H_{0i} is equal to one if the null hypothesis is true and to zero when the alternative is true, while M is the number of MCMC iteration.

Given $\widehat{\pi}_i$, estimated as the number of times when $H_{0i} = 1$ divided by the length of the simulation run, and given a pre-specified set of rejected null hypotheses, S, we can estimate the expected proportion of false discoveries as the empirical mean of the estimated posterior probabilities $\widehat{\pi}_i$'s belonging to S:

$$\widehat{FDR} = \frac{\sum_{i \in S} \widehat{\pi}_i}{D_S} \tag{3.12}$$

where D_S is the number of discoveries, i.e. the cardinality of S. $\widehat{FDR}$ is an estimate of the expected FDR conditional on data; recall expression (1.10) in chapter 1.

To make clear the terminology used we say in advance that in the sequel of the work we will not use the term "expected" every time we we want to refer to an estimate of the expected FDR conditional on data. To simplify we will refer to both the estimate and the estimator of FDR as $\widehat{FDR}$. Moreover, in discussing simulation results we will sometimes denote $\widehat{FDR}$ as the "estimated FDR".

After obtained the $\widehat{FDR}$ we can use it for two purposes. First, for merely predicting the proportion of false discoveries among a pre-specified set of areas declared at high-risk; for instance, in cases when for a particular areas is required a risk evaluation as regards several cause disease SMRs collected there. Then we can calculate the $\widehat{FDR}$ on the set of all cause disease SMR by averaging their respective $\widehat{\pi}_i$ values, that are all relative to that area (considering all those $\widehat{\pi}_i$'s as signaling a discovery). The second purpose may be more interesting: determining a rule of selection of high risk areas (for rejecting null hypotheses) based on the knowledge of the estimated proportion of false discoveries we will achieve in doing such rejections. The interesting aspect of

such rules is that they are not arbitrary thank you of the possibility to control the multiple testing error. Indeed, since the $\widehat{FDR}$ is a prediction of the expected proportion of errors, it allows us to control the maximum accepted proportion of false discoveries, that is, roughly speaking, the number of false positives we can at most live with. This is useful for instance when an overall conclusion about the risk level of all regions of the map is needed, and when the overall conclusion that the map contains high-risk areas need not be erroneous even if some of the null hypotheses are falsely rejected.

For decision (or selection) rule we mean a tool for calling areas either at high-risk or at null-risk. An easy way we suggest to determine such a kind of rule is pre-fixing a desired False Discovery Rate value, say $FDR = c$, (this means that we do not want to obtain an error greater than c in doing rejection) and operate in order to select as many as areas such that the $\widehat{FDR}$ (calculated on their respective $\widehat{\pi}_i$) is non-lower than c. We call an $\widehat{FDR}$ based selection (or decision) rule, a function of the $\widehat{\pi}_i$'s, the estimated posterior probabilities that the null hypothesis is true, and of t_π, a cut-off value for such posterior probabilities. The t_π value is actually a threshold: if $\widehat{\pi}_i \leq t_\pi$ an area i is declared a high-risk area and assigned to a set of discoveries, say S.

$$t_\pi = \inf\{\widehat{\pi_k} : \frac{\sum_{j=1}^{k} \widehat{\pi}_j}{k} \geq c\} \tag{3.13}$$

$$1 \leq k \leq \sum_i I(y_i \geq e_i)$$

The threshold t_π corresponding to an $\widehat{FDR} = c$ based selection rule is the smallest $\widehat{\pi_k}$ that yields $\widehat{FDR} = \frac{\sum_{j=1}^{k} \widehat{\pi_j}}{k} \geq c$. Such rule will identify k high-risk areas (discoveries) at the level of $FDR = c$, hence we can expect in average $k \cdot c$ false discoveries.

The strategy of building the rule following the idea of constraining the $\widehat{FDR}$ to be non-lower than c (the desired FDR level) is aimed to assure the "conservative" estimation of the FDR in the sense explained by Storey (2002) and showed in expression (1.4). However, the procedure (3.13) is not enough to assure the conservative estimation as it involves an estimated value of the expected FDR. Since $\widehat{FDR}$ is obtained through the estimated posterior probabilities $\widehat{\pi}_i$'s worked out by the proposed model, we need to assure that *BYM mix* can accurately estimate the required level of $FDR = c$, without incurring in under-estimation. If for instance the true unknown FDR is greater than c (i.e. the model under-estimate the value $FDR = c$), it will result in the practitioner declaring an estimate of false discoveries lower than it actually is, so do not achieving the required multiple testing control. Thus, one of the aim of the simulation study introduced in next chapter is to identify which levels of FDR are accurately estimated by the proposed model in several

simulated spatial scenarios that are frequent in practice, in order to suggest which levels of $\widehat{FDR}$ can be appropriate to build a conservative $\widehat{FDR}$ based decision rule.

Chapter 4

Simulation study

We saw that the *BYM mix* model proposed in chapter 3 can allow for an epidemiologist to estimate the False Discovery Rate in making decisions about a large set of null hypotheses (where the generic is $H_{0i} : r_i = 1$) evaluated in N spatial regions of a map under study. *BYM mix* model makes the same assumptions of the classic Besag York and Mollièe model (BYM) largely implemented for relative risks point estimation, but in addition can achieve the control of the expected FDR conditional on data through the assumption of a mixture for each log relative risk distribution. By applying the *BYM mix* model we aim to move from a point estimation inferential context to a multiple hypothesis testing set up, where we are mainly interested in evaluating two competitive hypotheses in each area, absence of risk *vs* possible presence of a higher risk. As said, there is however a point of contact between BYM and *BYM mix*, i.e. the capability of both models to provide point estimates of the true relative risk vector $\boldsymbol{r}$.

The main motivation of the simulation study is the evaluation of the *BYM mix* performance in diverse spatial contexts that can frequently occur in practice. There are not similar proposals in spatial epidemiologic literature, so we cannot make performance comparisons with other models, except for what concerns the relative risk estimation issue. Comparing true relative risk estimates by *BYM mix* and by BYM can be useful to understand if the proposed model can actually be of interest in practical applications, in the sense that can provide more information w.r.t. to disease mapping models usually employed.

We now give a short preview of the simulation study set up, a more detailed description being in the next sections. What we are mainly interested in, is checking the *BYM mix* model in several scenarios differing for the following factors: relative risks spatial correlation degree, relative risk level, size of areas, number of true alternative hypotheses. So, we firstly choose a map of N spatially contiguous regions in which we create each scenarios by simulating counts from a given set of known expected counts $\boldsymbol{e} = (e_1, ..., e_N)$. To create scenarios which differs for the spatial correlation degree

between risks $\boldsymbol{r} = (r_1, ..., r_N)$, we need to know the adjacency matrix of the chosen map, i.e. a symmetric $N \times N$ matrix where the generic element w_{ij} is 1 if i and j are neighbouring areas and zero if they are not (details about the way to build differently spatially correlated scenarios are in section 4.1.2). Hence for each area of the map we know the expected count and its neighbours. Then, for each scenario we build what we will call an "auxiliary" set of expected counts according to the required mentioned factors levels that we control in the simulation study (formally described in section 4.1.1). Finally, given a scenario specific auxiliary set of expected counts we can simulate as many as we need datasets of counts by independently sampling from the Poisson distribution in each area, the mean of the Poisson being the area specific auxiliary expected count.

As an example we describe how, starting from a basic set of known expected counts $\boldsymbol{e} = (e_1, ..., e_{341})$ collected in 341 areas, we obtained a dataset relative to a scenario with the following characteristics: the number of true null hypotheses were 19 out of the 341; the relative risk in areas where the alternative is true was 1.5, i.e. a 50% increment with respect to the null hypothesis risk; the risks were spatially uncorrelated. We generated the counts dataset by a two stage procedure. First, we create the auxiliary expected counts by choosing the 19 alternative hypotheses and multiplying their expected count by 1.5. To choose such an areas with a 1.5 risk level in order to generate a spatially uncorrelated risk pattern we exploit the idea of raising the risk in non contiguous areas (Richardson et al., 1995); see details in section 4.1.2. The second stage consists in simulating the actual dataset of 341 counts by sampling in each area i from the Poisson distribution with either mean $(e_i \cdot 1.5)$ (if i is one of those 19 areas where the null hypothesis is false) or mean (e_i) (if i is an area where the null hypothesis is true).

4.1 Objectives of the simulation study

The simulation study aims to evaluate the three targets that the *BYM mix* model can achieve:

a. estimation of FDR for a given set of areas declared as high-risk areas;

b. selection of high-risk areas by means of FDR based decision rules;

c. estimation of the true relative risk value in each area.

We now comment each above point both anticipating concepts related to the simulation study and recalling issues discussed in earlier chapters.

Point **a** refers to the ability of the proposed model to estimate the FDR. We will primarily check how close $\widehat{FDR}$ is to the true FDR that we known by simulation. Note point **a** and point **b** are related: as long as we can accurately estimate the FDR, we are able to determine $\widehat{FDR}$

based selection rules without incurring wrong declarations about the number of discoveries and the proportion of false discoveries among the discoveries themselves. On this note, one interesting question to investigate by simulation is whether or not there are FDR levels that are well estimated by the model, particularly in the small areas and spatially correlated scenarios which we are mostly interested to. Finding such FDR level as accurately estimated, that is to say conservatively estimated (without incurring in under-estimation), would make us able to suggest reliable $\widehat{FDR}$ based selection rules in many frequent small areas applications.

As regards point **b** it is worth recalling that an area is declared as a high-risk area when $\widehat{\pi}_i \leq t_\pi$, where t_π is an arbitrary chosen cut-off probability (or threshold) for the $\widehat{\pi}_i$'s. In section 3.3 we proposed an empirical but non arbitrary decision rule for determining such threshold and we called it $\widehat{FDR}$ based decision (or selection) rule. Precisely, the practitioner can fix a priori the desired FDR level (actually the error he decides he can live with) and then finding the cut-off value that realizes an estimate of the expected FDR conditional on data that is not lower than what was pre-fixed. The point **b** raises, indeed, another issue: the sensitivity and specificity of FDR based selection rules. This is a rather important point concerning the evaluation of the *BYM mix* performance since what we eventually need is a method to both adequately control the FDR and obtain the highest possible power in detecting true high-risk areas providing an acceptable level of specificity. Thus, we shall also quantify the sensitivity, and the correspondent level of specificity, for each possible FDR based selection rule in the diverse simulation contexts.

As regards the sensitivity/specificity issue, we have the chance to recall some basic concepts on multiple testing and explain how the FDR control differs from an unadjusted multiple testing procedure. Sensitivity (hereafter sn) is defined as the probability that an area is declared at high-risk given that it is such (probability of correctly rejecting a hypotheses), whereas specificity (hereafter sp) is the probability that a null-risk area is correctly declared as such (probability of correctly not rejecting a hypotheses). We can denote $1 - sp$ as the False Positive Rate (FPR): it corresponds to what is called in literature the Per Comparisons Error Rate (PCER) (see Benjamini and Hochberg, 1995). Controlling this quantity means non considering the multiple testing issue at all since it is actually the multiple testing analogous of the 'size' (α) fixed ex ante in the Neyman-Pearson single hypothesis test setting (i.e. the probability of rejecting a null hypothesis when it is true). Indeed, in a multiple testing setting FPR is the proportion of false positives among the set of true null hypotheses. The quantity we aim to control, the FDR, is instead the proportion of false positives among all positives. As an example to understand how the last two error measures yield different information, let us assume we have defined a rule (a test statistics and a critical value) to make inference on a number of null hypotheses. Suppose to reject hypotheses controlling

that in average $FPR \leq 0.05$, then it means that 5% of the times we will wrongly consider a true null hypothesis as rejected; testing 1000 null hypotheses we can expect up to 50 false discoveries. In many practical contexts this information (which actually correspond to an unadjusted procedure) is too poor. If, instead, we are willing to control the FDR at the 0.05 level, and, for instance, 100 null hypotheses out of the 1000 tested are rejected then this will results in about 5 false positives (if 500 are called significant it will results in 25 errors, etc). As claimed in chapter 2, in our case of study controlling the FDR gives useful information.

Point **c**) recalls us that *BYM mix* can also estimate relative risk values. They are obtained via MCMC estimation and appear as smoothed values compared to the maximum likelihood estimates, the smoothing being due to the Bayesian borrowing of strength between prior information and empirical data. Mostly in small areas cases the borrowing of strength can sometimes hide high relative risks because of the over-smoothing phenomena; i.e. the variance of the set of posterior relative risk values is underestimated. The point **c** is a suitable ground to make a comparison between BYM and *BYM mix* about the degree over-smoothing of the posterior relative risk estimates.

4.1.1 Factors controlled by simulation

We aim to evaluate the above mentioned three targets of the *BYM mix* model with respect to factors that typically play a big role in a spatial analysis based on $SMRs$. Such factors are the spatial correlation degree, the size areas (hence whether or not the expected counts are small values) and the true relative risk value in areas where the null hypothesis is false. Moreover, we expect that also the number of the true alternative hypotheses plays a role in estimating the FDR, hence controlling it as well. However, the factors which we will focus on with a special interest are the spatial correlation, the size areas and the true relative risk level in areas where the alternative is true.

Since the proposed model is originally constructed to address the lack of fitting of the Poisson model for the presence of positive spatial correlation between relative risks, we expect it works better when spatial correlation is strong than it is weak. We believe the model is less appropriate in case where risks are spatially uncorrelated because we introduce a non needed theoretical complexity. We will dedicate section 4.1.2 to describe in details the way we control the spatial correlation degree. In this section we introduce the other factors controlled in the simulation study, that are:

- $n = (\frac{N}{5}, \frac{N}{20})$, the number of areas where we incremented the relative risk value;
- $\theta = (1.5, 2, [1.2 \div 2])$: the relative risk value in areas where we incremented it;
- $SF = (0.5, 1, 5)$: a scale factor multiplying each area expected count (e_i) in order to vary

areas size.

The factor n corresponds to the number of true alternative hypotheses generated by simulation all over the map. If N is the total number of areas, we will achieve $N - n$ areas where the null hypothesis is true and n where it is not. In the latter areas, the relative risk is forced to be greater than 1; we recall in fact that the alternative hypothesis we are interested in is $H_{1i} : r > 1$. Varying the number of true alternative hypothesis we want to check if it can influence the FDR goodness of estimation, or the power of $\widehat{FDR}$ based decision (or selection) rules in identifying such true alternative hypotheses, or again the over-smoothing degree of relative risks estimates. Since in practical cases it is more frequent to find a small number of true alternative hypotheses, we decided to fix for the factor n values corresponding to a 20% ($\frac{N}{5}$) and a 5% ($\frac{N}{20}$) of true high-risk areas among the N total areas of the map. Thus we will generate exactly $n = 69$ true high-risk areas in the former case $n = 19$ true high-risk areas in the latter.

We want moreover to check the *BYM mix* model performance with respect to the true relative risk value θ arisen in area where the alternative hypothesis is forced to be true; in this simulation study context we will denote the relative risk as θ, instead of r, to stress the fact that θ is always greater than 1, denoting the relative risk in true high-risk areas. As regards the choice of the θ level, we follow some simulation studies present in literature and conducted for examining the characteristics of several disease mapping models (Richardson et al., 2004; Lawson et al., 2000). In those works even large values for the relative risks were checked for exploring model behavior in a wide range of cases. For our study, we generate scenarios with three levels of θ: $\theta = 1.5$, $\theta = 2$ and $\theta = (1.2 \div 2)$, where the latter means that θ will vary between a risk of 1.2 to a risk of 2 across the n areas. To not affect results interpretation, we however guaranteed that the average relative risk of scenarios where $\theta = (1.2 \div 2)$ is kept constant between scenarios differing for spatial correlation degree; i.e in all scenarios with $\theta = (1.2 \div 2)$ we operated the simulation of differently spatially correlated scenarios in order to obtain in average the same θ value (around 1.6).

Finally, as regards the factor size areas (SF), we expect the model behaves in a different way according to the magnitude of the expected count values on which observed (i.e. simulated) counts are generated by simulating. We believe that making inference on the true alternative hypothesis may be more difficult if the area i is small rather than big; the *BYM mix* model may over-estimate the π_i's. If the expected count e_i is small we have weak empirical evidence in favor of the alternative hypothesis even if it is actually true. We chose to make the scenarios varying for areas size by multiplying the set of basic expected counts for a scale factor SF taking respectively 1, 0.5 and 5. $SF = 1$ allow us to maintain the originally chosen map of expected count, $SF = 0.5$ build a map with twice smaller areas and $SF = 5$ identify scenarios where the areas are much bigger than

previous cases. Below we give a summary of the set of expected counts for each SF level:

```
> SF = 0.5
> summary(e.original*0.5)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.5496   5.5280   8.9970  20.7400  15.1700 749.7000

> SF = 1
> summary(e.original)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   1.099   11.060   17.990   41.470   30.340 1499.000

> SF = 5
> summary(e.original*5)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   5.496   55.280   89.970  207.400  151.700 7497.000
```

We consider the $SF = 0.5$ and $SF = 1$ cases as representative of small areas scenarios. Note as in $SF = 0.5$ (SF=1) half the number of areas have an expected counts lower than 9 (lower than 18). $SF = 5$ in our mind refers scenario where the small areas issue is no longer a problem. In such case the model could be considered no more necessary, since for SMRs collected at a big scale a traditional p-value may be good at evaluating the null hypothesis. We believe the model is useful for addressing small areas case, being the borrowing of strength between area-specific posterior probabilities (or in the case of the BYM model between area specific relative risks) fruitful when empirical information is not uniform across areas. We will still simulate counts from the scenarios where $SF = 5$, being however aware that our main target is the BYM *mix* model performance evaluation in small areas and spatially correlated scenarios.

4.1.2 Spatially correlated scenarios

To simulate scenarios with different degree of spatial correlation between risks we need to simulate a number n of areas where the alternative hypothesis is true, such that these n high-risk areas are positively spatially correlated. First step is the choice of the basic map on which simulating datasets. We need a map containing many areas, since we want to focus on applications where many tests are performed, one for each area. Moreover the BYM *mix* model is aimed to address small areas case study, hence we feel it is important to use a real map containing small areas (we consider a small area an area where $e_i < 5$), instead of using a regular grid. Following Lawson et al. (2000) we think important two recommendations:

1. the small areas have to have similar size and shape;

2. the small areas have not to show a clear spatial structure over the map (the tendency to be close each other).

If not so, artifacts may arise; as Gelman and Price (1999) argued, Bayesian inference on disease rates may lead to spatially correlated relative risk estimates for the only effect of spatially correlated expected counts. This could also be the case for the posterior probabilities estimates $\widehat{\pi}_i$'s computed by the *BYM mix* model, since it exploits the empirical information in the neighbouring areas through spatially autocorrelated random effects $\boldsymbol{u} = (u_1, ..., u_N)$.

We want to spend more comments on these two recommendations, these being valid also for the proposed model which focuses on both relative risk estimates ($\widehat{\boldsymbol{r}}$) and posterior probabilities estimates ($\widehat{\boldsymbol{\pi}}$). As regards point **2.**, the Bayesian shrinkage operated by disease mapping models in general, yields in each area a posterior disease rate that is a compromise between the observed region disease rate and both the mean disease rate for the entire map (global mean) and the neighborhood mean (local mean), with the relative weighting of observed and mean rates being dependent on the expected count e_i (that is low for small areas). Hence, if a lot of small areas are contiguous, i.e if small areas have a spatial structure, we would obtain a smoothing effect that will result in too uniform disease rates even if the true underlying rates are not uniform. In other words, posterior estimates of relative risks will be spatially correlated even if the true risks actually are not. Point **2.** ought to be at most as possible achieved for an appropriate interpretation of the simulation results. As regards point **1.**, the small areas are those where the shrinkage effect is more emphasized because, where the empiric information is poor, the relative risk is more shrunk towards local (the neighborhood) and global (the whole map) means. The idea that small areas ought to have approximately the same number of neighbours is again aimed to avoid the artifacts above mentioned: small areas with many neighbours would receive more prior information than small areas isolated. As a brief digression, we think interesting being able to consider all the small areas having the same chance to receive the prior information and avoid any artifacts; indeed, measuring the sensitivity as a function of the size of the areas would be an interesting point to investigate in future.

However, starting from a real map consisting of areas with irregular shape it is difficult follow the above recommendations at all. We believe, however, the choice to simulate from a real map of $N = 341$ known expected counts relative to lung cancer death all over five years in Emilia-Romagna municipalities can be appropriate. As regards point **1.** see figure 4.1 where an histogram is shown of the number of neighbours in small areas (we just considered areas where $e_i < 5$) for Emilia Romagna map. As regards point **2.** see instead figure 4.2 which highlights blue-tone colored small

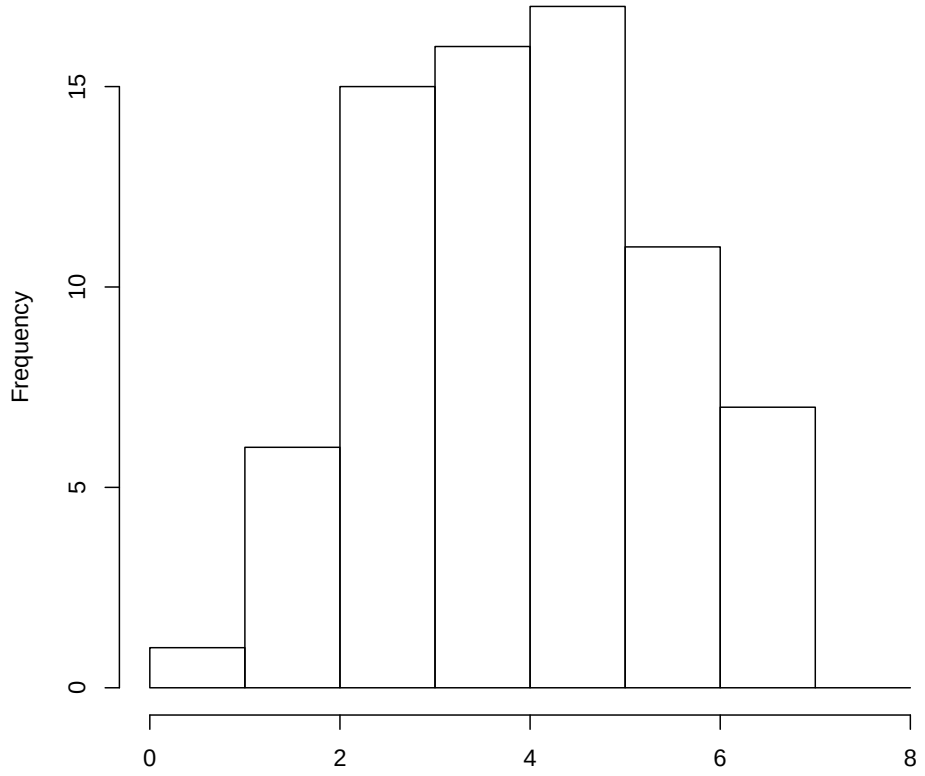


Figure 4.1: Histogram of the number of neighbors for areas where $e_i < 5$. The majority of small areas have a number of neighbors between 3 and 6.

areas; they seem to not show extreme tendency to aggregate each other.

For each level of the three factors n, θ and SF, we generated three different spatial scenarios, moving from a situation where spatial correlation is weak to one where it is strong. We choose some areas in the basic map, and in those areas we increment the relative risk value by setting $r = \theta$. For generating different spatial correlated scenarios we follow a simple principle used in Richardson et al. (2004): incrementing relative risks in contiguous areas make us able to create spatial correlated risks, whereas incrementing risks in not contiguous areas can yield a lower degree of spatial correlation. The idea of creating spatial correlation pattern by arbitrarily choosing high-risk areas is necessary for being able to compute the true value of FDR, since to work it out we need to know whether or not an area, where the null hypothesis has been rejected by a given decision rule, is one of the n true high-risk area. To creating scenarios following the above idea we need to know the adjacency matrix of the Emilia Romagna map. We shall see in the next section an

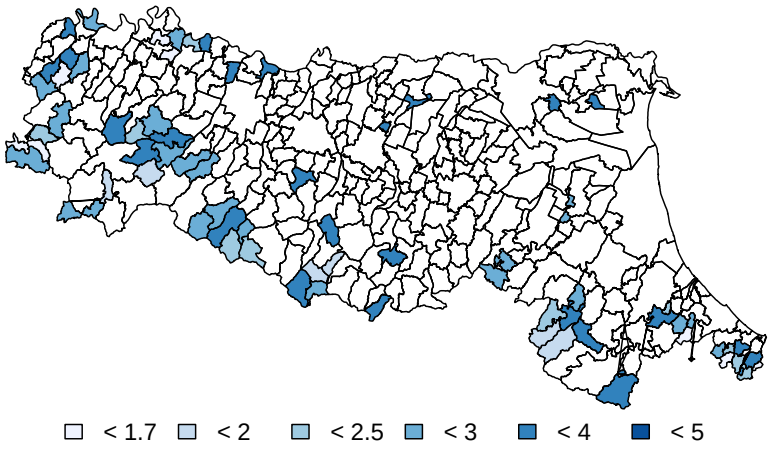


Figure 4.2: Map highlighting areas where $e_i < 5$. Small expected counts show a non strong degree of spatial correlation.

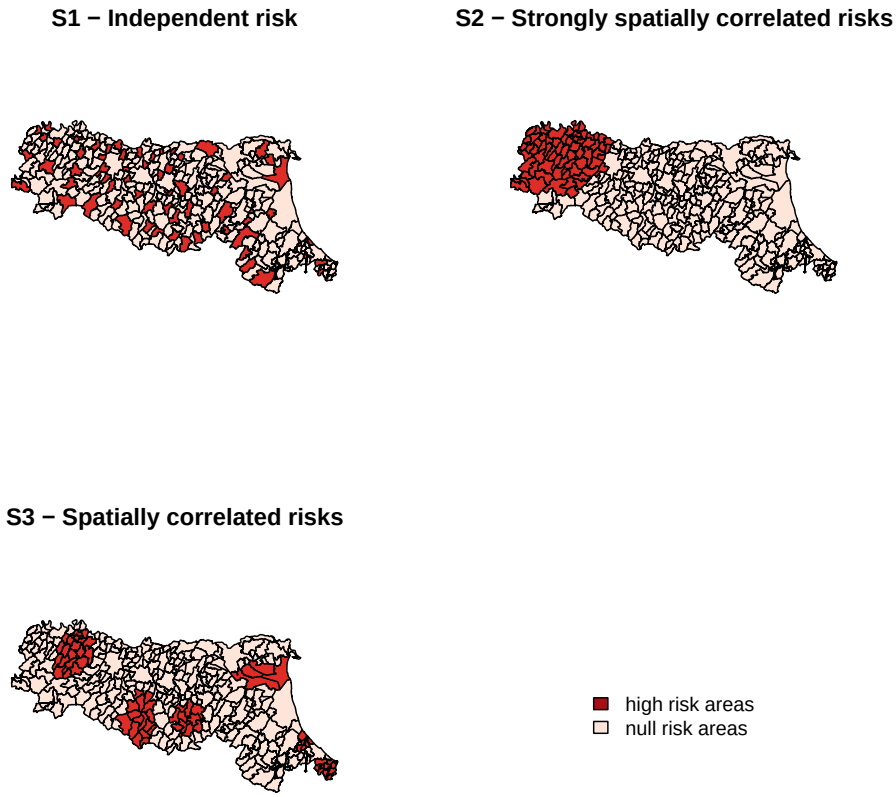


Figure 4.3: The three spatial scenarios ($S1$, $S2$ and $S3$) for $n = 69$

example of how generating such risks by sampling counts from a Multinomial distribution model.

The simulation of spatial count patterns with several degrees of spatial correlation raises the following difficulty: we need to generate spatial correlated counts and force either $r = \theta > 1$, in areas where we simulated the alternative hypotheses (2.6), or $r = 1$ in areas where we simulated the null hypothesis (2.5). Hereafter, we will denote true high-risk areas as HR areas and true null-risk areas (also background areas) as NR areas. To compact notation, we will call θ_{HR} (> 1) the relative risk in HR areas, while θ_{NR} $(= 1)$ the relative risk in background areas. We will call $S1$, $S2$ and $S3$ the three scenarios generated. In $S1$ the n HR areas are chosen controlling that they are not contiguous, in $S2$ all the n HR areas are chosen as contiguous, whereas $S3$ reflects an intermediate situation where the n HR areas are aggregated in a few number of clusters. Illustration of the three spatial scenarios in the case of $n = \frac{N}{5}$ ($n = 69$ out of the 341) and $n = \frac{N}{20}$ ($n = 19$ out of the 341) are shown in figures 4.3 and 4.4 respectively.

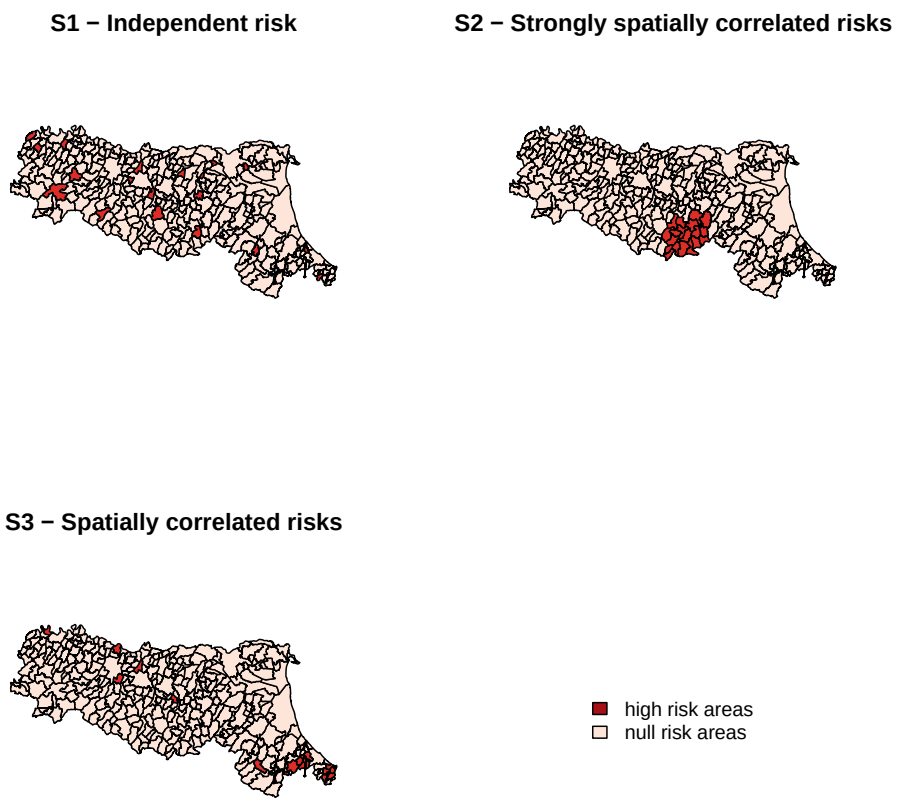


Figure 4.4: The three spatial scenarios ($S1$, $S2$ and $S3$) for $n = 19$

The main difficulty is building such three spatial scenarios without changing other factors levels, i.e. n, θ and SF. This is fundamental to appropriately interpret results relative to scenarios that differs for the only spatial correlation degree. This issue is made even more difficult by the fact that we want some of the HR areas generated in $S1$, $S2$ and $S3$ to be small areas; this, obviously, for evaluating the performance of the *BYM mix* model in small areas case studies that are frequent in practice. To properly make comparisons between $S1$, $S2$ and $S3$, keeping fixed each other factors we need to make two important constraints in choosing the HR areas in the map. Firstly, the sum of the HR areas expected counts has to be maintained approximately constant between $S1$, $S2$ and $S3$ so that we will introduce the same amount of risk in each of them. Secondly, the proportion of small areas (recall we want to introduce small areas among the set of generated HR areas) has to be approximately the same between the scenarios, the latter to avoid misinterpretation of simulation results. Indeed, introducing more small areas in $S2$ than in $S3$, for instance, could confound the simulation summary results that we want only being due to the different spatial correlation degree between $S2$ and $S3$ themselves. In figure 4.5 are shown histograms of expected counts relative to the areas where we generated the alternative hypothesis. We see that the proportion of areas where e_i is lower than 10 or 20 is approximately the same. Only in $S2$ scenario for $n = 19$ we have a slight lower proportion of areas with expected count lower than 10, but we believe it cannot affect the sensitivity results so much, especially if we think that it is the case where the number of true high-risk areas is small ($n = 19$). The total number of scenarios we generated is 3 x 3 x 2 = 54, that are 3 spatial correlated scenarios ($S1$, $S2$, $S3$), 3 levels of size of areas ($SF = 1$, $SF = 5$, $SF = 0.5$), 3 levels of relative risk values increment in HR areas ($\theta = 1.5$, $\theta = 2$ $\theta = [1.2 \div 2]$), 2 levels of the number of true null hypotheses ($\frac{N}{5} \simeq 69$, $\frac{N}{20} \simeq 19$). Depending on what we want to focus on we can show results fixing the factor under examination and varying all the others.

4.1.3 Multinomial sampling vs Poisson sampling

Counts were drawn from a Multinomial distribution in order to make the constraint $\sum_i e_i = \sum_i y_i$. More precisely, the constraint we actually impose is $\sum_i e_i \times SF = \sum_i y_i$, since we want SF to be a scale factor allowing for producing different size of areas.

$$(Y_i, ..., Y_N) \sim Multinomial\left(\sum_i e_i \times SF \times \theta_i, \boldsymbol{\tau}\right) \tag{4.1}$$

In the above notation $\boldsymbol{\tau}$ is a N component vector, in our case the generic element being $\tau_i = \frac{e_i \times SF \times \theta_i}{\sum_i e_i \times SF \times \theta_i}$. Marginally each Y_i is a Binomial random variable with parameters $(\sum_i e_i \times SF \times \theta_i)$ and τ_i. The term $(e_i \times SF \times \theta_i)$ is what we above denoted as the auxiliary expected count, since it is the expected count by which we can generate the area i count, relative to the scenarios having

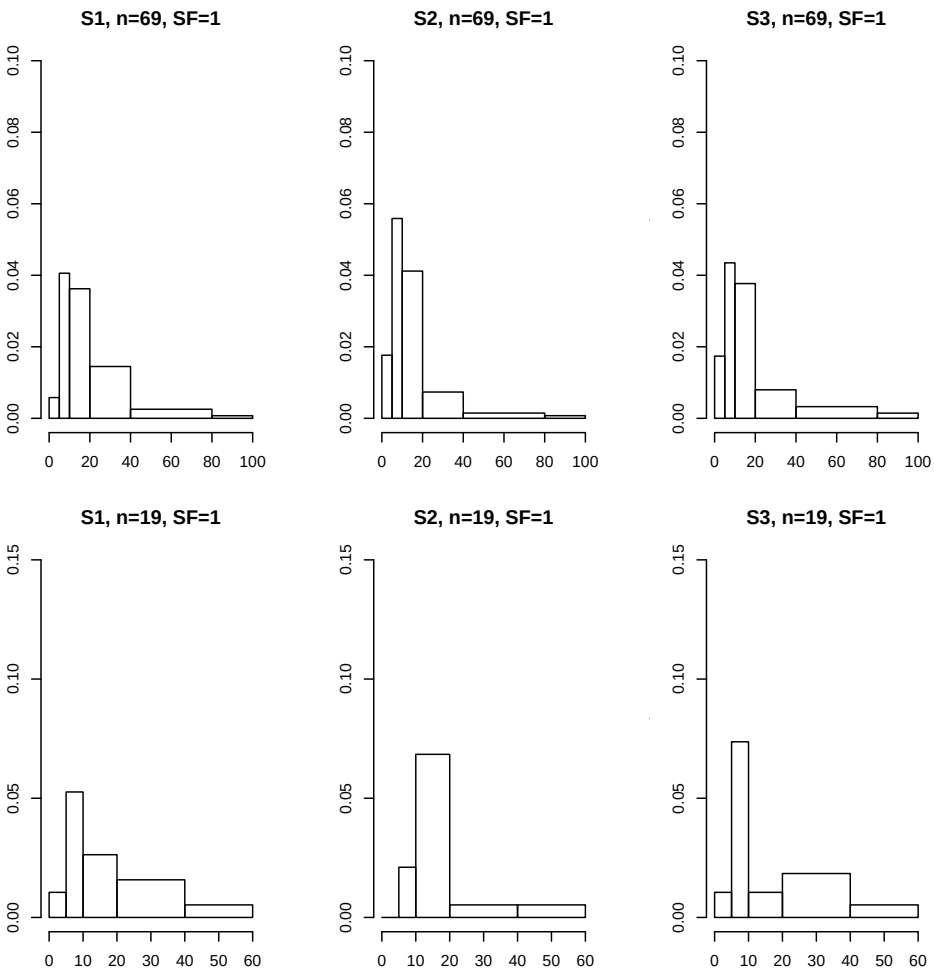


Figure 4.5: Histograms relative to the expected counts (of areas where the alternative hypothesis is true) for all spatial scenarios ($S1$, $S2$ and $S3$) and all n scenarios (n=19, n=69)

pre-chosen SF and θ levels; an example follows. Let us suppose we want to generate 100 datasets, relative to the Emilia-Romagna map, for a given level of SF and θ factors. To this aim we need to divide the sum of the known expected counts relative to the Emilia Romagna map ($\sum_i e_i \times SF$) across the $N = 341$ areas where each probability τ_i depends on ($e_i \times SF$) and on θ_i. The latter is equal to either θ_{HR} or θ_{NR}, depending on if area i is an HR or a NR area. For each HR area we sampled 100 values by using $\theta_{HR} > 1$, while for every background area we drew 100 values with $\theta_{RN} = 1$. For instance, in the scenario where $n = \frac{N}{5}$ (69 true alternative hypotheses out of 341 hypotheses), $SF = 0.5$ (very small areas) and $\theta = 1.5$ we will sample 100 values from a Multinomial sampling distribution. The generic simulated count is a fraction of the total $\sum_i e_i \times SF$ proportional to a probability of $\tau_i = \frac{e_i \times 0.5 \times 1.5}{\sum e_i \times 0.5 \times \theta_i}$, if i is an HR area, or to $\tau_i = \frac{e_i \times 0.5 \times 1}{\sum e_i \times 0.5 \times \theta_i}$ if i is instead a background area (NR).

Sampling the counts from a Poisson distribution is an alternative way, even though it cannot maintain the constraint between summation over expected and observed (here simulated) counts. The Poisson sampler reproduces the case where the reference rates for the disease under study are taken from an external standard population; as a result we will obtain HR areas with relative risk values around θ_{HR}, and background areas with risks leveling out at the constant risk of 1. With the model (4.1), instead, the total amount of simulated counts is constrained to be equal to the total amount of expected counts. The probability of drawing counts for area i depends on $\frac{e_i \times \theta_i}{\sum_i e_i \times \theta_i}$. Hence, in the HR areas we will not generate relative risks around θ_{HR}, but in average around the ratio $\frac{\theta_{HR}}{\theta_{NR}}$ that will be a bit lower than θ_{HR}. In other words, in scenarios where $\theta = 1.5$ we will have as number of observed disease cases as we were in scenarios where $\theta = 2$ because the total generated counts is the same, just the relative risk of HR areas relatively to background areas will differ. As a result, we will obtain areas with relative risks either greater or lower than 1 according wether they are HR or NR areas as if we had $SMRs$ computed by means of internal reference rates.

We believe that *BYM mix* model performance cannot be affected by considering internal or external standardization, however simulating from the sampling model (4.1) can be useful for building scenarios with different n factor levels without changing the θ factor levels, hence avoiding misinterpretation of the results. In fact, sampling from a Poisson model would augment the risk in HR areas ($\theta_{HR} > 1$) and would leave unchanged the risk in NR areas ($\theta_{NR} = 1$). So, in scenarios where $n = \frac{N}{5}$, the total amount of risk in the map will be larger than in scenarios where $n = \frac{N}{20}$ because of the bigger number of HR areas. By sampling from (4.1) the total observed counts must always be equal to the total expected counts, hence simulation results about scenarios having different values for n, and the same values for each other factors, will only depend on the varying

number of true alternative hypotheses as desired.

4.1.4 MCMC based inference

For each of the 54 scenarios (populations) we generated 100 datasets (elements) to make allowance for sampling variability. For each dataset, i.e. for each set containing 341 observed counts (sampled by model (4.1) concerning the particular scenario) and 341 expected counts (the auxiliary expected counts concerning the particular scenario), the model was implemented in OpenBugs to get estimates of the 341 components vector of $\boldsymbol{\pi}$ and $\boldsymbol{r}$. We have denoted the estimates of the posterior probability of H_{0i} as $\widehat{\pi}_i$ since it is worked out as the mean of a sample of 0 and 1 values from the posterior distribution $[H_{0_i}|\boldsymbol{y}]$. We used BRugs package of R (version 0.4 - 1) for running OpenBUGS version 3.0.2.

As regards the choice of burn in period and number of sampled values to consider for calculating the estimates, we checked the convergence in the described below way. Indeed, we cannot analyze all 54 scenarios so we checked one dataset belonging to each critical scenario, i.e. where the convergence may be thought of being slower because of the weak empirical information. So, for all levels of n and all spatial scenarios ($S1$ $S2$ $S3$), we considered as critical scenarios those with small areas (SF=0.5) and small risk levels $\theta = 1.5$. Once evaluated the number of iterations necessary for the convergence in datasets coming from the mentioned scenarios, we decided to fix a conservative value and implement it automatically in all simulations. The burn in period was fixed at 4000, after having checked autocorrelation plots of nodes H_{0i}, visual investigation of the trace plots, and calculation of the Gelman-Rubin convergence statistics; the latter two was computed after having run three chains with different starting values for each parameter with a prior distribution ($\boldsymbol{H_0}$, $\boldsymbol{u}$, $\boldsymbol{v}$, α, τ_u, τ_v, $\boldsymbol{\phi}$). The number of iterations for calculating estimates was fixed at 8000, obtaining an effective sample size around 2000 for almost all H_{0i} parameters of interest.

4.2 Evaluating model performance

In this section we introduce measures and tools for analyzing simulation results about the *BYM mix* model performance concerning the three targets discussed in section 4.1. One of them is the goodness of estimation of the FDR. We want firstly give two important points on the FDR computation. As already said we are interested in testing two competing hypotheses in each area, the null being the relative risk is 1 and the alternative being the area relative risk is greater than 1. Our alternative is not a bilateral hypothesis, i.e either lower or greater than 1. As discussed in chapter 3, we aim to evaluate such test for each area i by working out by the *BYM mix* model an estimate of the probability that the null hypothesis is true ($\widehat{\pi}_i$). We also noted that, for the

way the model has been formulated, a small $\widehat{\pi}_i$ value denotes a deviation from the null hypothesis, but such deviation can be in the direction of a relative risk both greater and lower than 1 (i.e. a bilateral alternative). Since our definition of the alternative hypothesis, we cannot consider a small $\widehat{\pi}_i$ as a possible discovery if it comes from a lower-risk areas. Hence, we need to restrict the set of potentially rejectable hypotheses to the set of areas where $y_i \geq e_i$. If not so, the denominator of the False Discovery Rate (i.e. the number of discoveries) would become greater than due, and we would achieve an under-estimation of it. Note that we do not make the same restriction when we estimate the π_i's by the model since all observation are needed for that. Therefore, the practitioner willing to use the proposed model should do such operation after having estimated the model, and before determining an $\widehat{FDR}$ based decision rule for selecting high-risk areas. Doing so, the set of discoveries achieved at the prefixed FDR level will contain high-risk areas.

The other point is on the computation of the FDR via simulation. In the following in this section we will denote a quantity as "realized" if it is worked out by simulation, and with "estimated" when it is obtained by the *BYM mix* model. Working out the FDR realized by simulation is possible since the simulation study allows us to create the population from which we can calculate the "true" FDR. We want however to clarify that FDR is not a measurable quantity, so we could not conduct an experiment to infer its value. FDR is a quantity that contains a random variable both in numerator and in denominator, but such random variable depends on a decision about the null hypothesis. Therefore, in our setting, the FDR can be calculated after having chosen a threshold t_π for the posterior probabilities, even though it is in principle computable for any given set of $\widehat{\pi}_i$'s as argued in section 3.2, without concerning any selection rules that tells us how to build such set of $\widehat{\pi}$'s. However we say in advance that in the presentation of the results we will show graphs where the FDR is plotted against a threshold t_π (also denoted as cut-off probability) and is relative to the set of discoveries found out selecting an area i if $\widehat{\pi}_i \leq t_\pi$. We will compare the estimated FDR (i.e. $\widehat{FDR}$), calculated averaging all the $\widehat{\pi}_i$'s lower than t_π, with the realized (or true) FDR computed as the proportion of the number of discoveries (i.e. the number of areas where $\widehat{\pi}_i < t_\pi$) that are actually HR areas among the discoveries themselves. In order to calculate the denominator of the realized FDR we need to know the spatial scenarios $S1$, $S2$, $S3$, that is we need to know whether or not each area is a HR (see the maps illustrating HR and NR areas in figures 4.3 and 4.4).We will formally write this down in the next section.

To sum up the concepts above, we say that though the FDR can be in principle (and in practice) computed for any given set of selected $\widehat{\pi}_i$'s, regardless of the rule used to select them, we will focus attention on $\widehat{FDR}$ achieved in π_i's monotonic sets. We define a $\widehat{\pi}_i$'s monotonic set to be a set of ordered $\widehat{\pi}_i$'s, i.e. $\widehat{\pi_1} \leq \widehat{\pi_2} \leq ... \leq \widehat{\pi_m}$ where m is the number of areas such that $y_i \geq e_i$ (number of

possible discoveries). For instance, in the FDR *vs* t_π graph, for all t_π values in the horizontal axis we will plot the FDR calculated on the set composed by all $\widehat{\pi}_i$'s non-greater than t_π. We will focus on such $\widehat{\pi}_i$'s monotonic sets because of the practical examples we think the epidemiologist can be interested in. For instance, the need to select high-risk areas having collected SMR's relative to a particular disease in many spatial regions; according to the threshold t_π at which the $\widehat{\pi}_i$'s are selected, we achieve different sets of discoveries, each one with its $\widehat{FDR}$. Thus the practitioner can declare a set of discoveries (a set of high-risk areas) for any given $\widehat{FDR}$ level. Aim of the simulation is to investigate in different scenarios the $\widehat{FDR}$ levels that are close to the true FDR level, in order to know which $\widehat{FDR}$ based decision rules can be recommended as they do not yield loss of sensitivity or loss of specificity.

If more than one disease cause SMR are collected in each area, the idea of measuring the FDR in $\widehat{\pi}_i$'s monotonic sets can be less interesting as in such as case the practitioner might prefer to work out the $\widehat{FDR}$ in each area to evaluate its generale risk state; note we can in principle average the $\widehat{\pi}_i$'s belonging to different diseases since the $\widehat{FDR}$ can be computed in any set of discoveries.

4.2.1 Measures introduced for evaluating model performance

We describe the measures useful for evaluating performance about the three targets of the proposed model. All the formulas above introduced are not averaged over the 100 simulated datasets, except the sensitivity and specificity formulas.

Relative Risk estimate, ($\widehat{r_i}$). The proposed model can estimate r_i as well as the Besag York Mollìe model. For each area i, it is a Monte-Carlo mean:

$$\widehat{r_i} = \frac{\sum_{j=1}^{M} (\exp(1 - H_{0ij})(\alpha_j + u_{ij} + v_{ij}))}{M} \tag{4.2}$$

where M is the number of $MCMC$ iterations.

Estimated posterior probability that the null hypothesis is true, $\widehat{\pi}_i$. Such a value is computable for each area i as a Monte-Carlo mean:

$$\widehat{\pi}_i = \frac{\sum_{j=1}^{M} I(H_{0ij} = 1)}{M} \tag{4.3}$$

where M is the number of $MCMC$ iterations.

Number of Discovery, (D). We compute D as the number of selected areas, i.e. the number of areas declared at high-risk by a given selection. Determining a selection rule means choosing a threshold for the $\widehat{\pi}_i$'s; a possible way is through (3.13). Thus, the number of discoveries D

conditional to a t_π value is:

$$D = \sum_i I(\widehat{\pi}_i \leq t_\pi) \tag{4.4}$$

The indicator function is equal to 1 when $\widehat{\pi}_i \leq t_\pi$.

Estimated False Discovery Rate, ($\widehat{FDR}$). This value is a prediction of the proportion of false discoveries among the discoveries. We recall the formula (3.12), where given a set S of selected areas posterior probability estimates, $\widehat{\pi}_i$'s, of cardinality D_S we have:

$$\widehat{FDR} = \frac{\sum_{i \in S} \widehat{\pi}_i}{D_S} \tag{4.5}$$

Realized (or true) False Discovery Rate, (FDR). FDR is the number of discoveries that actually do not belong to the set of HR areas, divided by the number of discoveries. Given a set S of selected areas posterior probability estimates, $\widehat{\pi}_i$'s, of cardinality D_S we have:

$$FDR = \frac{\sum_{i \in S} I(\widehat{\pi}_i \leq t_\pi, \theta_i = 1)}{D_S} \tag{4.6}$$

The indicator function takes 1 when both $\widehat{\pi}_i \leq t_\pi$ and area i is a background area ($\theta_i = \theta_{NR} = 1$). In discussing results it will be denoted as true FDR.

Cut-off value for the $\widehat{\pi}_i$'s given $\widehat{FDR}$, (t_π). As already mentioned a possible way to determine a cut-off (or a threshold) t_π is pre-fixing a value for the FDR, say $FDR = c$, and selecting as many discoveries as possible such that $\widehat{FDR} \geq c$. We recall the selection rule formulation (3.13):

$$t_\pi = \inf\{\widehat{\pi_k} : \frac{\sum_{j=1}^{k} \widehat{\pi}_j}{k} \geq c\}$$

$$1 \leq k \leq \sum_i I(y_i \geq e_i)$$

the cut-off value, t_π, is the smallest $\widehat{\pi_k}$ that yields $\widehat{FDR} \geq c$.

We explain how t_π can easily be calculated in practice. We select all areas where $y_i \geq e_i$. Then we make an ordered list of their corresponding posterior probabilities $\widehat{\pi}_i$'s; such a list will actually contain a number of increasing $\widehat{\pi}_i$'s values. From such a list, we can build each $\widehat{\pi}_i$'s monotonic set by fixing t_π equal to each $\widehat{\pi}_i$ in turn, and counting discoveries by formula (4.4); we thus obtain a number of $\widehat{\pi}_i$'s monotonic sets equal to the number of $\widehat{\pi}_i$'s in the list. Then, we calculate $\widehat{FDR}$ on each π_i's monotonic set, hence obtaining a list of increasing $\widehat{FDR}$ that actually matches the list of the increasing $\widehat{\pi}_i$'s. By simply observing the increasing $\widehat{FDR}$ values we can find an $\widehat{FDR} = c$ based decision rule. It will be a function of the vector $\boldsymbol{\pi}$ and of t_π as in the usual form introduced in chapter 1, where the threshold t_π is equal to the $\widehat{\pi}_i$ corresponding to a $\widehat{FDR}$ value greater or equal than c, c being the pre-specified FDR level which we want to control. If $\widehat{FDR}$ is close to the

true FDR we can be confident in declaring that the set of discoveries found by the rule is affected by a percentage of false discoveries equal to c. We will show the extent of degree we can trust the FDR estimation for all possible choice of t_π in discussing results of the simulation; see section 4.3.

False Positive Rate, FPR. It is the probability that a null-risk area (NR area) is wrongly declared as a high-risk area:

$$FPR_i = P(\pi_i \leq t_\pi | \theta_i = 1) = \frac{\sum_{k=1}^{100} I(\pi_{ik} \leq t_\pi, \theta_i = 1)}{100} \tag{4.7}$$

$$FPR = \frac{\sum_{i:\theta_i=1} \frac{\sum_{k=1}^{100} I(\pi_{ik} \leq t_\pi, \theta_i = 1)}{100}}{\sum_i I(\theta_i = 1)} \tag{4.8}$$

We recall that $\theta = 1$ means that the null hypothesis i is true, and that when π_i is lower than a cut-off t_π we are rejecting the null hypothesis. It is the probability of a false positive computed as an average over all 100 datasets. We can compute (4.7) for each background area, and (4.8) averaging over all background areas.

Note that $sp_i = 1 - FPR_i$ is the specificity of the test evaluated in area i, i.e. the probability of declaring i as a null-risk area when $\theta_i = 1$. The specificity can be computed as:

$$sp_i = P(\pi_i > t_\pi | \theta_i = 1) = \frac{\sum_{k=1}^{100} I(\pi_{ik} > t_\pi, \theta_i = 1)}{100} \tag{4.9}$$

$$sp = \frac{\sum_{i:\theta_i=1} \frac{\sum_{k=1}^{100} I(\pi_{ik} > t_\pi, \theta_i = 1)}{100}}{\sum_i I(\theta_i = 1)} \tag{4.10}$$

Sensitivity, (sn**).** It is the probability that an HR area is correctly declared as high-risk areas, i.e. the probability of a true positive:

$$sn_i = P(\pi_i \leq t_\pi | \theta_i > 1) = \frac{\sum_{k=1}^{100} I(\pi_{ik} \leq t_\pi, \theta_i > 1)}{100} \tag{4.11}$$

$$sn = \frac{\sum_{i:\theta_i>1} \frac{\sum_{k=1}^{100} I(\pi_{ik} \leq t_\pi, \theta_i > 1)}{100}}{\sum_i I(\theta_i > 1)} \tag{4.12}$$

In showing results we will take in account sensitivity averaged over all HR areas, and specificity averaged over all NR areas. Measuring sensitivity and specificity for each area i may be an interesting future development, mainly to observe the power achieved by FDR based selection rules in areas of different size.

4.2.2 Summary graphs

In what follows we describe graphs used to evaluate the performance of the *BYM mix* model for the usual three targets that the model can potentially achieve. Every measure plotted in graphs introduced below are averages over the 100 simulated datasets. A graph can be drawn for each of the 54 simulated scenario. However we shall show all 54 scenarios for only the FDR estimation issue, while we will plot graphs by grouping the three spatial correlation scenarios ($S1$, $S2$ and $S3$) in the same figure for summarizing simulation results relative to sensitivity of the $\widehat{FDR}$ based rules and goodness of relative risk estimation.

1. For investigating the ability to estimate FDR (point **a**)

   - plot the $\widehat{FDR}$ and the true FDR *vs* a cutoff value t_π.

   From this plot we can appreciate the goodness of estimation for the proposed model by examining the closeness of the estimated FDR ($\widehat{FDR}$) to the realized FDR (or true FDR). For each t_π threshold at which the posterior probabilities $\widehat{\pi}_i$'s can be cut, we obtain the respective $\widehat{FDR}$ by computing 4.5 conditionally to a discoveries set worked out given t_π itself. At the simulation stage we also obtain the true FDR by formula 4.6.

2. For investigating the ability to select high-risk areas (point **b**))

   - plot *sn*, *sp*, *vs* $\widehat{FDR}$

   Such a plot shows us the averaged values of sensitivity and specificity (averaged respectively on HR and NR areas, besides averaged over the 100 datasets) for any given $\widehat{FDR}$ based selection rule. As an example to show how such values can be plotted, let us suppose an $\widehat{FDR} = 0.10$ selection rule is wanted. Then, by the mechanism used for building a rule of the form (3.13) we can work out the threshold t_π that yields the biggest number of discoveries provided that $\widehat{FDR} \geq 0.10$. Such a threshold is then used for computing formula 4.12 and 4.10 and results of sensitivity and specificity plotted in correspondence to the value of 0.10.

   An interesting issue to consider is what level of sensitivity would be achieved if the *BYM mix* model could yield an estimate of FDR equal to the realized (or true) FDR. To this aim we shall show a slight different version of such a graph plotting both the sensitivity and the "potential sensitivity" in order to have an insight, for all scenarios, about which level of $\widehat{FDR}$ can be suitably used to determine $\widehat{FDR}$ based decision rules that gain the maximum powerful control of the multiple testing error. For this graph we will show all 54 scenarios results by aggregating all three spatial cases ($S1$, $S2$ and $S3$) in the same plot. (see figures A.7 and A.8

in Appendix). The main interest lies in checking if the *BYM mix* model allows for determining whether $\widehat{FDR}$ based selection rules are more sensitive in spatially correlated scenarios ($S2$ and $S3$), i.e. the cases where the model is thought of working well since it includes random effects spatially autocorrelated ($\boldsymbol{u}$). To the other side, we expect that sensitivity is lower in small areas ($SF = 0.5$ and $SF = 1$) and small θ values.

3. For investigating the ability to estimate the log relative risk (point **c**))

   - plot box-plots of log relative risks estimated by *BYM* and *BYM mix*.

   We will show such box-plots of the $\widehat{r_i}$ relative to HR areas grouping the three spatial scenarios in the same window to compare smoothing degree produced by both the *BYM mix* and *BYM* models: see figures A.9 and A.10 in 4.4.1.

4.3 Results

In this section we will respectively focus on goodness of estimation of FDR, sensitivity/specificity of $\widehat{FDR}$ based rules and goodness of estimation of the relative risk estimates.

4.3.1 The *BYM mix* performance on FDR estimation

Before commenting on the graphs, we attempt to interpret the over-estimation and the under-estimation of the False Discovery Rate. We try to figure out the reasons and consequences of both over-estimation and under-estimation. First of all, we expect the over-estimation of the FDR to be due to the presence of small areas. In HR small areas, though the true hypothesis is $\theta > 1$, we can have weak empirical evidence in favor of H_{10}, hence posterior probability of the null hypothesis may be over-estimated. Then, since the $\widehat{FDR}$ is computed by averaging $\widehat{\pi}_i$ values, the FDR will result be over-estimated as well. If for some FDR values we observe that $\widehat{FDR} \geq$ true FDR it always results in some kind of loss of sensitivity for such $\widehat{FDR}$ based rules; the practitioner, for a fixed number of discoveries, will declare a larger $\widehat{FDR}$ than actually it is, or analogously, for a fixed $\widehat{FDR}$, he will declare a lower number of discoveries. A loss of sensitivity is also likely to be caused by small θ values since in such scenarios the empirical evidence against the null hypothesis is weaker than in large θ scenarios.

Following such idea, the FDR estimation may be improved as the empirical evidence against H_0 of each area become stronger, i.e. as long as the factors θ and SF gets bigger. On the other hand, in the case where areas are very large and relative risk values in HR areas are high, we also may expect the Bayesian borrowing of strength to become too strong hence affecting the conclusions of the practitioner. This would result in the opposite problem: the under-estimation

of the π_i values, the consequent under-estimation of some FDR levels, hence a loss of specificity, i.e. rejecting the null hypothesis in areas where H_0 is actually true. For FDR values for which $\widehat{FDR} \leq$ true FDR, the $\widehat{FDR}$ selection rules will make the practitioner wrongly declare some areas as discoveries, because the true FDR value, i.e. the value that the practitioner cannot know, will actually be greater than that estimated by the *BYM mix* model.

As a final remark on over/under estimation of FDR it is worth noticing that since we want to achieve the control of a desired pre-specified FDR level, over-estimating is by far better than under-estimating the FDR. In the latter case the practitioner would declare less errors than he actually made, whereas with the former he would declare more errors than due that actually means not achieving the multiple testing control. As we said in section 3.3 what we need is a conservative estimation. On this note, an over-estimation at worst causes the $\widehat{FDR}$ selection rule adopted to not be as powerful as could be, that, in some sense, corresponds to pursue a conservative control of the FDR as is required. We will recall this in section 4.3.2.

We now discuss what is observed through simulation about the FDR estimation in the three different spatial scenarios by exploiting plots of the true FDR and $\widehat{FDR}$ *vs* t_π. We will look in this section at only two examples of both over-estimation and under-estimation, presenting all scenarios results in Appendix. In graphs here presented, we also draw two vertical line on the t_π values corresponding to $\widehat{FDR}$ equal to 0.05 and 0.10, to put in light the goodness of estimation of the FDR for such "traditional" values (figuring out that a practitioners could want to determine an $\widehat{FDR}$ based decision rule given such FDR values). We will compare averaged values of FDR (where the mean is calculated over the 100 simulated datasets as usual) and also point out some considerations about the true FDR and the $\widehat{FDR}$ distribution over the population of the 100 datasets by showing box-plots (containing the 100 FDR values for each threshold t_π in the horizontal line). Recall, in fact, that we have no unique FDR for a singular dataset, but we have an $\widehat{FDR}$ and a "realized" FDR for each fixed threshold t_π for the $\widehat{\pi}_i$'s.

In Figures 4.6 and 4.7 we investigate the combinations of factors $\{n = 69, \theta = 1.5, SF = 1\}$ and $\{n = 19, \theta = 1.5, SF = 1\}$. The horizontal axis reports all possible cut-off t_π, conditionally on which, in the vertical axis are calculated both the $\widehat{FDR}$ and the "realized" FDR. We also calculated 95% level confidence limits for $\widehat{FDR}$ assuming the 100 values to be normally distributed.

Such pictures show the tendency of the *BYM mix* model to over-estimate FDR in larger n scenario. In the smaller n scenario, instead, we achieve an accurate estimation for correlated cases $S2$ and $S3$ until some point (between 0.10 and 0.15) beyond which FDR becomes under-estimated.

To investigate the effect of having taken averaged values over the 100 datasets for summarizing results, figures 4.8 and 4.9 show the same plot as in Figure 4.7 but add underlying box-plots

of respectively the 100 "realized" and the 100 estimated FDR. By only looking at confidence bounds we realize green dashed lines are more spread out in n smaller scenarios as they yield more uncertainty about FDR estimation. Box-plots pictures show instead the empirical distribution of the true FDR and the $\widehat{FDR}$ according to each cut-off probability t_π in the horizontal axis. First, we note that true FDR (and $\widehat{FDR}$ as well) are not identically distributed regardless of the t_π values in the horizontal axis. The 100 "realized" FDR of figure 4.8 show strong right skewness for small cut-off values, almost all the mass of probability is around 0. They become nearly symmetric for t_π values greater than approximately 0.10 though still more spread out than a normal distribution. Instead, the 100 $\widehat{FDR}$ of figure 4.9 follow a quite symmetric distribution, probably more concentrated around the median than the normal distribution, and except in $S1$ scenario, they are almost identical distributed. An interesting note is that for small cut-off t_π, the empirical distribution of $\widehat{FDR}$ is more spread out in the independent risk case ($S1$) than in the dependent risk cases ($S2$ or $S3$): this signals that the FDR estimators is generally more precise in scenarios more favorable to the model itself, i.e. the spatially correlated risks scenarios.

By looking at figures from A.1 to A.6 in section 4.4.1 we can get a global insight about FDR estimation in all 54 scenarios. We give a brief a description of what can be found in one of the six figures. We have nine graphs of $\widehat{FDR}$ and true FDR *vs* t_π representative of 3 spatial scenarios (S1, S2, S3) times 3 θ scenarios; keeping fixed the factors n and SF. Figure A.1 for example focuses on nine scenarios with $n = 69$ (around a 20% of true high-risk areas) and $SF = 0.5$ (small areas). Each row, containing three figures, corresponds to a different θ value, the first on the top being $\theta = 1.5$, the second $\theta = [1.2 \div 2]$ and the third $\theta = 2$. Each column, containing three figures, corresponds to a different spatial scenario, the first column on the left being $S1$ (independent risks), the second $S2$ (strongly spatially correlated risks) and the third on the right $S3$ (spatially correlated risks).

We comment on the results of figures (A.4, A.5 and A.6) which are relative to $n = 19$ scenarios. We see as the model tends to under-estimated the FDR for some non-small values, but mostly for independent risks cases ($S1$), high θ values (for ex. $\theta = 2$), high SF levels (big areas scenarios like $SF = 5$). Instead, in correlated scenarios, small θ values and small areas case we often achieve an accurate estimation, except a small over-estimation for some t_π values corresponding to high $\widehat{FDR}$ values. In latter scenarios an $\widehat{FDR} = 0.10$ based selection rule can be suggested. Instead, in $SF = 5$ scenarios (non small areas) we get strong under-estimation, even for small $\widehat{FDR}$. In such scenarios we can determine selection rule based on just very small $\widehat{FDR}$ in order to avoid a huge loss of specificity; for instance, an $\widehat{FDR} = 0.10$ based selection rule cannot be suggested.

We focus now on the first three figures (A.1, A.2 and A.3) that are relative to $n = 69$ scenarios. The over-estimation is more frequent in $n = 69$ scenarios, see also figure 4.6 that corresponds to the

first row of figure A.2. The most over-estimated FDR lies in the strongly spatially correlated case $S2$. Over-estimation is less in the intermediate correlated case $S3$ and is even lower in independent risk case $S1$. Note that the "realized" FDR is by far the lower in $S2$ case, but more interestingly note that in $n = 69$ scenarios the true (or realized) FDR is in general lower than in $n = 19$ scenarios; see Figure 4.10 for a comparison of FDR goodness of estimation grouping the two n level in the same window. Thus, it seems that, moving from scenarios where the true alternative hypothesis is $n = 19$ to scenarios where $n = 69$ (out of the total $N = 341$ areas) would potentially allow for achieving less errors in selecting high-risk areas given a threshold for the $\widehat{\pi}_i$'s. But unfortunately the *BYM mix* model cannot take advantage of that, in fact, for a decrement of the "realized" FDR values we do not meet the same decrement for $\widehat{FDR}$ values estimated by the $\widehat{\pi}_i$'s worked out by *BYM mix*. Note that the spread between true and estimated FDR is stronger in risk correlated scenarios ($S2$ and $S3$) than in $S1$; we will see in next section as this bigger spread does not mean that $\widehat{FDR}$ based rules of spatially correlated scenarios are the less sensitive. By looking at figures (A.1, A.2 and A.3) we see over-estimation is made less evident for θ and SF high levels scenarios (where we have non small areas and high θ value in HR areas). Since the over-estimation, in such $n = 69$ scenarios we achieve a conservative control, we do not necessarily need to determine rules that control only very small FDR values in order to avoid a loss of specificity (as in $n = 19$ scenarios). We can fix $\widehat{FDR} = 0.10$ or $\widehat{FDR} = 0.15$ based selection rules being however aware that we cannot avoid to incur a loss of sensitivity, unless for high θ and high SF levels. Finally, also for the $n = 69$ case, we report in Figure 4.11 box-plot underlying picture of the $\widehat{FDR}$ and "realized" FDR. Also in this case the "realized" FDR is clearly right asymmetric, but oppositely to the $n = 19$ case such FDR (given t_π) empirical distributions are more concentrated. Indeed, this is due to the fact that we have a larger number of true alternative hypotheses and we will generally reject more null hypotheses, so making both numerators and denominators of FDR values higher and eventually obtaining less variability among the 100 datasets than in the $n = 19$ scenarios; see Figure 4.12.

4.3.2 The *BYM mix* power in identifying at risk-areas by $\widehat{FDR}$ based selection rules

We show results of sensitivity/specificity of FDR based selection rules by plotting the sn (4.12) and sp (4.10) against each possible FDR based selection rules in the horizontal axis. In appendix are shown results about all 18 scenarios grouping $S1$, $S2$ and $S3$ in the same window plot, see Figure A.7 and A.8. Here we focus the discussion on figure 4.13 that plots all spatial scenarios sensitivity and specificity results in the same window, for one of the cases analyzed in this section

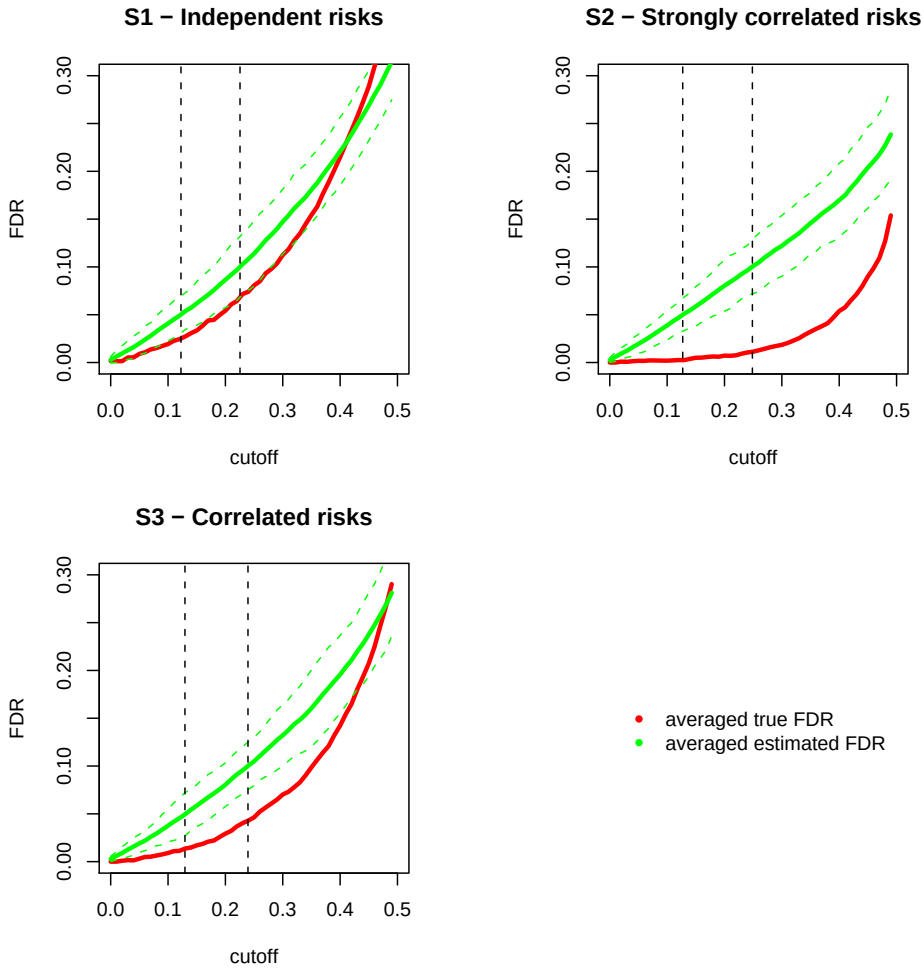


Figure 4.6: *BYM mix* model, $n = 69$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. The dark dashed vertical lines signal the cut-off value corresponding to having determined a selection rule based on $\widehat{FDR}$ equal to 0.05 and 0.10 respectively. We could suggest to use both selection rules, even if they are affected by a lack of sensitivity since the very conservative FDR estimation.

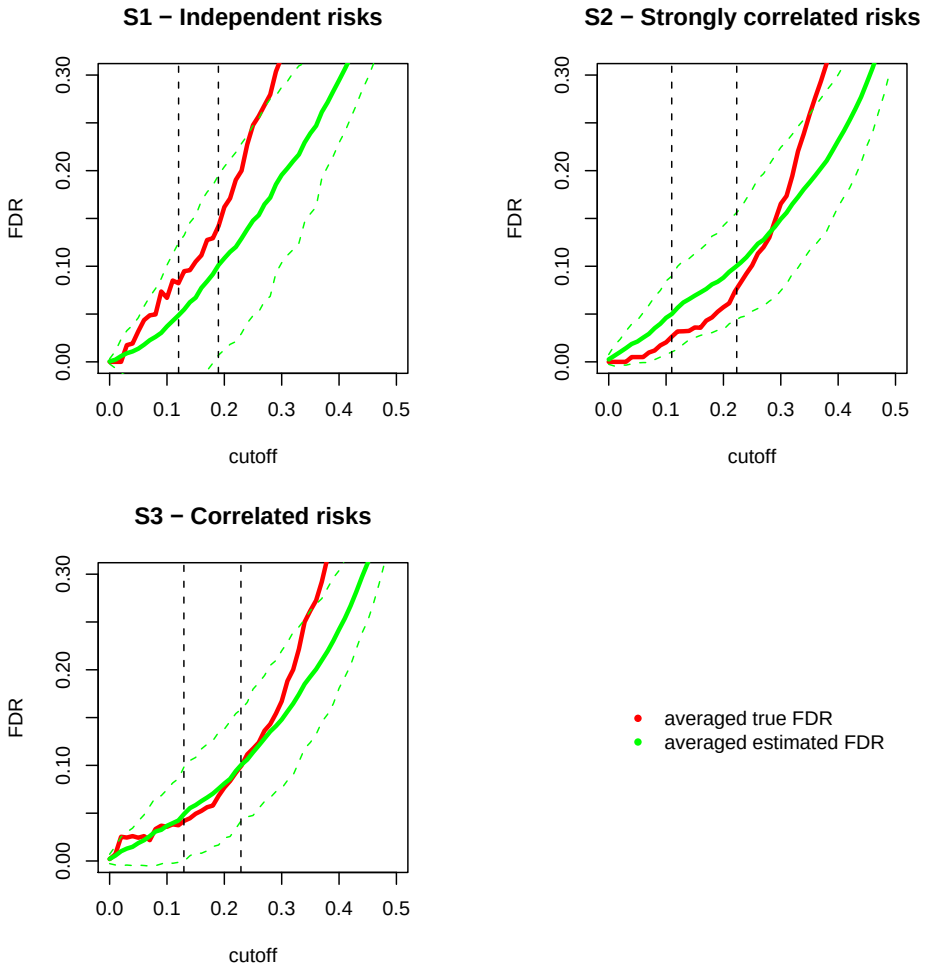


Figure 4.7: *BYM mix* model, $n = 19$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. The dark dashed vertical lines signal the cut-off value corresponding to having determined a selection rule based on $\widehat{FDR}$ equal to 0.05 and 0.10 respectively. We can suggest to use both selection rules in spatially correlated cases. Moreover an $\widehat{FDR} = 0.15$ based rule could be used as well without incurring in loss of specificity since the accurate FDR estimation (non-under estimation).

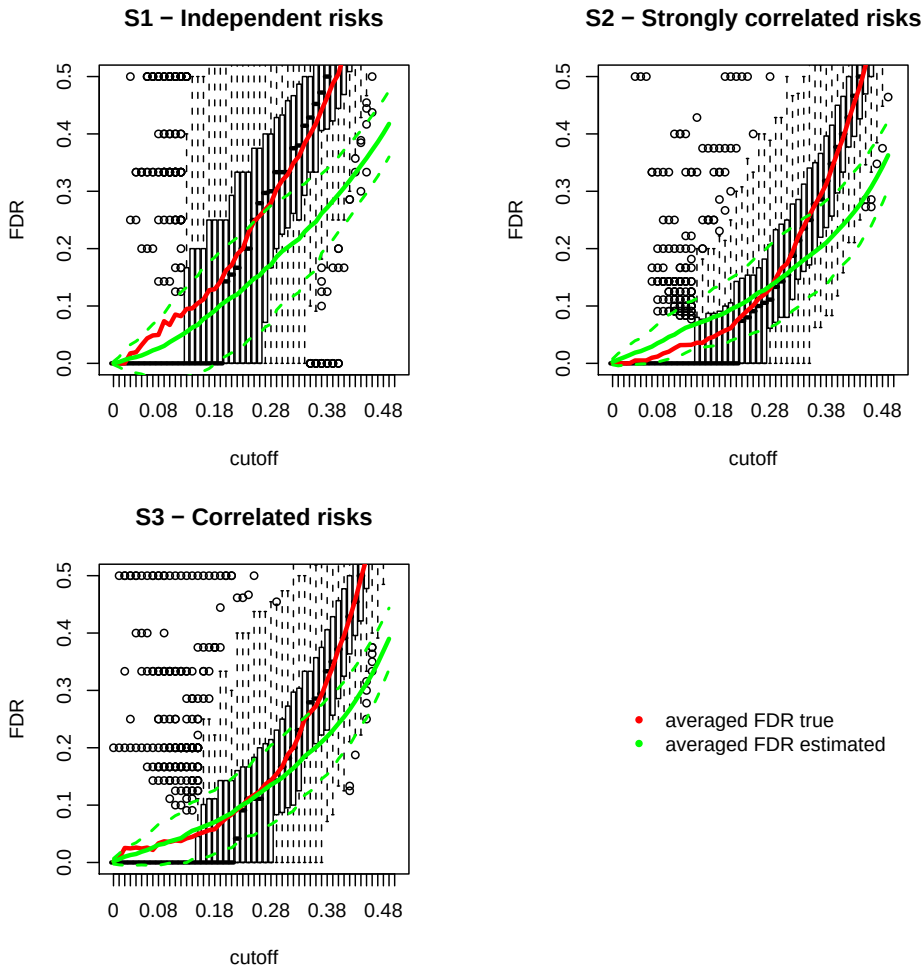


Figure 4.8: *BYM mix* model, $n = 19$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. Box-plots are relative to "realized" FDR empirical distributions conditional on each t_π.

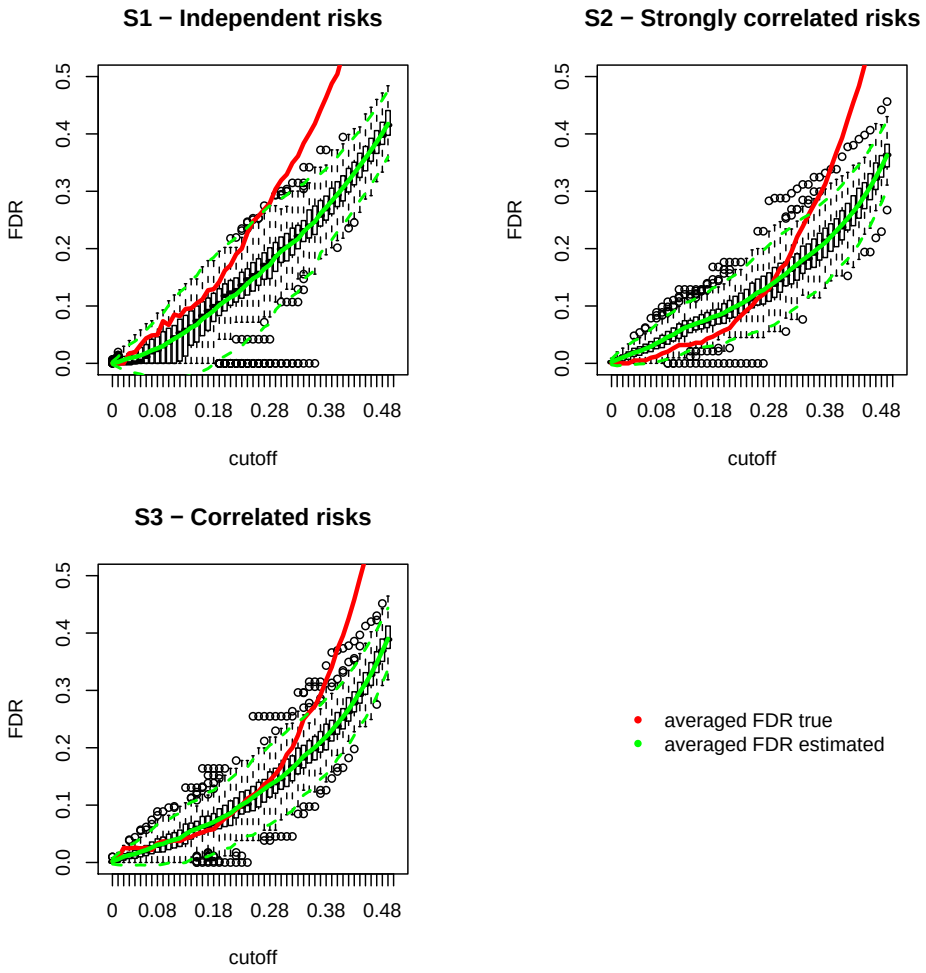


Figure 4.9: *BYM mix* model, $n = 19$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. Box-plots are relative to $\widehat{FDR}$ empirical distributions conditional on each t_π.

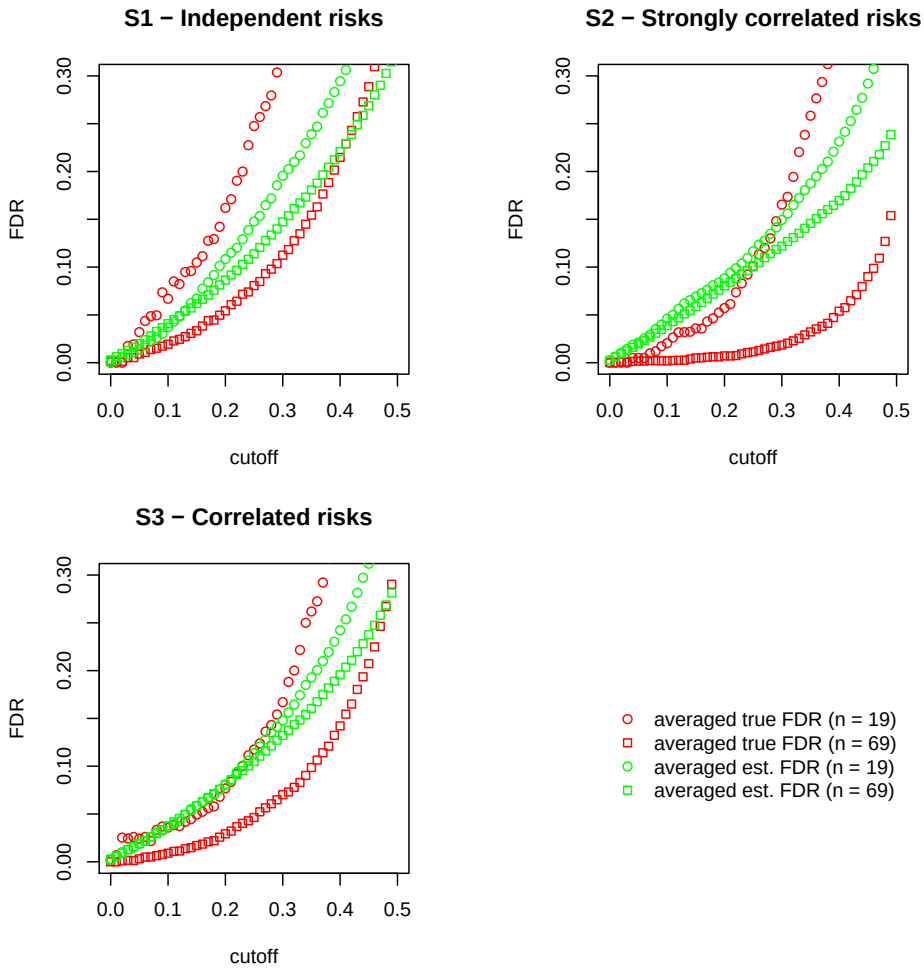


Figure 4.10: *BYM mix* model, both $n = 69$ and $n = 19$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. Moving from $n = 19$ to $n = 69$ we see a decrement of the true FDR values (red lines), whereas the $\widehat{FDR}$ values (green lines) are almost unaffected.

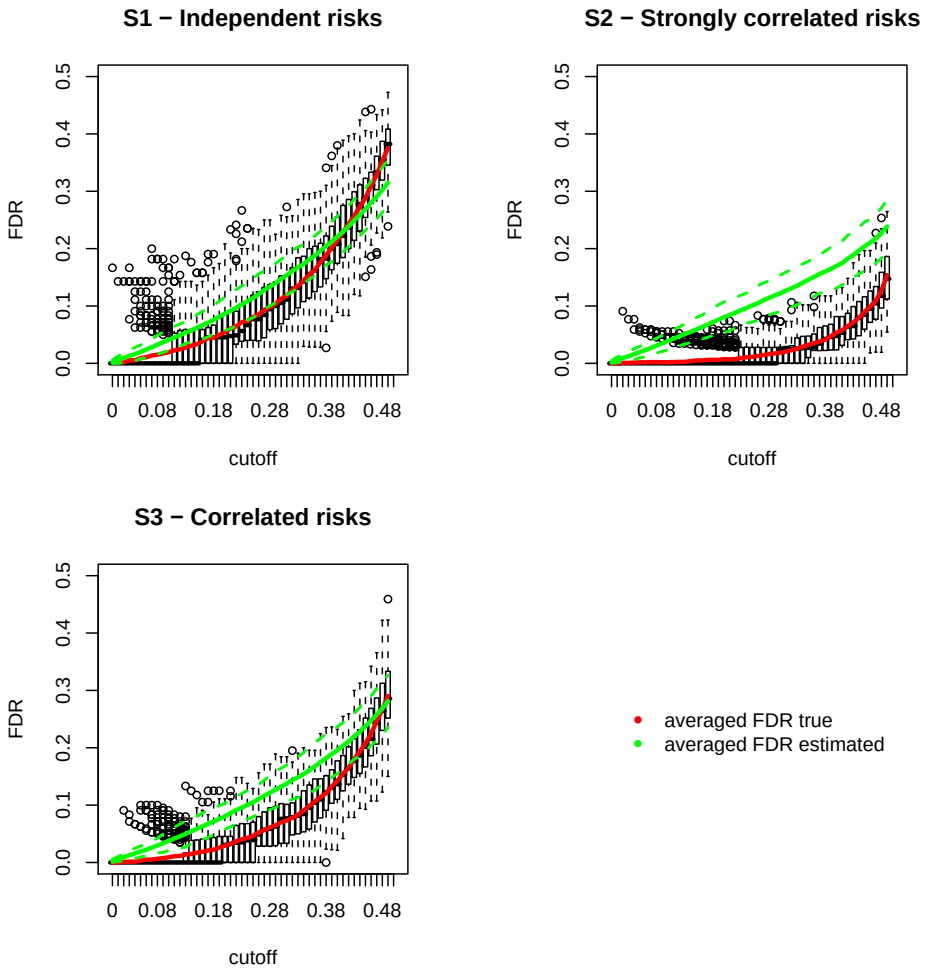


Figure 4.11: *BYM mix* model, $n = 69$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. Box-plots are relative to "realized" FDR empirical distributions conditional on each t_π.

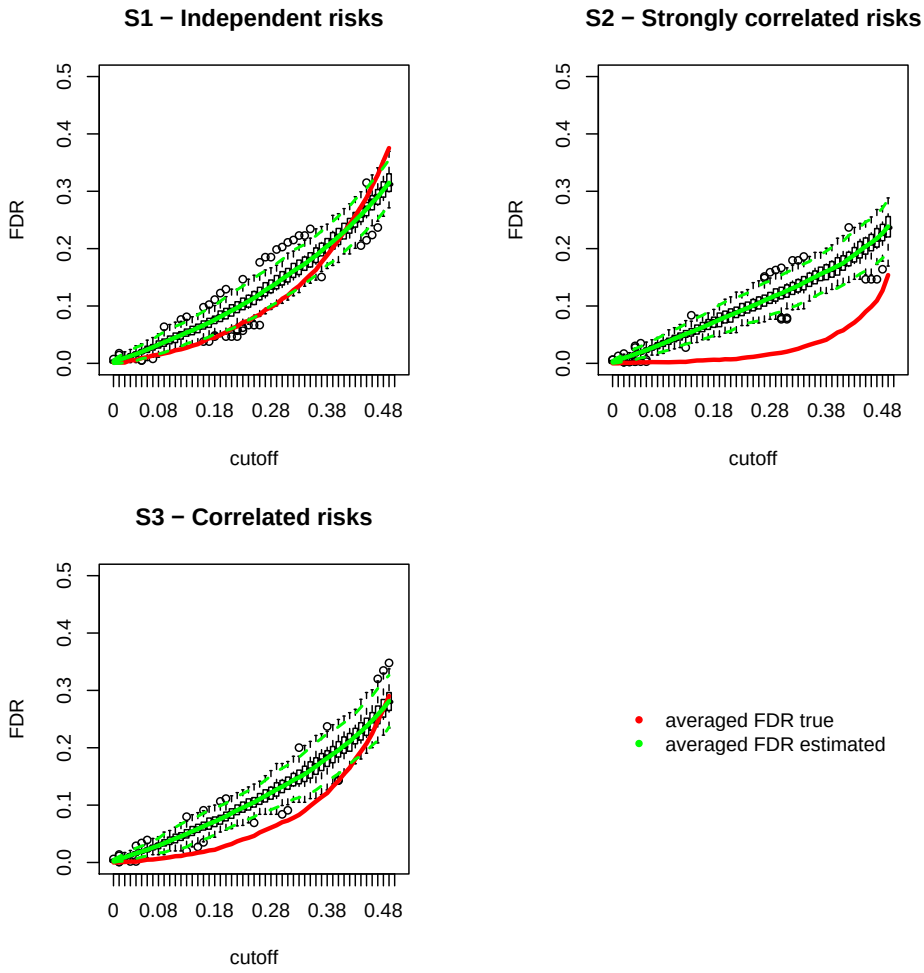


Figure 4.12: *BYM mix* model, $n = 69$, $\theta = 1.5$, $SF = 1$. FDR, $\widehat{FDR}$ vs t_π for $S1$, $S2$, $S3$ spatial scenarios. Box-plots are relative to $\widehat{FDR}$ empirical distributions conditional on each t_π.

($n = 19$, $SF = 1$ and $\theta = 1.5$). Moreover, we plot the "potential sensitivity" (potential *sp*) and the "potential specificity" (potential *sp*). As an example to understand this concept, let us suppose to fix an $\widehat{FDR} = 0.05$ based selection rule. The practitioner, after having obtained the $\widehat{\pi}_i$ by the *BYM mix* model, can choose the threshold t_π by selecting as many areas as possible such that $\widehat{FDR} \geq 0.05$; see (3.13). The *sn* and *sp* values that correspond to the 0.05 value in the horizontal axis has been calculated conditional on t_π. The potential *sn* an potential *sp* are instead calculated conditional to a different threshold, i.e. the threshold found out with the same mechanism as in (3.13) but considering the FDR realized by simulation instead of the $\widehat{FDR}$. If the model under-estimate the FDR the potential *sn* will be lower than the actually achieved *sn* (we will obtain a loss of specificity), whereas if the model over-estimate the FDR the potential *sn* will be greater than the actually achieved *sn* (loss in sensitivity). Thus, we can know the sensitivity potentially achievable if the *BYM mix* model had have carry out an unbiased estimated FDR. For instance, in figure 4.13 for scenarios $S3$ the potential and achieved sensitivity values are approximately equal since the FDR is accurately estimated, whereas for scenarios $S2$, $\widehat{FDR} > 0.15$ based rules decrease their specificity since the model under-estimates FDR levels greater than 0.15; recall figure 4.7 to check the under-estimation beyond 0.15 FDR levels.

An interesting point is that we reach more sensitivity in correlated spatial scenarios; for instance, in $S2$ and $S3$ scenarios, given a $\widehat{FDR} = 0.1$ based selection rule we obtain a sensitivity between 40% and 60% maintaining yet high levels of specificity. This is due to the flexible 2nd stage prior introduced in the *BYM mix* model that allows to take advantage of cases where a positive spatial correlation between risks is present.

Look at all results in Appendix (figures A.7 and A.8) for a global view on the sensitivity issue in all scenarios. We give a brief a description of what can be found in one of the two figures. We have nine sensitivity/specificity graphs representative of 3 SF scenarios ($SF = 0.5$, $SF = 1$, $SF = 5$) times 3 θ scenarios; keeping fixed the factor n. Figure A.8 for example focuses on nine scenarios with $n = 19$ (around a 5% of the true high-risk areas). Each row, containing three figures, corresponds to a different θ value, the first on the top being $\theta = 1.5$, the second $\theta = [1.2 \div 2]$ and the third $\theta = 2$. Each column, containing three figures, corresponds to a different SF value, the first column on the left being $SF = 0.5$ (very small areas), the second $SF = 1$ (small areas) and the third on the right $SF = 5$ (non-small areas). In each of the nine graphs the three spatial scenarios ($S1$, $S2$, $S3$) are grouped to emphasize the different degrees of sensitivity that they can achieve.

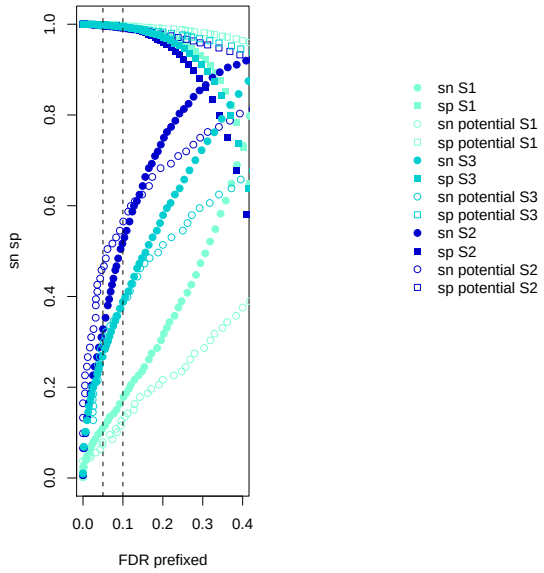


Figure 4.13: *BYM mix* model, $n = 19$, $\theta = 1.5$, $SF = 1$, all spatial scenarios. *sn*, *sp*, potential *sn*, potential *sp*, *vs* $\widehat{FDR}$. Looking at the dashed lines we see that an $\widehat{FDR} = 0.10$ based rule yields a sensitivity from 0.4 to 0.6 in spatially correlated risks scenarios ($S2$ and $S3$) with high level of specificity; independent risk scenarios achieve small power. Note that when the potential *sn* is bigger than *sn* (the achieved or actual sensitivity) we are over-estimating FDR; in the opposite case we are under-estimating FDR (see the corresponding loss of specificity). $\widehat{FDR} > 0.15$ based rules lose specificity.

4.3.3 The *BYM mix* performance on relative risk estimation

The idea of making inference on all area-specific null hypotheses and at the same time on all area-specific relative risk values is appealing. We saw there are contexts, not far from practical applications, where $\widehat{FDR}$ based selection rules are good, in the sense that they allow for a conservative control.

Besides controlling the FDR, or equivalently selecting high-risk areas such that the FDR is non-greater than a prefixed level, the practitioner may be interested into the evaluation of relative risk values too. Thus, we pursue a comparison between the classic BYM and BYM *mix* models in terms of closeness between posterior estimates $\widehat{r}_i$'s and true relative risk values (in the n areas where the alternative hypothesis is true). As it is well know, Bayesian estimation often leads to posterior estimates that are over-shrunk towards a global mean, in this particular models both towards a global and a local (the neighborhood) mean. By looking at box-plots in Figure 4.14 we can see that, in the usual three spatial scenarios $S1$, $S2$ and $S3$, the posterior relative risks of the HR areas are less smoothed when they are calculated by BYM *mix*; see the BYM *mix* box-plot (mix) is closer to the nominal true risk (red line) than the classic BYM box-plot (cl). This fact is almost spread in all scenarios unless the strongly spatial correlated cases ($S2$) for $n = 69$ where the classic BYM model works slightly better. In general, we can say that by applying the proposed model we do not produce a stronger degree of over-smoothing with respect to the Besag York and Mollié model. Figure A.9 and A.10 in Appendix, give all scenarios results. As usual, we give a brief description of what can be found in one of the two figures. We have nine windows representative of 3 SF scenarios ($SF = 0.5$, $SF = 1$, $SF = 5$) $\times$ 3 θ scenarios, keeping fixed the factor n. Figure A.9 for example focuses on nine scenarios with $n = 19$, (around a 5% of the true high-risk areas). Each row, containing three figures, corresponds to a different θ value, the first on the top being $\theta = 1.5$, the second $\theta = [1.2 \div 2]$ and the third $\theta = 2$. Each column, containing three figures, corresponds to a different SF value, the first column on the left being $SF = 0.5$ (very small areas), the second $SF = 1$ (small areas) and the third on the right $SF = 5$ (non-small areas). In each of the nine windows the three spatial scenarios ($S1$, $S2$, $S3$) are grouped to emphasize the different over-smoothing degrees that they yield.

4.4 Conclusive remarks on the simulation study results

What we are interested to highlight in summarizing the simulation results is firstly the scenarios where we can trust the FDR estimation in order to trust the conclusion we make by means of a $\widehat{FDR}$ based rule. Secondly, we are also interested to know what we may expect if the scenarios

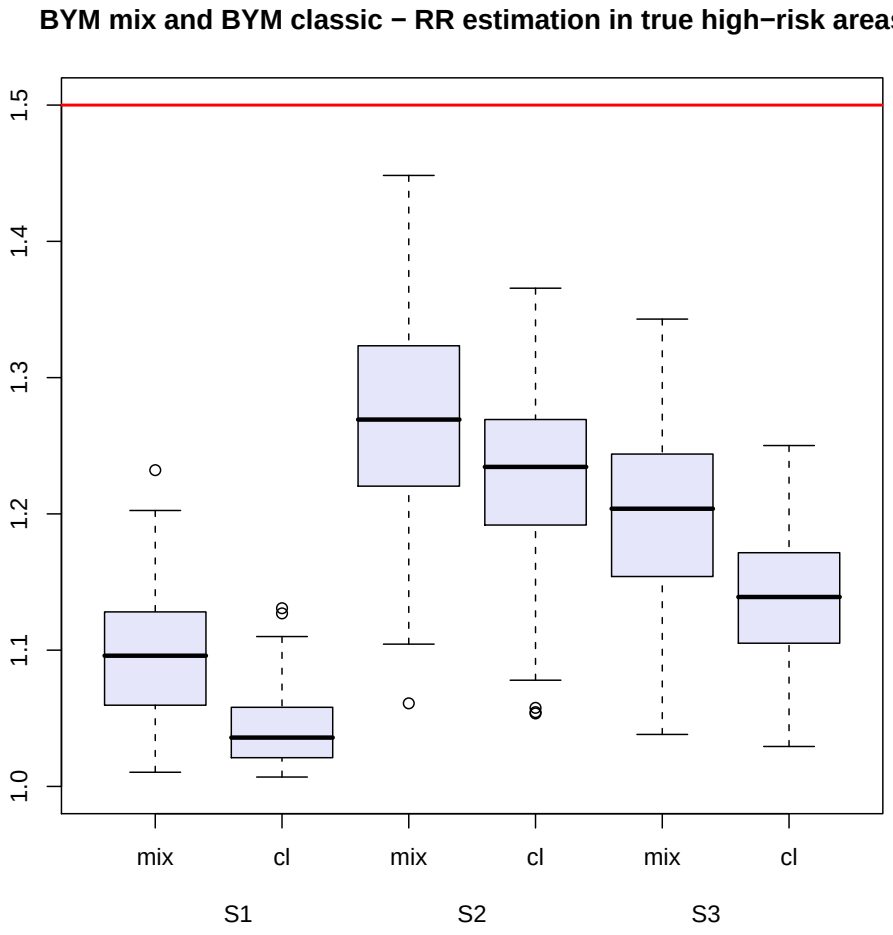


Figure 4.14: *BYM mix* model, $n = 19$, $\theta = 1.5$, $SF = 1$, all spatial scenarios. Box-plots of relative risk values in true high-risk areas. The red line is the true nominal value $\theta = 1.5$ in such scenarios. More precisely the true value is a bit lower since the multinomial model generates counts yielding in HR areas a risk value around $\frac{\theta_{HR}}{\theta_{NR}}$. The degree of over-smoothing is weaker for the *BYM mix* posterior relative risk estimates (mix) than the classic *BYM* posterior relative risk estimates (cl).

where we obtain an accurate FDR estimation change with respect to the factors we control by simulation. On this note we say in advance that we will mostly focus on interpretation of results according to changes in areas size factor (three levels: $SF = 0.5$, $SF = 1$ and $SF = 5$) and spatial correlation factor (three levels: $S1$, $S2$ and $S3$), that are the two sources of variability which the *BYM mix* model aims to capture.

As regards the first point, we saw the small areas, small θ, small n and strong spatially correlated scenarios (both $SF = 0.5$ and $SF = 1$, $\theta = 1.5$, $n = 19$, both $S2$ and $S3$) yield the most accurate FDR estimation. To the other hand, they are not the scenarios where we achieve the most power in detecting the true alternative hypotheses. Power increases as long as risks ($\theta > 1.5$) and size areas ($SF = 5$) get larger, but as a result of this we have two drawbacks: first, we loose specificity even though not dramatically; second, we can trust the estimation of only low values of FDR, even much lower than 0.05 because of the borrowing of strength between contiguous areas posterior probabilities. Borrowing of strength in scenarios where risks are spatially uncorrelated ($S1$), expected counts are not small ($SF = 5$) and risks are high ($\theta > 1.5$) is not necessary and actually produces artifacts because we have enough information by only the empiric observation of crude SMRs.

Therefore, it seems there is a tradeoff between power achievable and goodness of the FDR estimation (more precisely, in goodness of estimation for a wide range of FDR values). We now explain this issue. In practice, the practitioner who chooses, say a $\widehat{FDR} = c$ selection rule, has two targets:

1 selecting areas providing that the true unknown FDR will not be greater than c;

2 the selection rule chosen is powerful maintaining an acceptable level of specificity.

Target **1** corresponds to look for a conservative FDR estimation; in fact, the practitioner, after selecting discoveries, would not want to declare an estimate of the FDR lower than the true FDR (he does not want $\widehat{FDR} <$ true FDR), because if so he will declare a number of errors that is lower than due, hence not achieving the FDR control. This is the reason why over-estimating the FDR is not as bad as under-estimating it, because in the former case we at worst will achieve a too much conservative FDR control ($\widehat{FDR} \geq$ true FDR as in expression (1.4)). The result of over-estimating FDR is indeed the loss of sensitivity, while the result of an under-estimation is the loss in specificity. The latter two points are connected with the target **2**: the practitioner that chooses the $\widehat{FDR} = c$ rule needs such rule to be sensitive for an acceptable level of specificity. This seems a reasonable conservative strategy, since, as long as we know the decision about keeping on investigating by means of individual epidemiological studies is not generally addressed with

the introduction of loss functions which quantify the cost for a false discoveries and a false non-discoveries.

One of the main achievement of the simulation study is that it puts in light that in small areas scenarios ($SF = 0.5$, and $SF = 1$) and in non-high θ values scenarios ($\theta = 1.5$ or $\theta = [1.2 \div 2]$) a specificity around 0.95 is always achieved for selection rules based on $\widehat{FDR}$ values around 0.10 or 0.15 (in $n = 19$ scenarios) and 0.05 (in $n = 69$ scenarios). See figures A.7 A.8 in Appendix 4.4.1 which plot sensitivity and specificity *vs* the $\widehat{FDR}$. Moreover, as long as risks are spatially correlated both point **1** and **2** are easier achieved by the *BYM mix* model, that is in $S2$ and $S3$ spatial scenarios we observed more accurate FDR estimates and more powerful $\widehat{FDR}$ based selection rules than in independence risks scenarios $S1$. Such results suggest that applying the proposed model to control the FDR at 0.05 level may be recommended in many small areas application, and $\widehat{FDR} = 0.05$ based selection rules though poorly sensitive can be a non-arbitrary way to proceed with selecting high-risk areas and simultaneously achieving the FDR control. To argue the usefulness of this approach, it is worth noticing that scenarios where we may advise the practitioner to proceed in high-risk areas selection by means of a $\widehat{FDR} = 0.05$ (or even $\widehat{FDR} = 0.10$) based rule (at worst achieving a conservative FDR estimation) are frequent in practice; they are indeed the cases where there is the presence of small areas, spatial correlation between risks and low relative risk values. Another point in favor of the *BYM mix* model is its ability to estimate the relative risk as well, being hence able to provide two different kind of information to the practitioner. Besides the relative risk values (that we checked are well estimated in almost all scenarios comparing to those obtained by the classic *BYM* model) we can also make inference on many null hypotheses of absence of risk controlling the proportion of false discoveries among the declared discoveries. Therefore, in many frequent practical cases, the proposed model allows us to address, at the same time, a point estimation inference on relative risk values and a multiple testing procedure on many null hypothesis.

4.4.1 An application to a real dataset

We show an application to real data where the aim is to find possible high-risk areas. Observed and expected counts are relative to liver cancer morbidity cases recorded over five years in Emilia-Romagna municipalities. The below summary statistics on the expected counts tells us that small areas are present; around a half of areas have an expected count lower than 5.

```
> summary(e)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.3793   3.4860   5.6040  13.1300   9.4040 491.5000
```

note we are in a case similar to the $SF = 0.5$ scenarios.

In general, we saw via simulation that for small areas scenarios, $FDR = 0.05$ levels are conservatively estimated, hence FDR is not under-estimated. For instance, in $n = 19$ scenarios, for FDR values lower than 0.10 we observed a weak over-estimation, while in $n = 69$ scenarios we achieve a strong over-estimation unless for very small FDR levels. Thus, we believe a control of the FDR at a level 0.05 can be suggested, the level of specificity for $\widehat{FDR} = 0.05$ being generally high.

The OpenBugs code to estimate the model is the same as presented in section 3.2.1. Before drawing samples for inferences, we checked that posterior probability estimates and consequent FDR estimation did not change for different prior specifications of precision parameters τ_u and τ_v of the clustering and heterogeneity terms. We tried the $Gamma(0.5, 0.0005)$ and the $Uniform(0, 100)$ on the standard deviations. We used equal prior specification for both τ_u and τ_v not considering here the fair prior specification (3.8). Below we show the maps relative to the observed SMR's (figure 4.15), the posterior $\widehat{r_i}$'s estimates by BYM model (figure 4.16), the posterior $\widehat{r_i}$'s estimated by the proposed BYM *mix* model (figure 4.17), the discoveries at $\widehat{FDR} = 0.05$ level (figure 4.18). Note that the map of posterior relative risk estimates by BYM and by BYM *mix* are very close; it seems that BYM *mix* and BYM models are able to depict the same risk pattern. In addition, BYM *mix* model can provide the information about the number of discoveries at a pre-specified FDR level. In figure 4.18 different red colors identify the posterior probability (that the area is at null risk) estimates by the model. We find out 20 possibly high-risk areas by an $\widehat{FDR} = 0.05$ based selection rule, i.e. 20 high-risk areas being aware that one of them (the 5% percent) will probably be a false discovery.

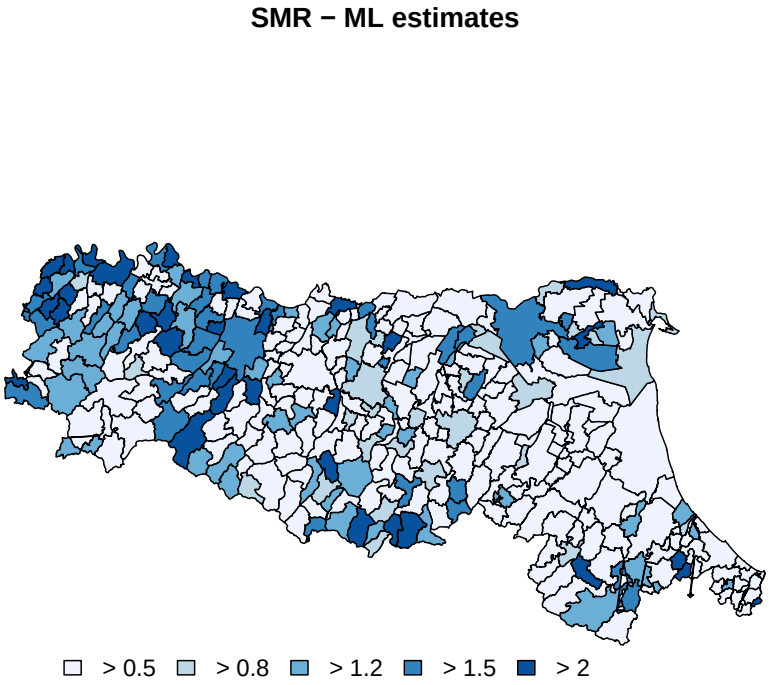


Figure 4.15: Map of SMRs.

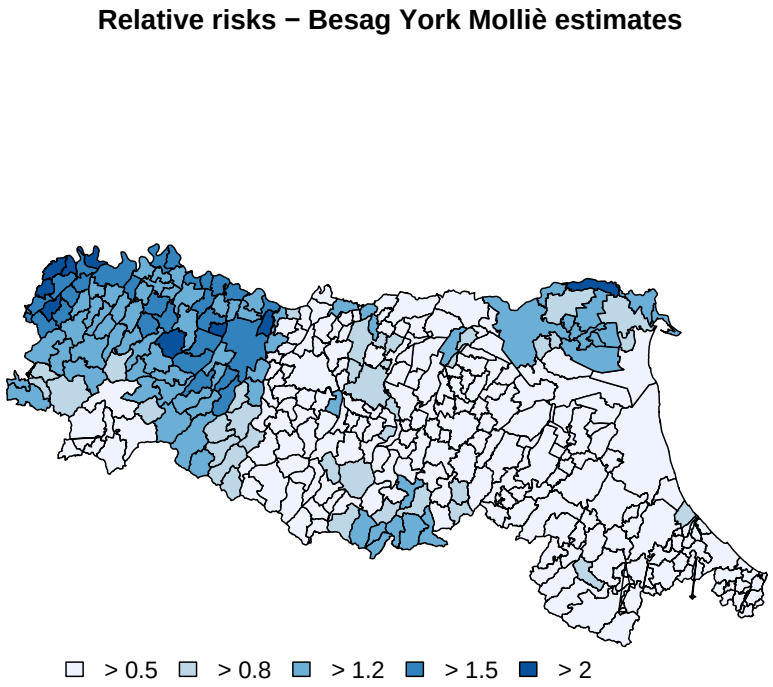


Figure 4.16: Map of posterior relative risk estimates by Besag York Molliè (BYM) model.

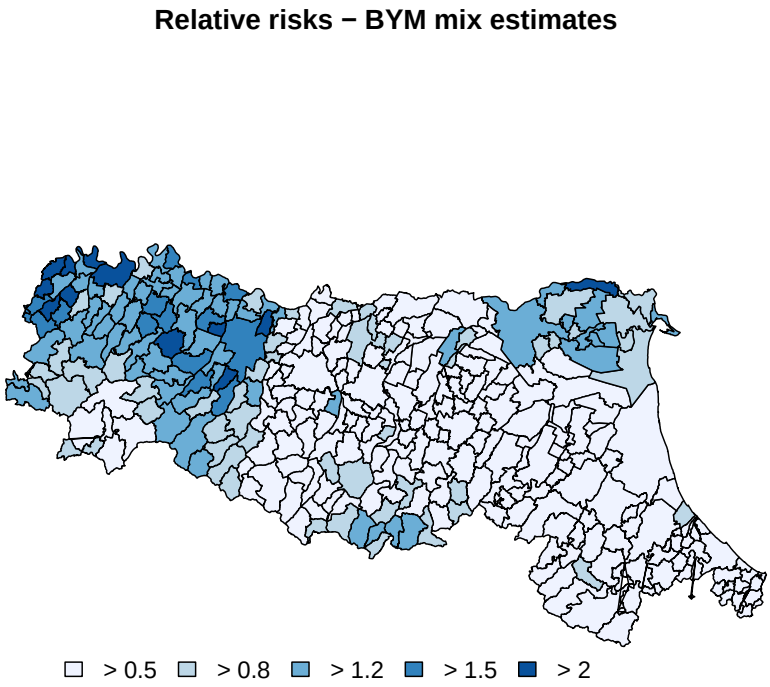


Figure 4.17: Map of posterior relative risk estimates by our proposal (*BYM mix*) model.

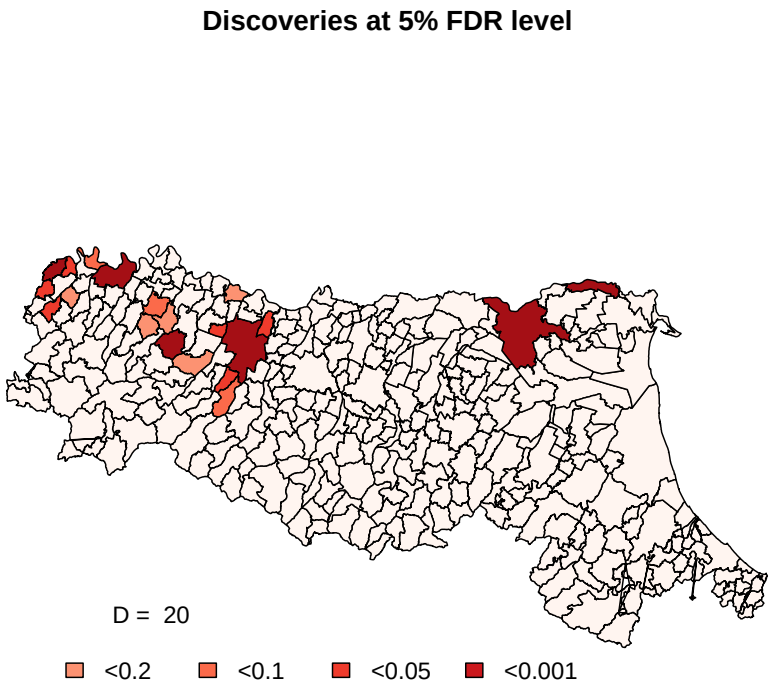


Figure 4.18: Map of discoveries (high-risk areas) selected by an $\widehat{FDR} = 0.05$ based decision rule. Legend is relative to the posterior probability estimated values.

Conclusions and perspectives

The idea of the work is to propose a new methodology for conducting a descriptive analysis in spatial epidemiology which is based on the control of the False Discovery. Since methodology developed can estimate both relative risks and FDR, this approach gives potentially more information than a disease mapping analysis which allows only visual inspection of risks on the map. Indeed, performing a multiple testing procedure on the multiplicity of risk indicators gives a more decision-oriented approach which yields two interesting features. Firstly, it means that inferences on many null hypotheses can be made controlling a global error that occurs in making rejections. On this note, the choice of the FDR as a global error measure is strategic as it allows a less conservative control than, for instance, controlling the $FWER$, that is the probability of at least a false discovery. In many multiple testing case studies where strict control is not of interest the FDR is fruitful; indeed, in the case under examination, the decision to conduct more investigations (on the whole map under study) need not be erroneous even if more than one null hypothesis is falsely rejected. Secondly, such an approach allows the selection of high-risk areas by means of non-arbitrary rules. By a non-arbitrary rule, we mean a rule that can achieve control of the FDR, such that the practitioner is aware of the proportion of false discoveries. To this purpose, a model which produces an estimate of the FDR allows the practitioner to reject null hypotheses by fixing a priori a desired FDR level. This can provide a valid alternative to the practice of selecting high risk areas from knowledge only of the posterior relative risk estimates without any concern about the multiple testing issue. Therefore, the main contribution of the work is the proposal of a model for simultaneously estimating posterior relative risks and controlling the FDR so being able to determine a desired $\widehat{FDR}$ based selection rule for detecting possibly high-risk areas while being aware of the errors we incur.

The proposed model is thought to address cases where a large datasets of SMRs is collected over many spatial contiguous regions, particularly when small areas are present and risks are spatially correlated. The model specifies each area-specific log relative risk distribution as a mixture of two components. Under the null hypothesis (of a relative risk equal to 1) the observed count is assumed as a realization from a Poisson with mean the expected count. Under the alternative

hypothesis (of a relative risk greater than 1) the Besag York Mollié model is assumed as true. The latter is usually employed in disease mapping applications since it has enough flexibility to capture the over-dispersion in the data. Then, through hierarchical modelling we can specify prior random effects that can capture the spatially structured and unstructured extra-Poisson heterogeneity. The Bayesian borrowing of strength between prior and empirical information allows us to evaluate any area-specific null hypothesis by means of all observations in the map, incrementing the power, especially in small areas. Posterior probability (that the null hypothesis is true) estimates are worked out by MCMC computation and form the basis for the estimation of the expected FDR conditional on data. Indeed, the posterior probability that the null hypothesis is true in area i is an estimate of the *typeI* error probability in declaring the area i as a high-risk area (or as a discovery). To evaluate the alternative hypothesis of an incremented risk we need to take in account only the posterior probabilities relative to areas where the observed count is non lower than the expected. Finally, an estimate of the expected FDR conditional on data can be obtained given any set of discoveries by averaging their respective posterior probabilities. Estimates of the relative risk values can be provided through MCMC computation as well.

The simulation study

For an $\widehat{FDR}$ based selection rule to be useful for the practitioner we need a model which yields a conservative estimate of the FDR. The simulation study was set up to answer the question whether the model can conservatively estimate the FDR in cases that are frequent in practice, that is small areas and spatially correlated risks cases. Results show that FDR is well estimated (at worst we get an over-estimation, hence a too conservative FDR control) in small areas, low risk levels and spatially correlated risks scenarios. In such scenarios we have good estimates of the FDR for all values less or equal than 0.10. The sensitivity of $\widehat{FDR}$ based decision rules is generally low but specificity is high. Thus, in such a scenario the use of $\widehat{FDR} = 0.05$ or $\widehat{FDR} = 0.10$ based selection rules can be suggested. In cases where the number of true alternative hypotheses (number of HR areas) is small, also $FDR = 0.15$ values are well estimated, and $\widehat{FDR} = 0.15$ based decision rules gain power maintaining high specificity. On the other hand, in non-small areas and non-small risk level scenarios the FDR is under-estimated unless for very small values (i.e. for $\widehat{FDR}$ much lower than 0.05); this results in a loss of specificity of a $\widehat{FDR} = 0.05$ based decision rule. In such scenarios $\widehat{FDR} = 0.05$ or, even worse, $\widehat{FDR} = 0.1$ cannot be recommended because the true proportion of false discoveries is actually much higher than that predicted by the model. As regards relative risk estimation, our model achieves almost the same results as the classic Besag York Molliè model, yielding in most cases a weaker degree of over-smoothing of the posterior estimates.

As regards the simulation study undertaken a number of different approaches could have been considered. The way we chose was aimed to strictly control the factors of interest: the number of true null hypotheses, the risk level, the size of areas. In particular simulating from a multinomial after generating the three spatial scenarios (with different degrees of spatial correlation), allows us to control a precise value of the risk level in areas where the alternative is constrained to be true. Moreover, to avoid misinterpretation of results we needed to make some constraints, such as keeping fixed the proportion of small areas (areas where $e_i < 5$), and that the sum of expected count between the three spatial scenarios was equal.

Future developments

A different simulation set up could be interesting under a different point of view even although it could not straightforwardly achieve the above objectives. For instance, simulating from the model itself or from the Besag York Mollié model would have been useful for generating datasets relative to a wider range of spatially correlated scenarios. This can be accomplished by manipulating the precision parameters of the prior random effects introduced at the third stage of the model. Indeed, the relative magnitude of these parameters can guide the spatially and non-spatially structured amount of extra-Poisson variability. Moreover, the magnitude of such parameters in themselves can guide the relative risk level in areas where by simulation it is greater than 1. On the other hand, as said, we cannot strictly control the factors of interest. Thus, we believe the choice undertaken allows us to create a wide range of scenarios, controlling precise factors of interest, for checking the model performance both in cases often encountered in practice (small areas scenarios) and less frequent (high risk values and non-small areas).

As regards FDR estimation in such epidemiological case studies, the model proposed does not have a model with which it is being compared. It will be interesting to extend other disease mapping models to the FDR estimation and compare their performance via simulation. A natural extension of the model proposed is the introduction of covariates for instance.

For the Besag York Mollié model, several works have been published in literature about the sensitivity to the choice of random effects precision parameters. On this note, for working out simulation results we chose Gelman's proposal of a uniform on the standard deviation instead of the usual conjugate Gamma distribution on the precision. As regards real data applications we checked the choice of the above two specifications did not change results for the number of high-risk areas discovered with an $\widehat{FDR} = 0.05$ based rule.

However, our model presents slightly different features w.r.t. the Besag York Mollié model since it introduces a mixture where only if the alternative hypothesis is true there is the Bayesian learning

on the clustering and heterogeneity terms. Indeed, we noticed by a simulation (whose results we did not show here) that our model under-estimates the precision of random effect terms if compared with the Besag York Mollié model posterior estimates. Thus a simulation study aiming to check the sensitivity of posterior probability estimates to different prior specifications is an interesting future project line, and may be pursued considering a number of datasets simulated with different precision parameters values.

Appendix A

All results

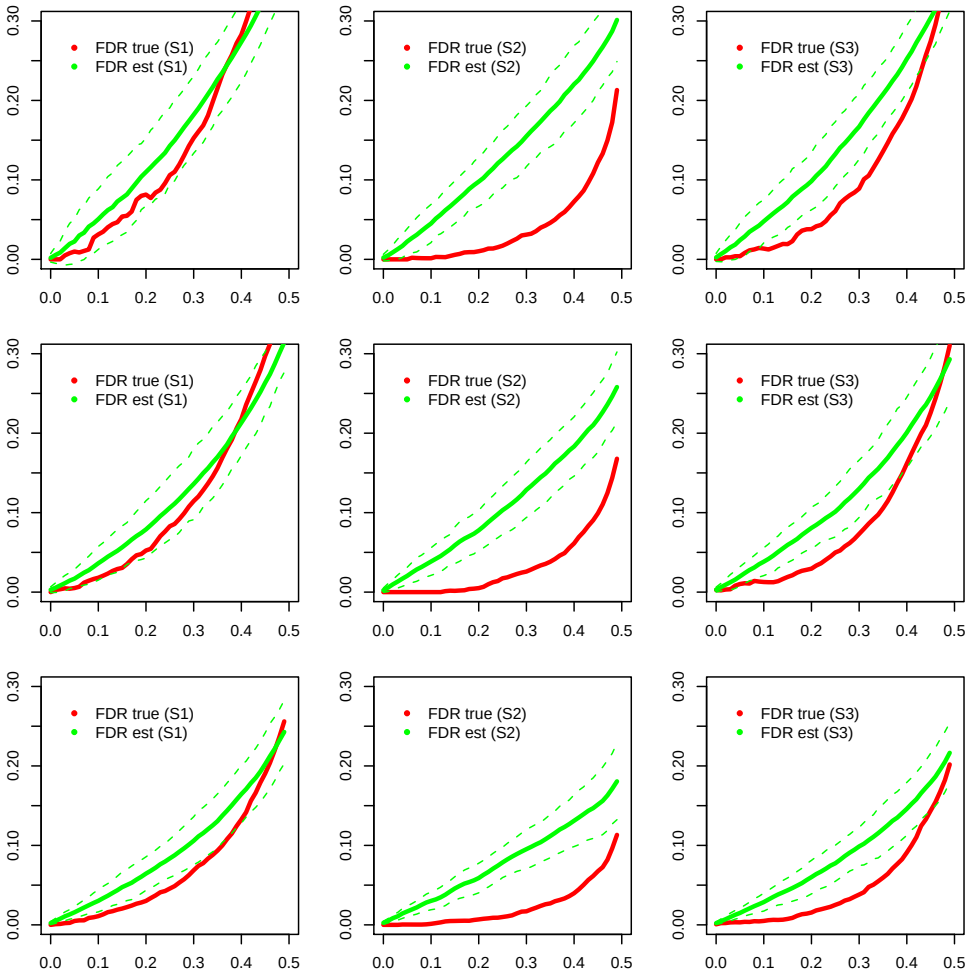


Figure A.1: *BYM mix* model, $n = 69, SF = 0.5$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. Here and in the following figure (A.2), we see that the most FDR over-estimation (conservative FDR estimation) is for spatially autocorrelated cases ($S2$ and $S3$). By looking at figure A.7 we however realize that sensitivity is greater for $S2$ and $S3$; hence the over-estimation degree informs us on the loss of sensitivity (in $S2$ and $S3$) with respect to what could be potentially achieved if $\widehat{FDR} \equiv$ true FDR.

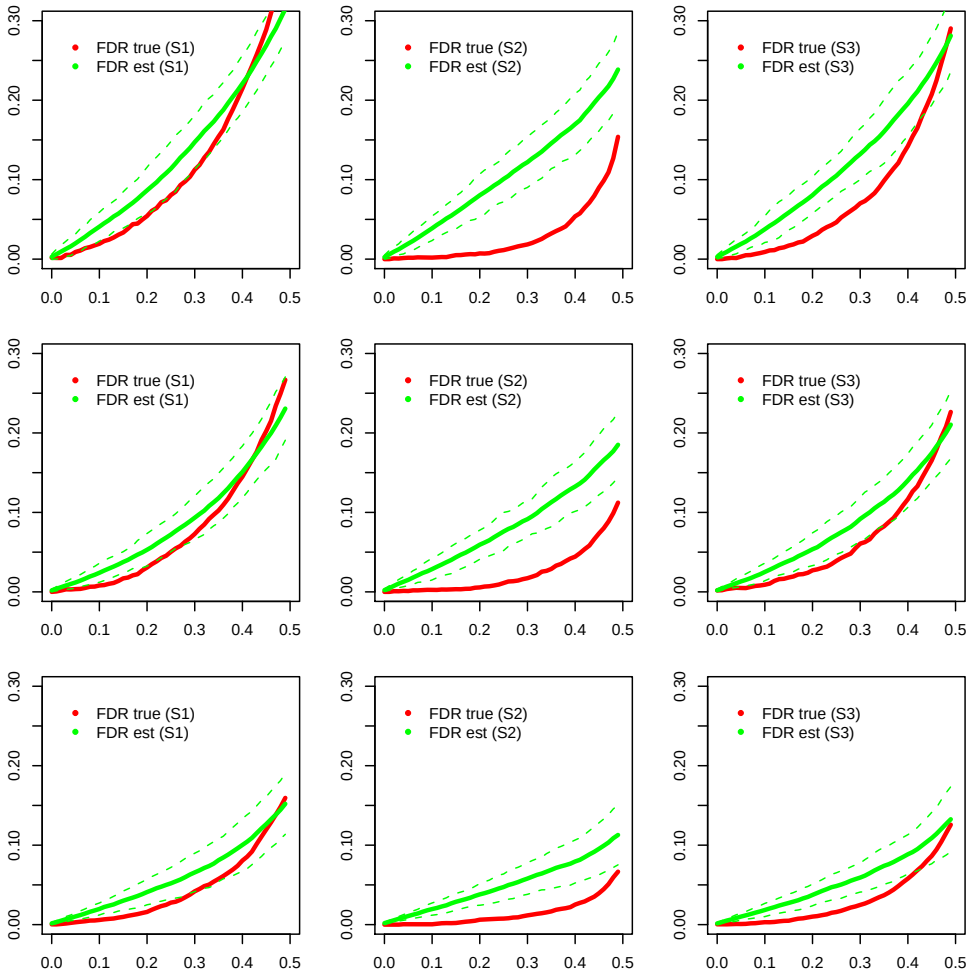


Figure A.2: *BYM mix* model, $n = 69, SF = 1$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. Look at comments for figure A.1.

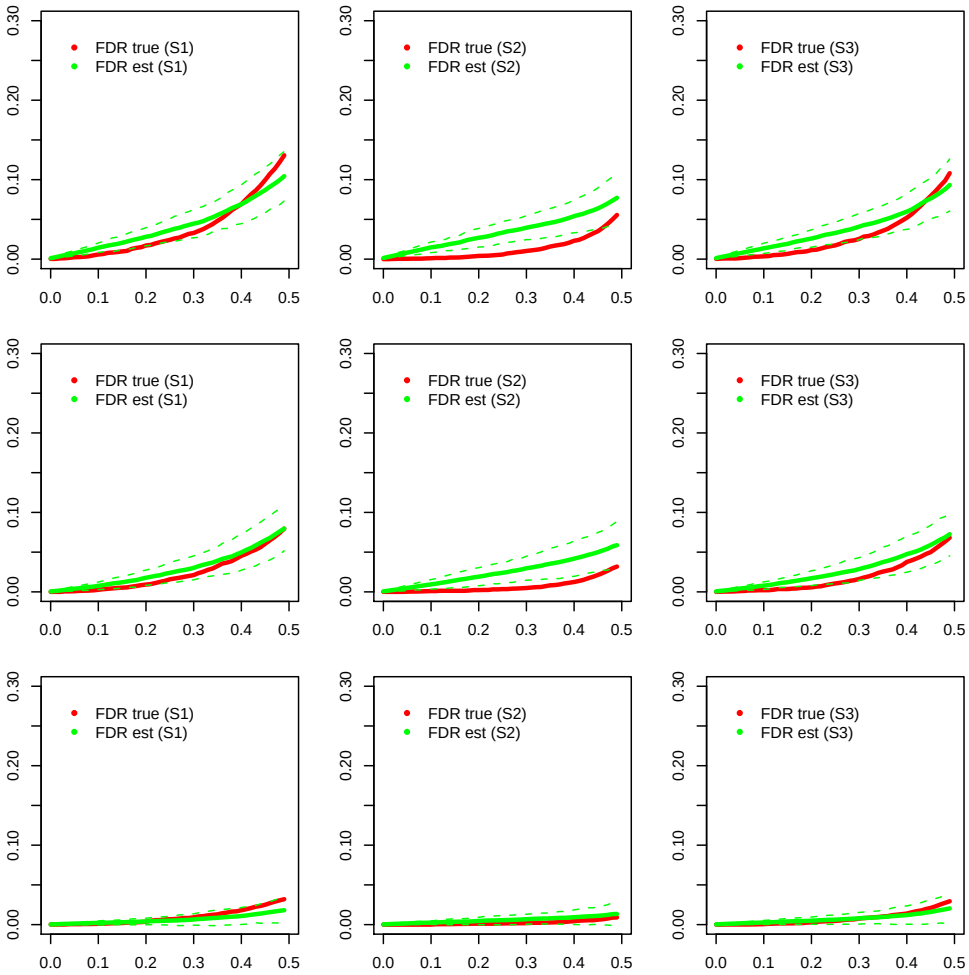


Figure A.3: *BYM mix* model, $n = 69, SF = 5$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. With such non-small areas scenarios we obtain a good FDR estimation but only for very small FDR values.

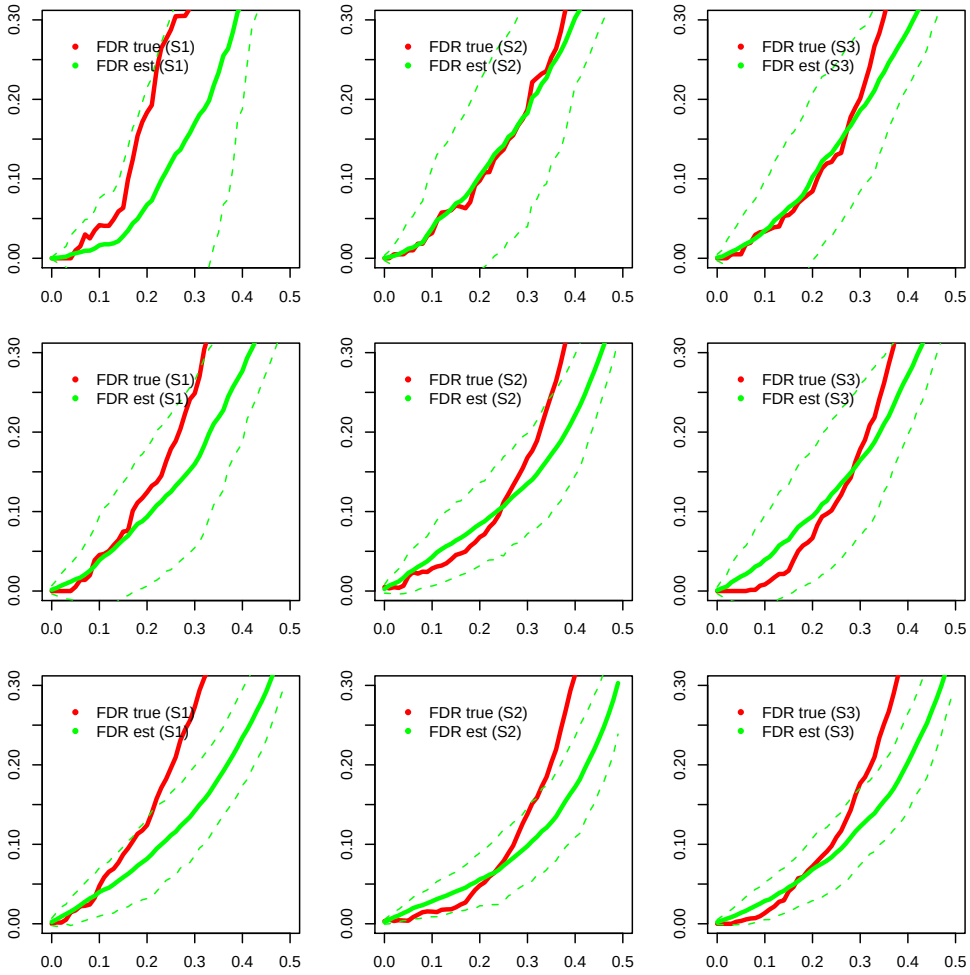


Figure A.4: *BYM mix* model, $n = 19, SF = 0.5$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. In spatially correlated risks ($S2$ and $S3$) and non-high θ scenarios (first two rows) we can suggest the use of $\widehat{FDR} = 0.10$ based selection rules. Higher FDR levels could probably yield a loss of specificity since the FDR under-estimation. These scenarios occur frequently in practice.

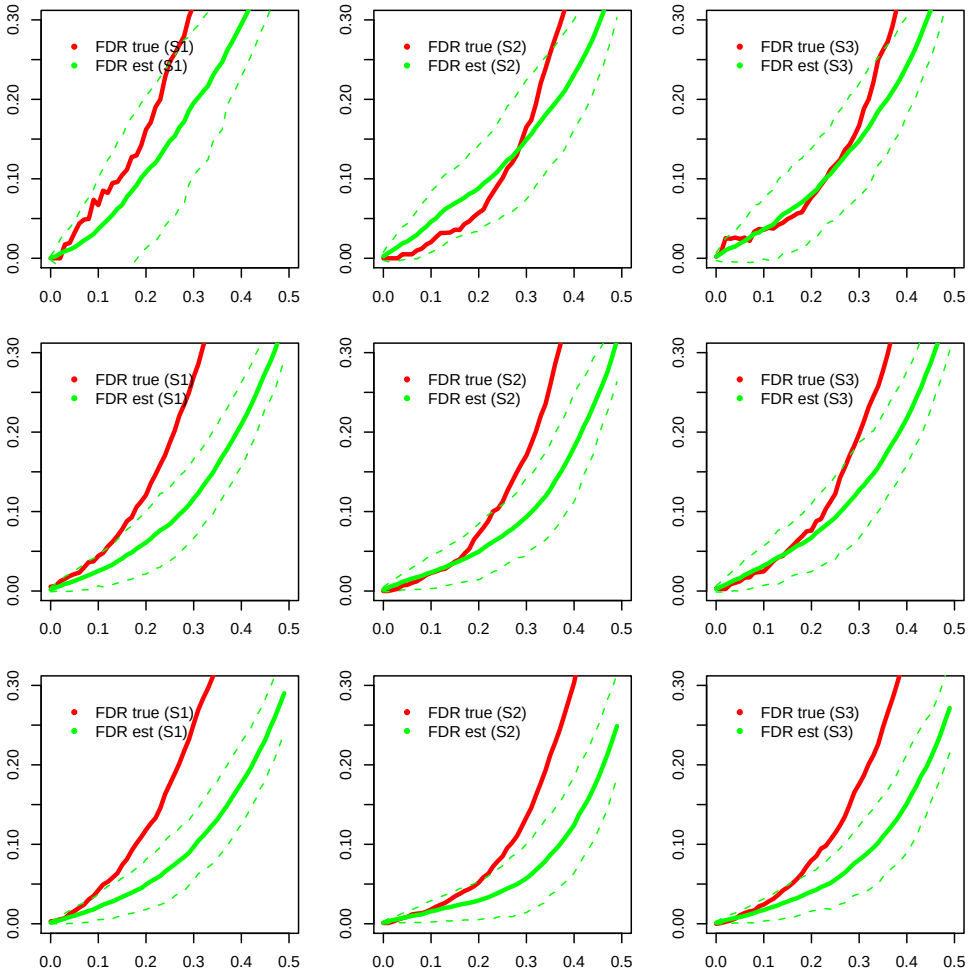


Figure A.5: *BYM mix* model, $n = 19, SF = 1$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. In spatially correlated risks ($S2$ and $S3$) and non-high θ scenarios (first two rows) we can suggest the use of $\widehat{FDR} = 0.05$ based selection rules. Higher FDR levels could probably yield a loss of specificity since the FDR under-estimation. These scenarios occur frequently in practice.

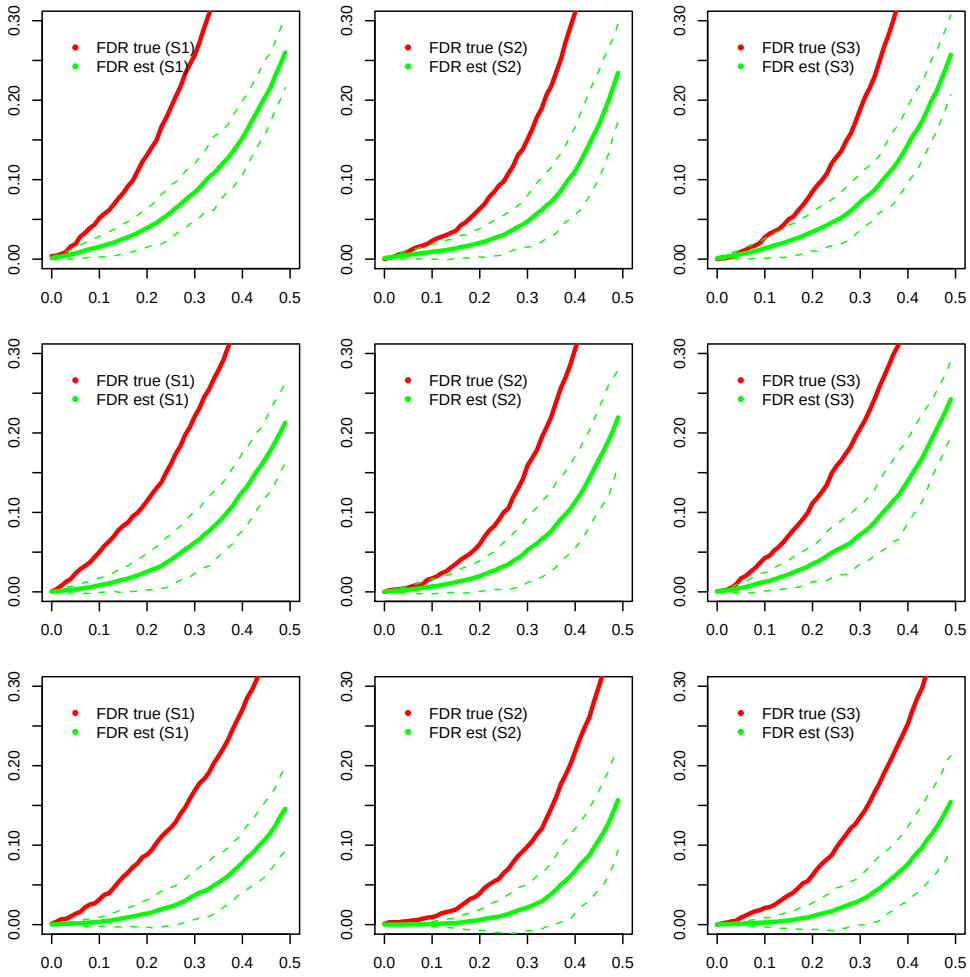


Figure A.6: *BYM mix* model, $n = 19, SF = 5$. In each of the nine figures is plotted FDR (FDR true) and $\widehat{FDR}$ (FDR est) vs t_π. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $S1$ $S2$ $S3$ scenarios. In such scenarios FDR tends to be under-estimated except for very small values.

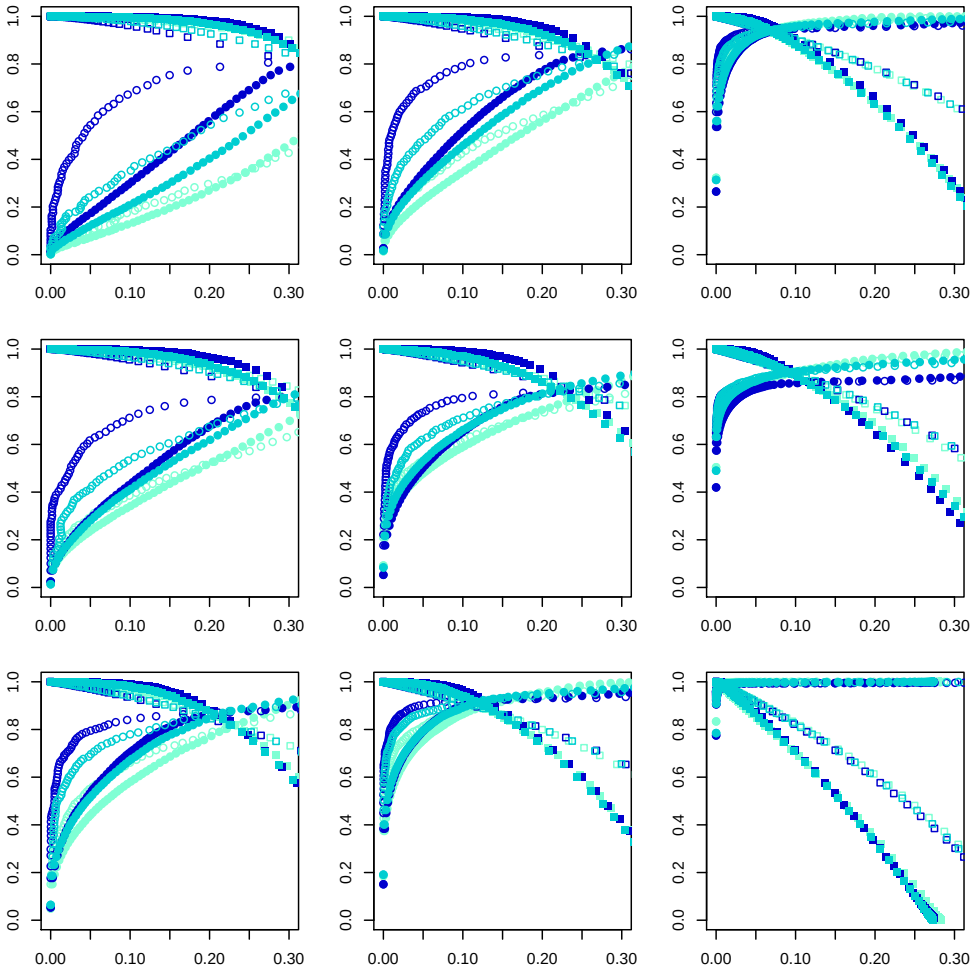


Figure A.7: *BYM mix* model, $n = 69$. In each of the nine figures it is plotted *sn*, *sp*, potential *sn*, potential *sp vs* $\widehat{FDR}$ for $S1$ (light-blue), $S2$ (dark-blue) and $S3$ (blue), see legend of figure 4.13. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $SF = 0.5$, $SF = 1$, $SF = 5$ scenarios. In such $n = 69$ scenarios the $\widehat{FDR}$ based selection rules are more sensitive in spatially correlated scenarios but they could potentially achieve more sensitivity (because of over-estimation of FDR). Non-small areas scenarios (third column) lose a lot of specificity for high FDR values; just very small FDR level are well estimated, the others being under-estimated for the "non-necessary" borrowing of strength (empirical information is not poor in big areas).

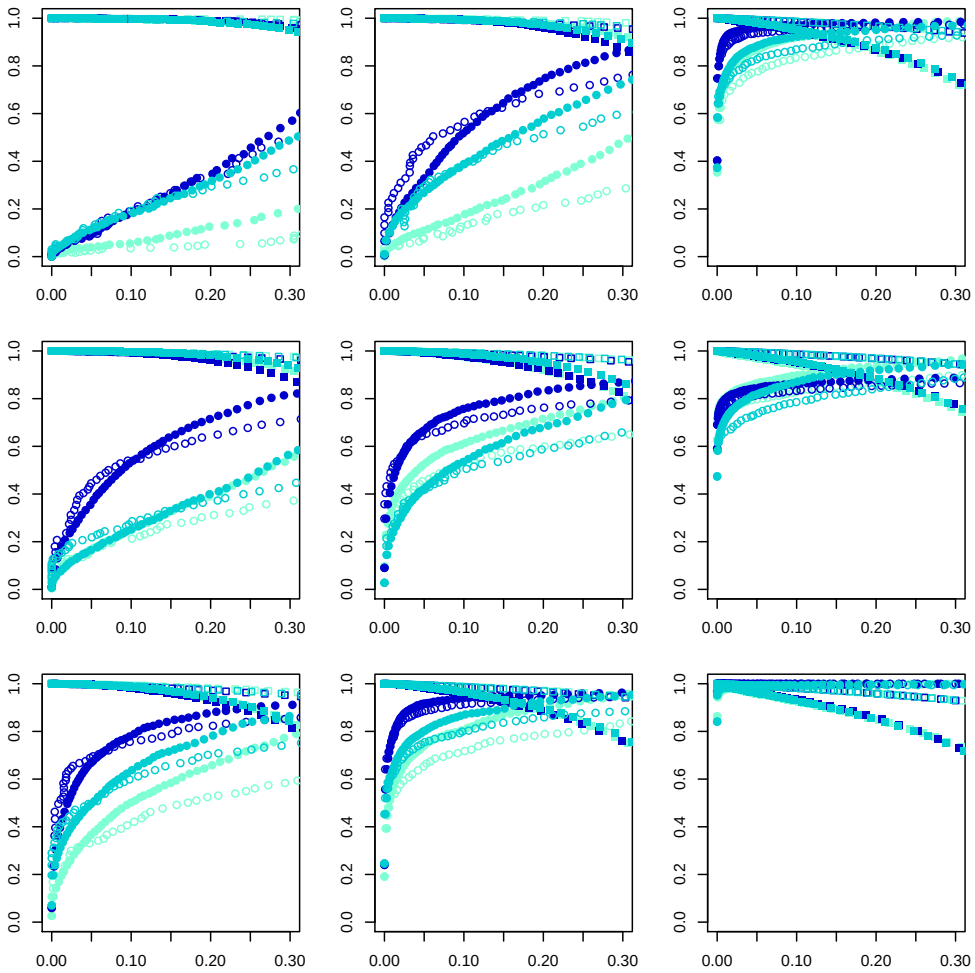


Figure A.8: *BYM mix* model, $n = 19$. In each of the nine figures it is plotted *sn*, *sp*, potential *sn*, potential *sp vs* $\widehat{FDR}$ for $S1$ (light-blue), $S2$ (dark-blue) and $S3$ (blue), see legend of figure 4.13. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $SF = 0.5$, $SF = 1$, $SF = 5$ scenarios. In such $n = 19$ scenarios the $\widehat{FDR}$ based rules tend to be more specific and apparently non less sensitive than in $n = 69$ case. More sensitivity can be achieved in spatially correlated scenarios. Sensitivity and potential sensitivity are almost equal, since the FDR is more accurately estimated in such cases (especially in $S2$ and $S3$).

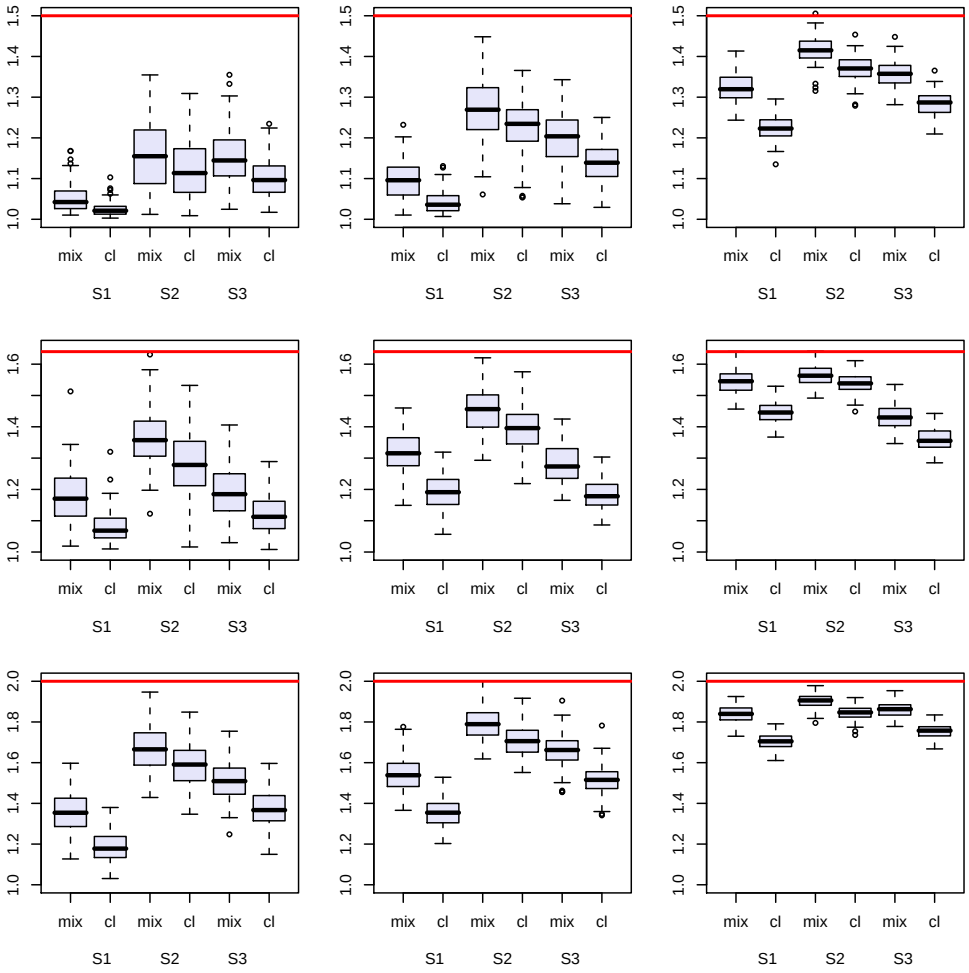


Figure A.9: *BYM mix* model, $n = 19$, all θ, all SF, all three spatial scenarios in each window. Box-plots of relative risk values in true high-risk areas (belonging to the 100 datasets) both for *BYM mix* (mix) and *BYM* (cl) models. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $SF = 0.5$, $SF = 1$, $SF = 5$ scenarios. The red line is the true risk value in such scenarios. *BYM mix* (mix) model yields a degree of over-smoothing lower than *BYM* (cl) in all scenarios.

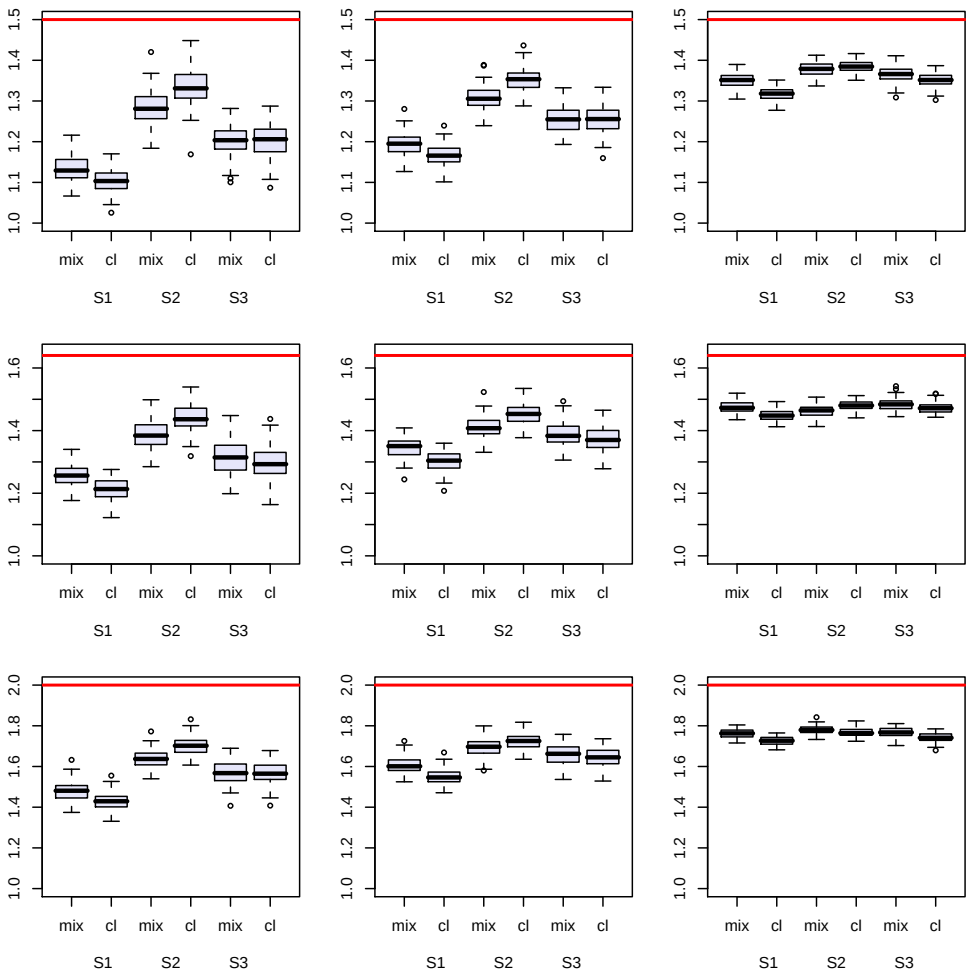


Figure A.10: *BYM mix* model, $n = 69$, all θ, all SF, all three spatial scenarios in each window. Box-plots of relative risk values in true high-risk areas (belonging to the 100 datasets) both for *BYM mix* (mix) and *BYM* (cl) models. The three rows from the top to the bottom corresponds to $\theta = 1.5$, $\theta = [1.2 \div 2]$, $\theta = 2$ values; the three columns from left to right corresponds to $SF = 0.5$, $SF = 1$, $SF = 5$ scenarios. The red line is the true risk value in such scenarios. *BYM mix* (mix) model yields a degree of over-smoothing lower than *BYM* (cl) in all scenarios except in the strongly spatially correlated risks case $S2$. Note in such $n = 69$ scenarios we reach less uncertainty in posterior relative risk estimates comparing to the $n = 19$ scenarios.

Bibliography

[1] Armitage P., (1971). Statistical methods in medical resear ch. Oxford, Blackwell.

[2] Banerjee, S., Bradley P. C., Gelfand A., (2004). *Hierarchical modeling and analysis for spatial data.*

[3] Benjamini, Y., Hochberg, Y., (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing *J. Royal. Statist. Soc. B*, **57**, 289-300.

[4] Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, **14**, 2411-2431.

[5] Berry, D., Hochberg, Y., (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, **82**, 215-227.

[6] Besag, J., York, J., Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1-21.

[7] Besag J., Kooperberg C.L., (1995). On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733-746.

[8] Besag, J., (1974). Spatial interaction and the statistical analysis of lattice system. *Journal of the Royal Statistical Society*, **36**, 192-236.

[9] Besag J., Newell J., (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A*, **154**, 327-333.

[10] Best N. G., Ickstadt K., Wolpert R. L., Briggs D. J., Combining models of health and exposure data: the SAVIAH study. In Elliott P., Wakefield J. C., Best N. G., Briggs D. J., (2000). *Spatial epidemiology: methods and applications.* Oxford University Press, 393-414.

[11] Best N., Richardson S., Thomson A., (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical methods in medical research*, **14**, 35-59.

[12] Biggeri A., Catelan D., Dreassi E., (2007). Epidemiologic surveillance and impact evaluation. *Proceedings of the 2007 intermediate conference SIS.*

[13] Broet, P., Lewin, A., Richardson, S., Dalmasso, C., Magdelenat, H., (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562-2571.

[14] Carlin, B. P., Louis, T. A., (2000). *Bayes and empirical bayes methods for data analysis*

[15] Catelan D., Biggeri A., (2008). A statistical approach to rank multiple priorities in Environmental Epidemiology: an example from high-risk areas in Sardinia, Italy. *Geospatial Health*, **3**, 81-89.

[16] Choynowski M., (1959). Maps based on probabilities. *Journal of the American Statistical Association*, **54**, 385-388.

[17] Cressie, N. A. C., (1993). *Statistics for Spatial Data.*

[18] Duncan, D.B., (1965). A Bayesian approach to multiple comparisons. *Technometrics*, **7**, 171-222.

[19] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, **96**, 1151-1160.

[20] Elliott P., Wakefield J., Best N., Briggs D., (2000). *Spatial Epidemiology: Methods and Applications*, Oxford.

[21] Farcomeni, A., (2006). More powerful control of the false discovery rate under dependence. *Statistical Methods & Applications*, **15**, 43-73.

[22] Frisén M., (2003). Statistical surveillance, optimality and methods. *International Statistical Review*, **71**, 403-434.

[23] Gelman, A., (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **2**, 1-19.

[24] Gelman, A., Price, P.N., (1999). All maps of parameter estimates are misleading. *Statistics in Medicine*, **18**, 3221-3234.

[25] Genovese, C., Wasserman, L., (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. Royal. Statist. Soc. B*, **64**, 499-518.

[26] Genovese, C., Wasserman, L., (2003). Bayesian and Frequentist Multiple Testing. *Biostatistics*, **5**, 155-176.

[27] Gòmez-Rubio, et al. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems*, **7**, 189-206.

[28] Greenland S., Robins J. M., (1991). Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemilogy*, **2**, 244-251.

[29] Haining R., Law J.,Griffith D., (2008). Modelling small area counts in the presence of over-dispersion and spatial autocorrelation. *Computational Statistics and Data Analysis*, article in press, doi:10.1016/j.csda.2008.08.014.

[30] Kelsall, J.E. and Wakefield, J.C. (1999). Discussion of "Bayesian models for spatially correlated disease and exposure data", by Best et al. *Bayesian Statistics* **6**, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), Oxford: Oxford University Press, p. 151.

[31] Kulkami, P.M., Tripathi, R.C., Michalek, J.E. (1998). Maximum (Max) and Mid-P Confidence Intervals and p Values for the Standardized Mortality and Incidence Ratios. *America Journal of Epidemiology*, **147**, 83-86.

[32] Kulldorff M., (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society A* , **164**, 61-72.

[33] Lawson A., et. al. (1999). Disease mapping and risk assessment for public health. *Wiley New York*

[34] Lawson A. et. al., (2000). Disease mapping models: an empirical evaluation. *Statistics in medicine*, **19**, 2217-2241.

[35] Mollié, A., (1996). Bayesian Mapping of Disease. *in Markov Chain Monte Carlo in Practice.*

[36] Muller, P., Parmigiani, G., Rice, K., (2006). FDR and Bayesian Multiple Comparisons Rules. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, 215-227.

[37] Newton, M., Noueriry, A., Sarkar, D., Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *in Bayesian Statistics 7*, 145-162.

[38] Pascutto C., Waakefiled J. C., Best N. G., Richardson S., B ernardinelli L., Staines A., Elliott P., (2000). Statistical issues in the analysis of disease mapping data. *Statistics in medicine*, **19**, 2493-2519.

[39] Perone Pacifico, M., Genovese, C., Verdinelli, I., Wasserman, L., (2004). False Discovery Control for Random Fields. *Journal of the American Statistical Association*, **99**, 1002-1014.

[40] Richardson, S., Thomson, A., Best, N., Elliott, P., (2004). Interpreting Posterior Relative Risk Estimates in Disease Mapping Studies. *Environmental Health Perspectives*, **112**, 1016-1024.

[41] Rolka H., Burkom H., Cooper G., Kulldorff M., Madigan D., Wong W.K., (2007). Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs, *Statistics in Medicine* , **26**, 1834-1856.

[42] Rothman, K., (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, **1**, 43-46.

[43] Rothman, K. and Greenland S., (1998). *Modern Epidemiology*, Lippincott Williams and Wilkins.

[44] Schervish, M. J., (1996). P Values: What They Are and What they Are Not. *The American Statistician*, **50**, 203-206.

[45] Schlattmann P., Böhning D., (1993). Mixture models and disease mapping. *Statistics in medicine*, **12**, 1943-1950.

[46] Scott, J, Berger J. (2003). An exploration of aspects of bayesian multiple testing. *Tech. rep. Duke university.*

[47] Shen W., Louis T. A., (1998). Triple-goal estimates in two-stages hierarchical models. *Journal of the Royal Statistical Society B*, **60**, 455-471.

[48] Storey, J.,(2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013-2035.

[49] Storey, J.,(2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society*, **64**, 479-498.

[50] Tsai S. P., wen C. P., (1986). A Review of Methodological Issues of the Standardized Mortality Ratio (SMR) in Occupational Cohort Studies. *International Journal of Epidemiology*, **15**, 8-21.

[51] Ulm K., (1990). A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *American Journal of Epidemiology*, **131**, 373-375.