Bayesian space–time analysis of health insurance data

Stefan Lang, Petra Kragler, Gerhard Haybach and Ludwig Fahrmeir
University of Munich, Ludwigstr. 33, 80539 Munich
email: lang@stat.uni-muenchen.de and andib@stat.uni-muenchen.de

Abstract

Generalized linear models (GLMs) and semiparametric extensions provide a flexible framework for analyzing the claims process in non-life insurance. Currently, most applications are still based on traditional GLMs, where covariate effects are modelled in form of a linear predictor. However, these models may already be too restrictive if nonlinear effects of metrical covariates are present. Moreover, although data are often collected within longer time periods and come from different geographical regions, effects of space and time are usually totally neglected. We provide a Bayesian semiparametric approach, which allows to simultaneously incorporate effects of space, time and further covariates within a joint model. The method is applied to analyze costs of hospital treatment and accommodation for a large data set from a German health insurance company.

Keywords: MCMC, semiparametric Bayesian inference, smoothness priors,treatment costs

1 Introduction

Actuarial applications of generalized linear models (GLMs) have gained much interest in recent years, see Renshaw (1994), and Haberman and Renshaw (1998) for a survey. In non-life insurance, they are used as a modelling tool for analyzing claim frequency and claim severity in the presence of covariates. Knowledge about these two components of the claims process is the basis for determining risk premiums. A characteristic feature of many applications is that they rely on traditional GLMs or quasi-likelihood extensions, assuming that the influence of covariates can be modelled in the usual way by a parametric linear predictor. However, as in our application to health insurance, the data provide detailed individual information for types of covariates where influence on claims is difficult or almost impossible to assess with parametric models. Firstly, the effect of metrical covariates, such as age of the policy holder, is often of unknown nonlinear form. Generalized additive models (GAMs) with a semiparametric additive predictor provide a flexible framework for statistical modelling in this case. Secondly, the data also include information on the calendar time of claims, and on the district where the policy holder lives. Neglecting these effects in modelling the claims process will lead to biased fits, with corresponding consequences for risk premium calculation, see Brockman and Wright (1992) for a discussion in the context of calendar time.

Therefore, statistical modelling tools are required which make thorough space-time analyses of insurance regression data possible and allow to explore temporal and spatial effects simultaneously with the impact of other covariates. We present a semiparametric Bayesian approach for a unified treatment of such effects within a joint model, developed in the context of generalized additive mixed models in Fahrmeir and Lang (2001a, b) and Lang and Brezger (2001). Our application investigates costs caused by treatment and accommodation in hospitals. However, the basic concepts are transferable to other costs for medical treatment, to claim frequencies and to other non-life insurances.

2 Semiparametric Bayesian inference for space-time regression data

2.1 Data

The space-time regression data from health insurance, which will be analyzed in the next section, consist of individual observations $(y_{it}, x_{it}, w_{it}, s_{it})$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, where y_{it} are costs for hospital treatment or for accommodation of policy holder i in month t, x_{it} is the age at calendar time t, w_{it} is a vector of categorical covariates such as gender, occupation group, type of disease, and s_{it} is the district in West Germany where the insured lives in month t. In general, other types of response variables y, in particular claim frequency, might be of primary interest, and x could be a vector of several metrical covariates.

2.2 Observation model

Since costs y_{it} are nonnegative, several distributional assumptions can be reasonable, see for example Mack (1998). We do not take into account zero-costs, so a Gamma or log-normal distribution is a common choice. While the former is often preferred in car insurance, a log-normal distribution gives a better fit to the health insurance data at hand. Therefore, we consider log-costs $z_{it} = \log(y_{it})$, and choose a Gaussian additive model $z_{it} = \eta_{it} + \epsilon_{it}$, with i.i.d. errors $\epsilon_{it} \sim N(0, \sigma^2)$, and predictor

$$\eta_{it} = f(x_{it}) + f_{time}(t) + f_{spat}(s_{it}) + w_{it}'\gamma, \quad i = 1, \ldots, n, \quad t \in T_i, \tag{1}$$

where $T_i \subset \{1, \ldots, T\}$ are the months with nonzero-costs $y_{it} > 0$. The unknown function $f(x)$ is the nonlinear effect of age x, $f_{time}(t)$ represents the calendar time trend, and $f_{spat}(s)$ is the effect of district $s \in \{1, \ldots, S\}$ in West Germany. We further split up this spatial effect into the sum

$$f_{spat}(s) = f_{struct}(s) + f_{unstr}(s)$$

of structured (spatially correlated) and unstructured (uncorrelated) effects. A rational for this decomposition is that a spatial effect is usually a surrogate of many underlying unobserved influential factors. Some of them may obey a strong spatial structure, others may be present only locally.

The last term in (1) is the usual linear part of the predictor, with fixed effects. To ensure identifiability, an intercept is always included into w_{it}, and the unknown functions are centered about zero.

Retransformation of the Gaussian additive model (1) for log-costs z_{it} gives a lognormal model for costs y_{it} with (conditional) expectation

$$E(y_{it}|\eta_{it}, \sigma^2) = \mu_{it} = \exp(\eta_{it} + \sigma^2/2), \tag{2}$$

i.e., we get a multiplicative model for expected costs.

Model (2) is closely related to a Gamma model for y_{it} with predictor (1) and an exponential link function. The models are special cases of generalized additive mixed models described in Fahrmeir and Lang (2001a).

2.3 Priors for functions and parameters

To formulate priors in compact and unified notation, we express the predictor vector $\eta = (\eta_{it})$ in matrix notation by

$$\eta = f + f_{time} + f_{struct} + f_{unstr} + W\gamma, \tag{3}$$

where f, $f(time)$ etc. are the vectors of corresponding function values and $W = (w_{it})$ is the design matrix for fixed effects. It turns out that each function vector can always be expressed as the product of a design matrix and a (high-dimensional) parameter vector. Using $f = X\beta$ as a generic notation for functions, (3) becomes

$$\eta = \cdots + X\beta + \cdots + W\gamma.$$

For fixed effects γ, we generally choose a diffuse prior, but a (weakly) informative normal prior is also possible. Constructions of the design matrix X and priors for β depend upon the type of the function and on the degree of smoothness. For metrical covariates, such as age and calendar time, random walk models, P-Splines and smoothing splines are suitable choices, structured spatial effects are modelled through Markov random field priors, and unstructured effects through i.i.d. normal random effects. In any case, priors for the vectors β have the same general Gaussian form

$$p(\beta|\tau^2) \propto \exp(-\frac{1}{2\tau^2}\beta' K \beta). \tag{4}$$

The penalty matrix K penalizes roughness of the function. Its structure depends on the type of covariate and on smoothness of the function, see Fahrmeir and Lang (2001a, b) and Lang and Brezger (2001) for details. The hyperparameter τ^2 acts as a smoothing parameter and controls the degree of smoothness. A highly dispersed inverse Gamma $IG(a, b)$ prior is a convenenient choice as a hyperprior. The same choice is made for the variance σ^2 of the errors ϵ_{it}. As usual, observations and priors are assumed to be conditionally independent.

2.4 MCMC inference

Estimation of functions and parameters is based on the posterior, which is defined by the observation model and the priors. Since the posterior is intractable analytically or numerically, inference is carried out via MCMC simulation. For the

Gaussian additive model (1) for log-costs, full conditionals are (high-dimensional) Gaussian or inverse Gamma distributions, so that Gibbs sampling is possible. The full conditional for a typical β is Gaussian with precision matrix P and mean m

$$P = \frac{1}{\sigma^2} X'X + \frac{1}{\tau^2} K, \quad m = P^{-1} \frac{1}{\sigma^2} X'(y - \tilde{\eta}),$$

where $\tilde{\eta}$ is the part of the predictor assossiated with the remaining effects. Efficient sampling can be achieved by Cholesky decompositions for band matrices (Rue, 2000) and is implemented in BayesX (Lang and Brezger, 2000). For non-Gaussian observation models, e.g., a Gamma model, additional MH steps are necessary.

3 Application to Health Insurance Data

The approach has been applied to a large space-time regression data set from a private health insurance company in Kragler (2000), with separate analyses for various types of health services. The data set contains individual observations for a sample of 13.000 males (with about 160.000 observations) and 1.200 females (with about 130.000 observations) in West Germany for the years 1991-1997. All analyses were carried out separately for males and females. Analyses for costs were based on Gamma or log-normal models, while frequencies of doctoral visits or of treatments in hospitals were modelled by logit regressions.

Supported by evidence from diagnostic model checks, we use Gaussian additive models (1) for the following space-time analyses of costs for health services in hospitals. In contrast to costs for doctoral visits, it turns out that the categorical covariates "occupation group" and "type of disease" are non-significant. Furthermore, separate analyses for the 3 types of health services (accommodation, treatment with operation, treatment without operation) are preferred to a joint model with type of service as a categorical covariate. Therefore, our analysis for the 6 subgroups, determined by the combinations of gender and type of service, uses a Gaussian additive model (1) for log-costs, where w_{it} contains only an intercept. The effect of age and the time trend are modelled by Bayesian P-splines (Lang and Brezger 2001), for the spatial effect a Markov random field prior with adjacency weights is used (Fahrmeir and Lang 2001b).

The effect of age is displayed in Figure 1 for each of the 6 groups. It differs between groups, showing that separate analyses are necessary to avoid confounding. For females, the effect on costs for accommodation increases monotonically over a wide range of age values. In contrast, the effects for treatment with or without operation have a different shape. Starting from a higher level in younger years, they decrease monotonically until about 45 years of age, where the effects starts to increase. A possible explanation might be the higher proportion of younger females staying in hospitals for births of their children: Mostly, they stay only for a few days with comparably low costs for accommodation, but with relatively higher costs for medical treatment. Figure 2 shows the effects of calendar time. While the effect on accommodation costs is more or less continuously increasing over the years, a corresponding increase of the effect on treatment costs until about 1995 is followed by an enormous decline. The reason for this decline might be changes in regulations

or laws for health insurance or health care. More discussion with experts is needed for a convincing explanation of this effect. Anyway, it becomes obvious that risk premium calculation based on data from this period has serious problems, when the calendar time trend is simply neglected. This is one of the main reasons, why we believe that careful space-time analyses of insurance data are needed, at least for monitoring purposes. The argument is confirmed by the results for regional effects. They are visualized in Figure 3 by "significance maps", constructed as follows: For each region 10% and 90% posterior quantiles of its estimated effect are calculated from the posterior. If the 10% quantile is positive, the regional effect is significantly positive; it is significantly negative if the 90% quantile is negative, and it is non-significant otherwise. Again, the maps reveal distinct patterns which motivate closer inspection by experts.

4 Conclusion

Our application demonstrates that a thorough space-time analysis of insurance data can reveal important features of the claims process which are not easily detected by traditional methods. Although we focussed on claim severity in health insurance, the concepts can also be used for modelling claim frequencies and for analyzing other non-life insurance data. First experience with modelling claim frequencies in health insurance shows that a direct transfer of models common in car insurance is at least problematic. We will investigate this in detail in future work.

References

Brockmann, M. and S. Wright (1992). Statistical Motor Rating: Making Effective Use of Your Data. *Journal of the Institute of Actuaries 119*, 457–543.

Fahrmeir, L. and S. Lang (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Appl. Statist. (JRSS C) (to appear)*.

Fahrmeir, L. and S. Lang (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Ann. Inst. Statist. Math. 53*, 11–30.

Haberman, S. and A. Renshaw (1998). Actuarial Applications of Generalized Linear Models. In D. Hand and S. Jacka (Eds.), *Statistics in Finance*. Arnold, London.

Kragler, P. (2000). Statistische Analyse von Schadensfällen privater Krankenversicherungen. Master's thesis, University of Munich.

Lang, S. and A. Brezger (2000). BayesX–Software for Bayesian Inference based on Markov Chain Monte Carlo Simulation Techniques. SFB 386 Discussion Paper 187, University of Munich.

Lang, S. and A. Brezger (2001). Bayesian P-splines. SFB 386 Discussion Paper 236, University of Munich.

Mack, T. (1998). *Schadensversicherungsmathematik*, Volume 28 of *Schriftenreihe Angewandte Versicherungsmathematik*. Verlag Versicherungswirtschaft, Karlsruhe.

Renshaw, A. (1994). Modelling the Claims Process in the Presence of Covariates. *Astin Bulletin 24*, 265–285.

Rue, H. (2000). Fast Sampling of Gaussian Markov Random Fields with Applications. Technical report, University of Trondheim, Norway.

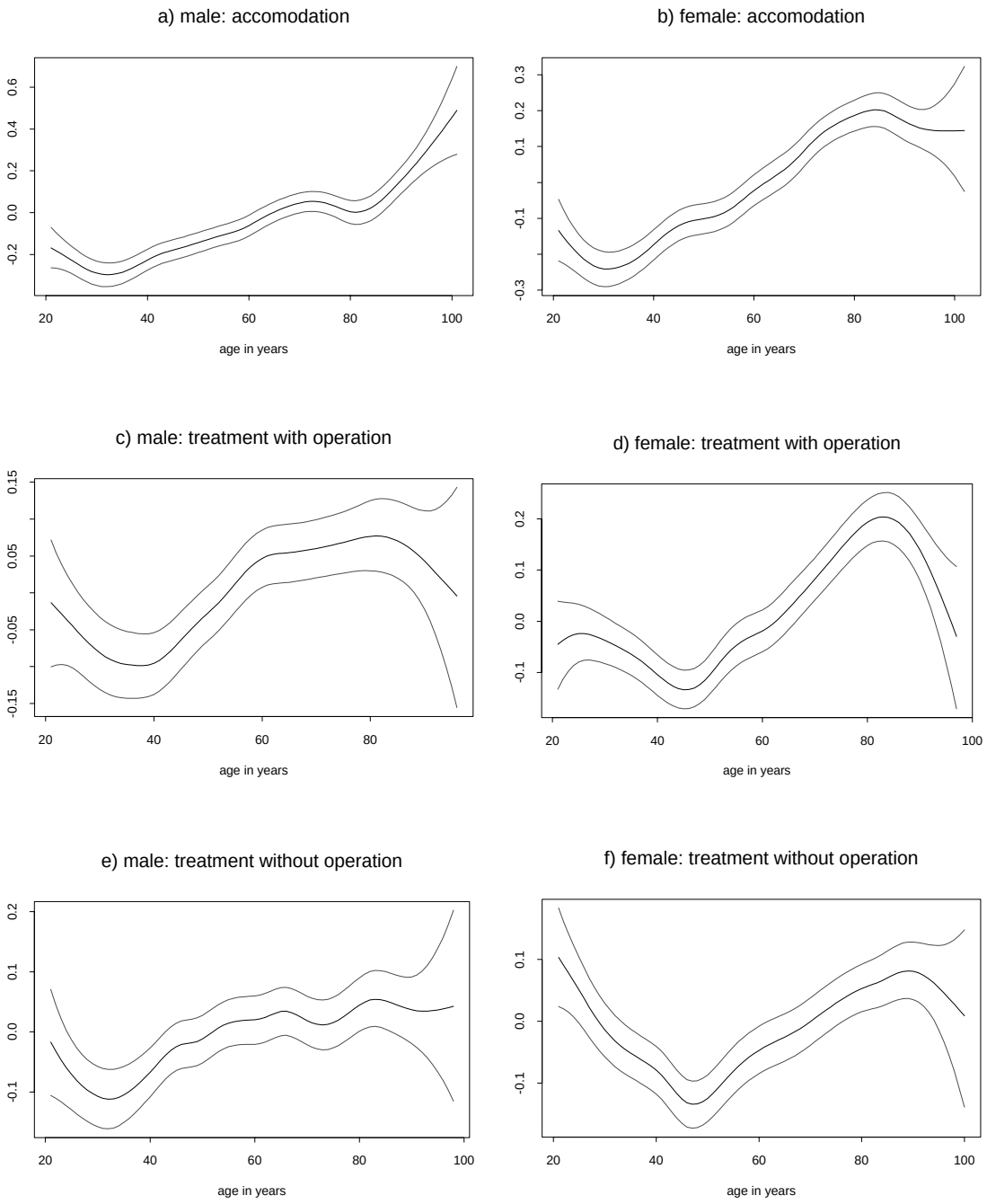

Figure 1: *Estimated effect of age. Shown is the posterior mean within 80 % credible regions.*

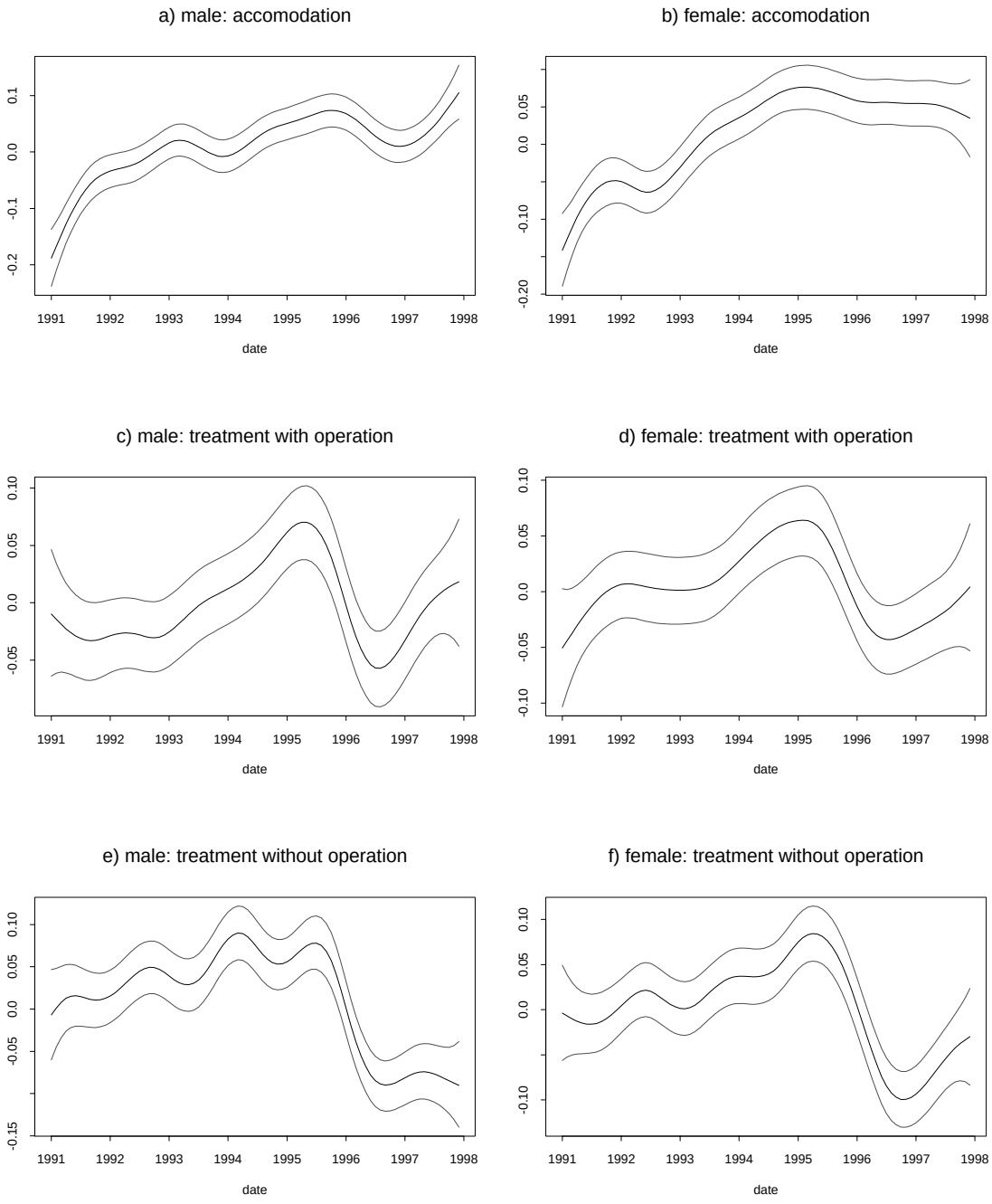

Figure 2: *Estimated time trend. Shown is the posterior mean within 80 % credible regions.*

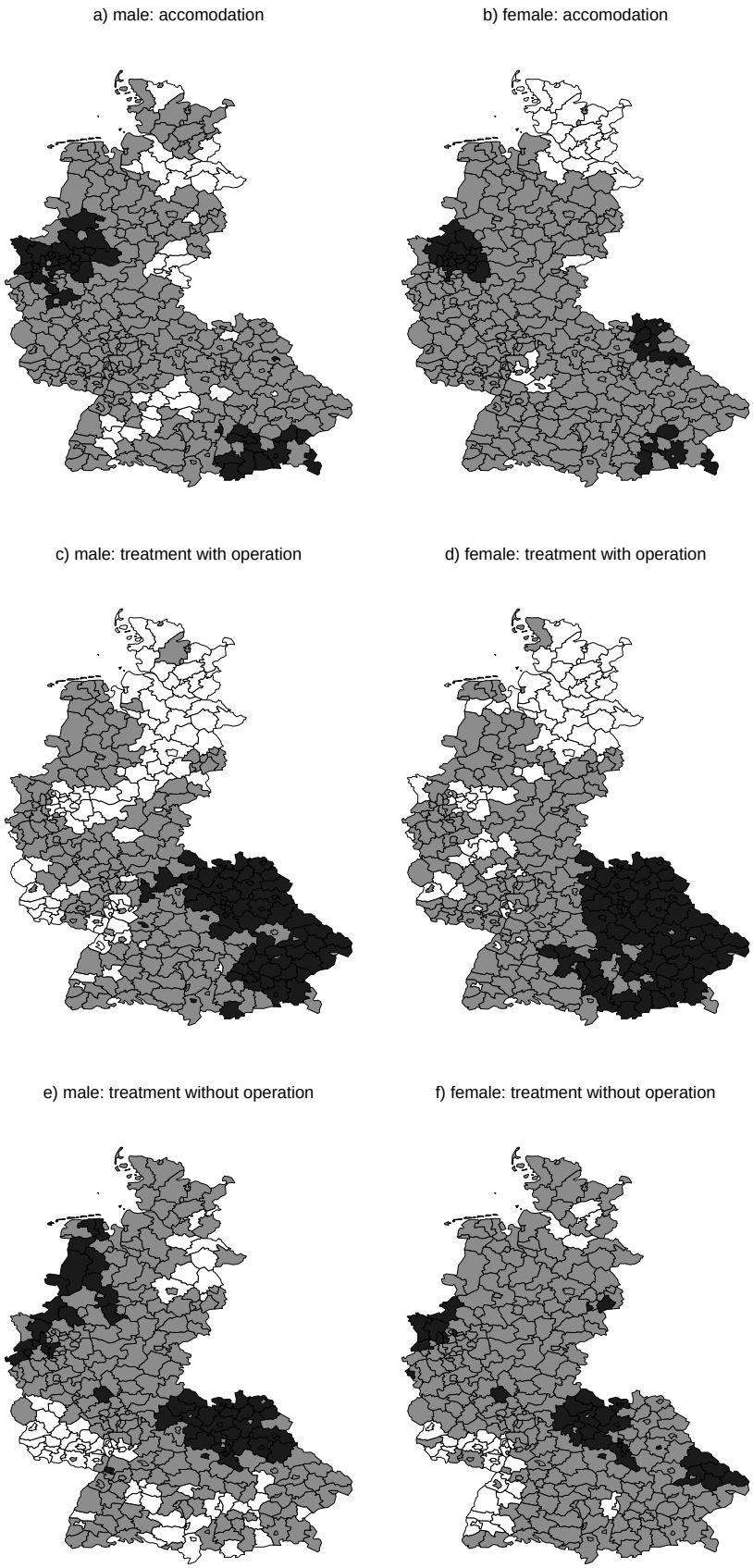

Figure 3: *Posterior probabilities of the structured spatial effect.*