Dipartimento di Statistica
"Giuseppe Parenti"

W O R K I N G P A P E R 2 0 1 0 / 1 2

Shared component models in joint disease mapping: a comparison via a simulation experiment

Emanuela Dreassi

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Shared components models in joint disease mapping: a comparison via a simulation experiment

Emanuela Dreassi

December 9, 2010

Abstract

Two models for jointly analysing the spatial variation of incidences of three (or more) diseases, with common and uncommon risk factors, are compared via a simulation experiment. In both models, the linear predictor can be decomposed into shared and disease-specific spatial variability components (named shared clustering and specific clustering respectively). The two models are the shared model on the original formulation that use exchangeable Poisson distribution as response multivariate variable and shared components model that use a Multinomial one. The simulation study shows that models behave similarly. However, Multinomial shared components model performs better for disease-specific spatial variability clustering terms but it is lower for the shared one.

1 Introduction

A great amount of the literature deals with disease mapping, as the statistical analysis of geographical patterns of disease. Any spatial variation may be explained by different risk factors, therefore disease mapping allows to state hypotheses concerning their aetiology. Interest in joint disease mapping increased over recent years: joint statistical modelling of several diseases on the same spatial location, with different and common aetiologies. Joint analysis highlights common and uncommon geographical patterns of risk and obtains more precise and convincing results.

Various attempts to consider simultaneously more than one disease have been made: by a multilevel model as Langford et al. (1999) and Leyland et al. (2000), or by an ecological regression approach where a disease represent a covariate of the model as Bernardinelli et al. (1997). However, the joint modelling approach seems to be more naive, as all diseases enter as response variables with reference to unobserved latent risk factors. More recently, joint modelling following a Multivariate Gaussian Markov random field has been proposed; see Gelfan and Vounatsou (2003) and Jin et al. (2005).

In this paper, we focus on a particular class of models: shared component models. Originally introduced by Knorr-Held and Best (2001), these models have been extended to more than two diseases by Held et al. (2005) and from exchangeable Poisson response to a Multinomial one by Dreassi (2007).

In Dreassi (2007) a Multinomial model (PL) is presented and compared with exchangeable Poisson model (SC) by a real example. In this paper, a simulation study is conducted to evaluate and compare more deeply the performances of both models.

The paper is organized as follows. Sect. 2 introduces the joint analysis with shared components model following exchangeable Poisson model (SC) and Multinomial models (PL). Sect. 3 describe the simulation experiment. Results are showed in Sect. 4 and conclusion in Sect. 5.

2 Shared components models

Shared components models highlight common and specific spatial components, allowing the linear predictor to be decomposed into shared and disease-specific spatial variability terms.

2.1 Shared components Poisson model

Let y_{ik} denote the number of death cases for k-th disease ($k = 1, \ldots, K$) and i-th area ($i = 1, \ldots, I$). Each y_{ik} is assumed to follow a Poisson distribution with parameters $E_{ik}\theta_{ik}$, where E_{ik} represent the expected cases in i-th area and k-th disease and θ_{ik} the relative risk. Following the standard model of Besag et al. (1991) on consider a log link for θ_{ik}

$$\log(\theta_{ik}) = \alpha_k + u_{ik} + v_{ik} \tag{1}$$

where α_k represents a cause-specific intercept, such as an overall risk level, u_{ik} is a spatially structured term, and v_{ik} a spatially unstructured term.

The prior distribution for the model parameters is as follows. The intercept α_k has a flat non-informative distribution. The heterogeneity terms v_{ik} are independent, each v_{ik} being Normal $(0, \lambda_{vk}^{-1})$ (λ_{vk} represents the precision parameter). Using Gaussian Markov random fields (GMRFs) models in order to cope the spatial structure, the clustering terms u_{ik} are modeled conditionally on $u_{l\sim ik}$ terms ($\sim i$ indicates adjacent areas to i-th ones, $l = 1, \ldots, I$ and n_i their number; where adjacent means that two areas share an edge or, for islands, that exists a boat connection), as Normal $(\bar{u_{ik}}, (\lambda_{uk} n_i)^{-1})$ where $\bar{u_{ik}} = \sum_{l\sim i} \frac{u_{lk}}{n_i}$.

The hyperprior distributions of the precision parameters λ_{vk} and λ_{uk} are assumed to be Gamma $(0.5, 0.0005)$ as suggested by Kensall and Wakefield (1999).

Following Knorr-Held and Best (2001) and Held et al. (2005), a model on the shared components formulation is considered: the structured spatial terms (clustering) u_{ik} in 1 are decomposed into a shared and a disease-specific effect. So, for example, when $K = 3$ each disease's clustering term

could be

$$\begin{aligned} u_{i1} = &\quad us1_i \times \omega_1 + us2_i \times \delta_1 + up_{i1} \\ u_{i2} = &\quad us1_i \times \omega_2 + us2_i \times \delta_2 + up_{i2} \\ u_{i3} = &\quad us1_i \times \omega_3 \end{aligned} \tag{2}$$

where $us1_i$ and $us2_i$ represent the shared clustering components (the know risk factors pattern) and up_{i1} and up_{i2} the specific ones. The scale parameters $\omega_1, \ldots, \omega_3$ and δ_1, δ_2 allow the shared components to vary per cause by a constant factor.

Terms $\log \omega_1, \ldots, \log \omega_3$ and $\log \delta_1, \log \delta_2$, constrained to $\sum_{k=1}^{3} \log \omega_k = 0$ and $\sum_{k=1}^{2} \log \delta_k = 0$, are assumed to be multivariate normal distributed with zero mean and variance covariance matrix respectively

$$\Sigma_\omega = \sigma_\omega^2 \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \tag{3}$$

$$\Sigma_\delta = \sigma_\delta^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{4}$$

Knorr-Held and Best (2001) consider $\sigma_\omega^2 = \sigma_\delta^2 = 0.17$. The $us1_i$, $u12_i$, up_{i1} and up_{i2} terms are modelled following a GMRF as described before.

2.2 Shared components Multinomial model

Dabney and Wakefield (2005) remade a proportional mortality model (see Breslow and Day (1987)) to the joint mapping of two diseases. Originally, this model was used when the population at risk is unknown. Instead of adopting a Poisson model with expected cases as offset, the model makes use of a simultaneous estimation of age and spatial effects that should be preferred to the Poisson one, since it includes variability in the age estimates. Proportionality is assumed in the model, so that sums over strata population are allowed; only a single parameter per confounder (i.e. age, sex, race) for each area is considered. Inference on the differences in log relative risks can be made without knowledge of the population counts of those at risk.

In Dreassi (2007), Following suggestions to highlight similarity and dissimilarity on spatial patterns by proportional mortality model and shared component model, an other shared component model is introduced: a Multinomial (or polytomous logit) (PL) model. In this model, a disease is regarded as reference category, and for each predictor on adopt the shared components model formulae. In the model on assume proportionality and then on consider the model for death for each disease, area and age-stratum without knowledge of the population at risk; the latter may be unknown or subject to data anomalies as migration or census undercount.

Let $y_{ij} = (y_{ij1}, \ldots, y_{ijk}, \ldots, y_{ijK})\prime$ be distributed according to a multinomial with parameters m_{ij} and probability vector $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijk}, \ldots, \pi_{ijK})\prime$, where $m_{ij} = \sum_{k=1}^{K} y_{ijk}$ and $\sum_{k=1}^{K} \pi_{ijk} = 1$. A polytomous logit model is

considered: each category probability is modeled as

$$\pi_{ijk} = \phi_{ijk} / \sum_{r=1}^{K} \phi_{ijr} \tag{5}$$

where each log odd

$$\log(\phi_{ijk}) = \alpha_k^{\diamond} + a_{jk}^{\diamond} + u_{ik}^{\diamond} + v_{ik}^{\diamond} \tag{6}$$

is decomposed additively into a disease-specific intercept $\alpha_k^{\diamond}$ (representing overall difference between k-th disease and K-th reference disease), $a_{jk}^{\diamond}$ a time-structured term by age and disease representing difference between k-th disease and reference category, and structured $u_{ik}^{\diamond}$ and unstructured $v_{ik}^{\diamond}$ spatial effects (again representing difference on the spatial structured and unstructured spatial terms between the disease k considered and the reference disease).

For the $a_{jk}^{\diamond}$ term a first order random walk with independent Gaussian increments is assumed (see Clayton (1996)). For the other terms prior are equal to SC model. This is a time-structured term by age that represents the additive effect of the j-th age for each disease on the log odds; conditionally to the adjacent (on time scale) terms $a_{jk} \sim$ Normal $(\bar{a_{jk}}, (\lambda_{ak} n_j)^{-1})$ where $\bar{a_{jk}}$ is the mean of the $(j-1)$-th and $(j+1)$-th terms and $n_j = 2$; for the extreme age classes $j = 1, 13$, $n_j = 1$ and $\bar{a_{jk}}$ is the $(j+1)$-th or $(j-1)$-th term. Hyperprior for precision parameter λ_{ak} are again assumed to be Gamma $(0.5, 0.0005)$ as suggested by Kensall and Wakefield (1999)

Representing, for example, the third disease the reference category (when $K = 3$), $\alpha_3^{\diamond} = 0$, $a_{j3}^{\diamond} = 0$ (for each age-class $j = 1, \ldots, 13$), $u_{i3}^{\diamond} = 0$ and $v_{i3}^{\diamond}$ (for each area $i = 1, \ldots, I$) has defined, as constraint for identifiability.

Note that terms $u_{i1}^{\diamond}$ and $u_{i2}^{\diamond}$ represent differences between first disease and reference category disease clustering, and between third disease and reference category disease clustering, respectively,

$$u_{i1}^{\diamond} = u_{i1} - u_{i3} \qquad \text{and} \qquad u_{i2}^{\diamond} = u_{i2} - u_{i3} \tag{7}$$

We consider a model where the difference structured spatial terms (clustering) in equation (6) are decomposed into a shared and a disease-specific effect (Held et al. (2005)). We can represent each clustering term for the first and second disease, respectively, as

$$u_{i1}^{\diamond} = us2_i \times \delta_1 + up_{i1} \qquad \text{and} \qquad u_{i2}^{\diamond} = us2_i \times \delta_2 + up_{i2} \tag{8}$$

where $us2_i$ is the shared clustering component and up_{i1} and up_{i2} is the disease specific one; both are distributed according GMRF models. Prior distributions are the same described before for SC model. Note that equation (7) and equation (8) imply

$$u_{i1} = u_{i3} + ua_i \times \delta_1 + up_{i1} \qquad \text{and} \qquad u_{i2} = u_{i3} + ua_i \times \delta_2 + up_{i2} \tag{9}$$

which is different from equation (2) because we are forcing to be $\omega_1 = \omega_2 = \omega_3$. Nevertheless, since these terms are constant on the space,

spatial patterns for us_i and difference between up_{i1} and up_{i2} are still informative. Interest is focused on the estimate of disease-specific spatially structured effects up_{i1} and up_{i2} because these are considered as latent variables denoting disease-specific risk factors.

3 Simulation study

To evaluate the proposed shared components models, considering exchangeable Poisson (SC) or Multinomial (PL) models, we conducted a simulation study.

The shared components models used for the simulation experiment has been conceived with reference to a specific application: three diseases, a common risk factor and another risk factor shared by two diseases only. In the present application, the incidence of the disease that shared only one risk factor represents the reference category for the Multinomial model (PL). Then a shared component, representing the second risk factor common only for the two diseases adjusted for the first risk factor, is considered. Finally, including disease specific terms in the predictors, the possibility of other different risk factors is investigated.

We used three different disease maps (each map with n=225 areas) taken square areas over a 15×15 grid. For each i-th area $i = 1, \ldots, 225$, and for each k disease $k = 1, 2, 3$, we generated 100 deaths counts from Poisson (100 θ^0_{ik}). We assumed that each $\log \theta^0_{ik}$ for $k = 1, 2, 3$ is equal to

$$\begin{aligned} \log \theta^0_{i1} = &\quad us1^0_i + us2^0_i + up1^0_i \\ \log \theta^0_{i2} = &\quad us1^0_i + us2^0_i + up2^0_i \\ \log \theta^0_{i3} = &\quad us1^0_i \end{aligned} \tag{10}$$

and a range of $us1^0$ from -0.20 to 0.22, of $us2^0$, $up1^0$ and $up2^0$ from 0 to 0.15. We fix ω_1, ω_2, ω_3, δ_1 and δ_2 equal to 1 and α and heterogeneity u_{ik} equal to zero.

Figure 1 shows the map of each clustering terms, shared $us1^0$ and $us2^0$ and specific $up1^0$ and $up2^0$ respectively; Figure 2 the three disease true map θ^0_k.

We estimate $us1$ $us2$ $up1$ $up2$ using (SC) shared poisson model and $us2$ $up1$ and $up2$ using (PL) multinomial model.

The marginal posterior distributions of the parameters of interest for both models are approximated by Monte Carlo Markov Chain methods.

The estimates for SC model are obtained using specific MCMC software. It uses joint updates of the latent spatial fields and it is able to incorporate sum to zero constraints in spatial fields explicitly in the prior and in the MCMC algorithm.

For PL model we used WinBUGS software (Spiegelhalter et al (2004)) in order to perform the MCMC analysis. The convergence of the algorithm has been evaluated using the test proposed by Gelman and Rubin (1992) for multiple chains for a subset of identifiable parameters (precision hyperparameters) for some simulation iteration. The algorithm seems to converge after a few thousand iterations. However, given also the very

Table 1: Average MSE and average variance from shared component Poisson model (SC) and shared components Polytomous model (PL) for shared components and specific components

clustering	AMSE (SC)	AMSE (PL)	AVAR (SC)	AVAR (PL)
$us1$	0.001559		0.000475	
$us2$	0.001555	0.001569	0.000121	0.000127
$up1$	0.001278	0.001226	0.000212	0.000185
$up2$	0.001277	0.001240	0.000223	0.000192

high number of (non monitored) parameters in the model, we decided to discard the first 200,000 iterations (burn-in) and to store for estimation 2,000 samples (one each 100) of the following 200,000 iterations.

The estimates obtained through the two models are compared using the average mean square error and the average variance, respectively

$$\text{AMSE} = \sum_{i=1}^{225}\sum_{j=1}^{100}\frac{(uo_{ij}^{*} - uo_{i}^{0})^2}{225\ \ 100} \quad \text{and} \quad \text{AVAR} = \sum_{i=1}^{225}\sum_{j=1}^{100}\frac{(uo_{ij}^{*} - \bar{uo}_{i}^{*})^2}{224\ \ 100} \tag{11}$$

where uo_{ij}^{*} denotes the estimates under model $*$ (SC or PL) for i-th area and j-th simulated data for the generic clustering term uo ($us1$, $us2$, $up1$ and $up2$); uo_{i}^{0} represent the true value for the generic clustering term and $\bar{uo}_{i}^{*}$ the average over all the simulations for uo_{i}^{0}.

4 Results

Figure 3 and Figure 4 show the average over the 100 simulated data of clustering terms estimates by the two different models. Each average map is on a 15×15 grid, with the same levels of gray (from -0.177 to 0.168).

Results about AMSE and AVAR are reported on Table 1. They suggest a similar behavior of the models; however PL performs better than SC for specific clustering terms while it is lower for shared clustering term. Results are consistent with previous analysis (Dreassi, 2007) on a real example: specific clustering estimated terms with PL model have smaller standard deviation respect to SC estimates.

Figure 5 and Figure 6 show the relative error maps. Denoting with $\text{abs}(\cdot)$ the absolute value, the relative error for each i-th area ($i = 1, \ldots, 225$) is defined as

$$\text{abs}\frac{\bar{uo}_{i}^{*} - uo_{i}^{0}}{uo_{i}^{0}}, \tag{12}$$

where $\bar{uo}_{i}^{*}$ is the average, over the 100 simulated data, of the estimates for a given model ($*$ states for SC o PL), and uo_{i}^{0} is the true clustering parameter (uo states for $us1$, $us2$, $up1$ and $up2$).

5 Conclusion

As stated in Dreassi (2007), the SC model for joint disease mapping is perhaps more 'natural' and 'elastic' than PL model: both risk factors are considered as shared components of the model, and both common clustering terms are allowed to vary per cause for a multiplicative constant factor. In turn, the PL model gives some advantages: it allows to analyse mortality data without knowing the population at risk and to consider variability on age effect estimates in the model. Using a particular disease as reference category, we can omit a GMRF for the shared terms common to all the diseases; using a Multinomial model instead than exchangeable Poisson for a multivariate problem seem to be more convenient.

Advantages and disadvantages for each model have been disregarded using an unrealistic, but particular simulation experiment. Accordingly, results from simulation give us information about the performances on estimating clustering terms.

The simulation study suggests that both models provide similar estimates. However, PL model behaves better for specific clustering terms, once hypotheses (even if strong, unfortunately) of this model are accomplished.

Acknowledgement

The author would like to express her gratitude to Håvard Rue, for making available the software for SC model.

References

Bernardinelli L., Pascutto C., Best N.G., Gilks W.R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, 16, 741-752.

Besag J., York J., Mollié A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Breslow N., Day N.E. (1987). The analysis of cohorts studies, in *Statistical Methods in cancer research, volume 2.*. Scientific publications n. 82, Lyon: International Agency for Research on Cancer.

Clayton D. (1996). Generalized linear mixed models, in Gilks WR, Richardson S, Spiegelhalter DJ (Eds) *Markow Chain Monte Carlo in practice*. London: Chapman & Hall, 275-301.

Dabney A.R., Wakefield J.C. (2005). Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, 14, 83-112.

Dreassi E. (2007). Polytomous Disease Mapping to detect uncommon risk factors for related diseases. *Biometrical Journal*, 49 (4), 520-529.

Gelfand A., Vounatsou P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4, 11-25.

Gelman A., Rubin D.R. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Held L., Natário I., Fenton S.E., Rue H., Becker N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research*, 14, 61-82.

Jin X., Carlin B.P., Banerjee S. (2005). Generalized Hierarchical Multivariate CAR Models for Areal Data. *Biometrics*, 61 (4), 950-961.

Kelsall J.E., Wakefield J.C. (1999). Discussion of "Bayesian Models for Spatially Correlated Disease and Exposure Data", by Best *et al.*, in *Bayesian Statistics 6*, Bernardo *et al.* (eds.), Ney York Oxford University Press, 151.

Knorr-Held L., Best N. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 164, 73-86.

Langford I.H., Leyland A.H., Rasbash J., Goldstein H. (1999). Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistics C - Applied Statistics*, 48, 253-268.

Leyland A.H., Langford I.H., Rasbash J., Goldstein H. (2000). Multivariate spatial models for event data. *Statistics in Medicine*, 19 (17-18), 2469-2478.

Spiegelhalter D.J., Thomas A., Best N., Lunn D. (2004). *WinBUGS User Manual, Version 1.4.1.*

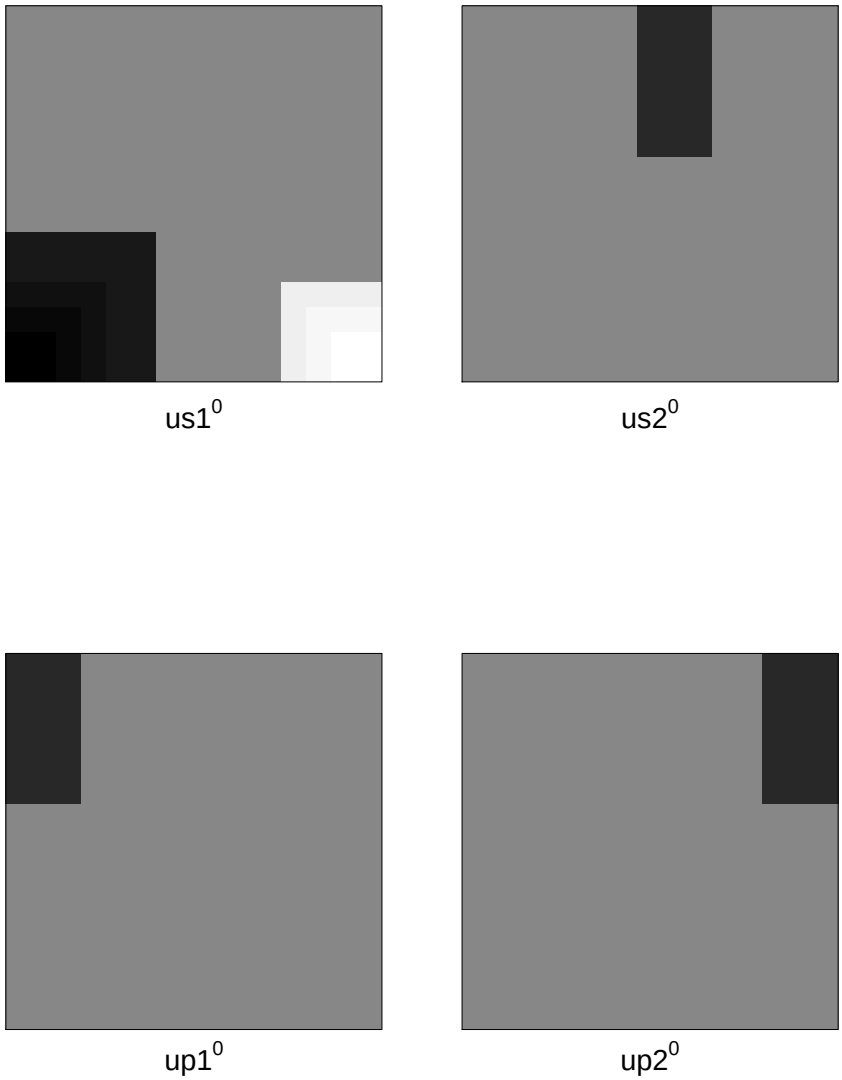
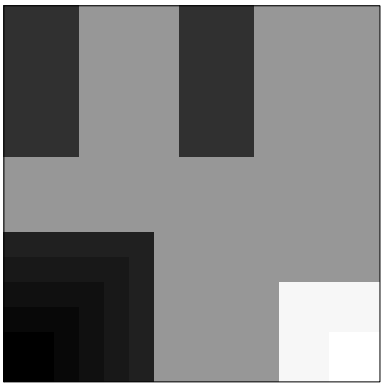
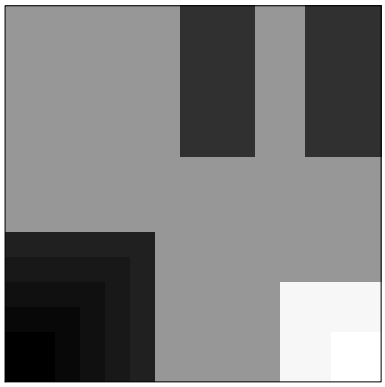


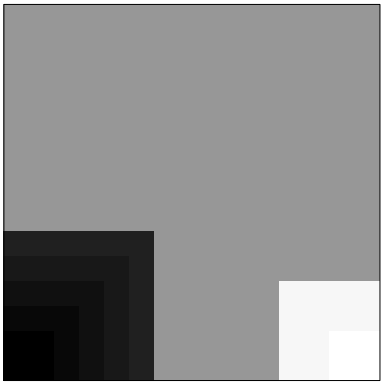
Figure 1: True clustering terms



$\theta^0 1$



$\theta^0 2$



$\theta^0 3$

Figure 2: True map of disease

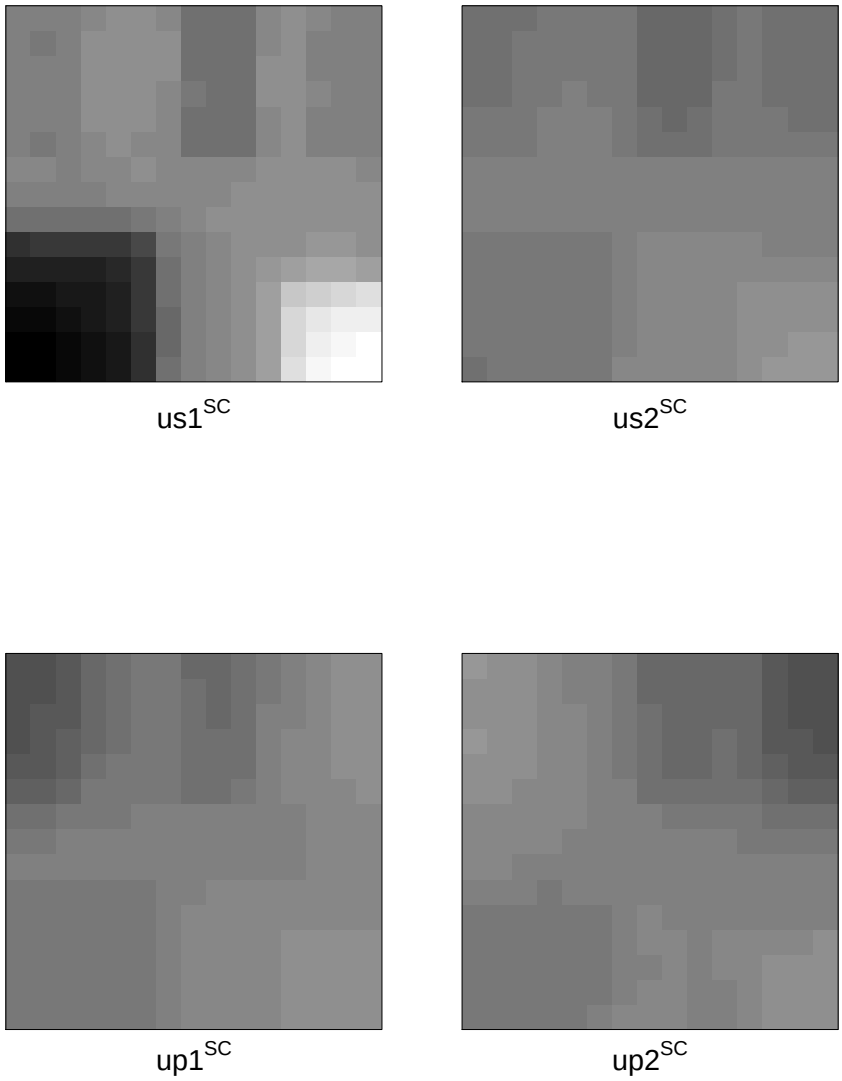


Figure 3: Estimated clustering terms with SC

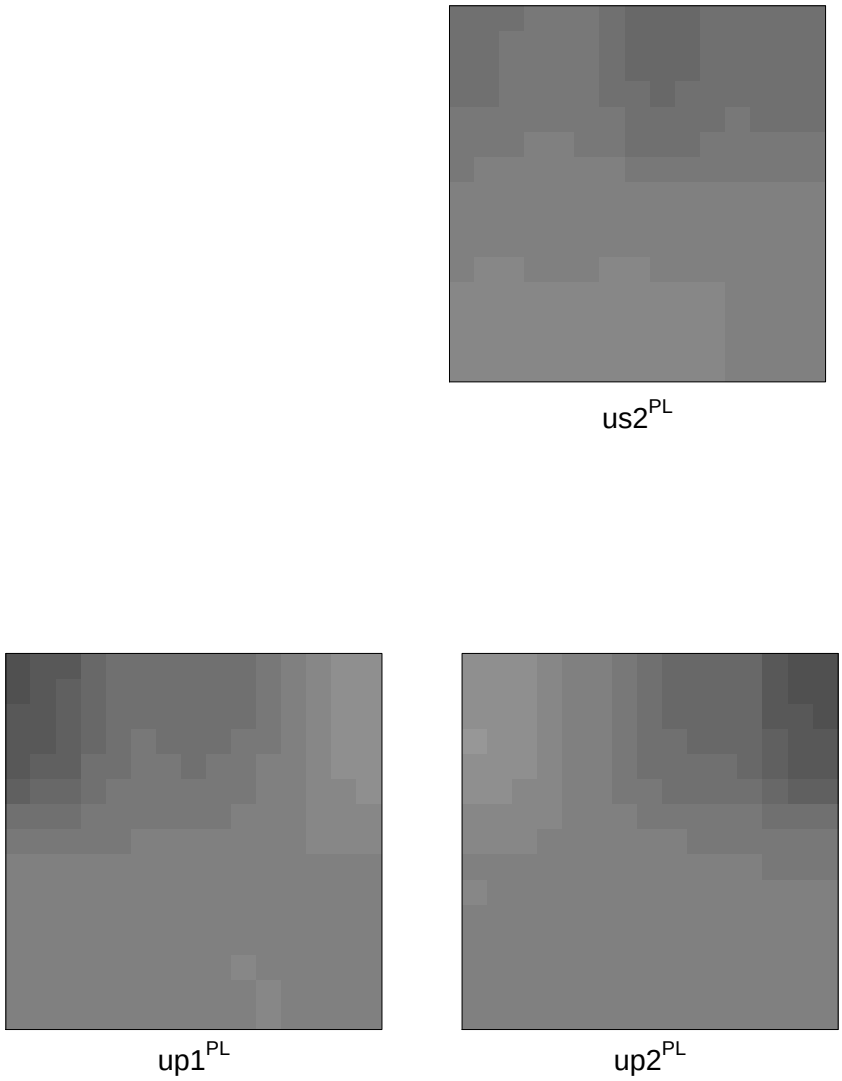


Figure 4: Estimated clustering terms with PL

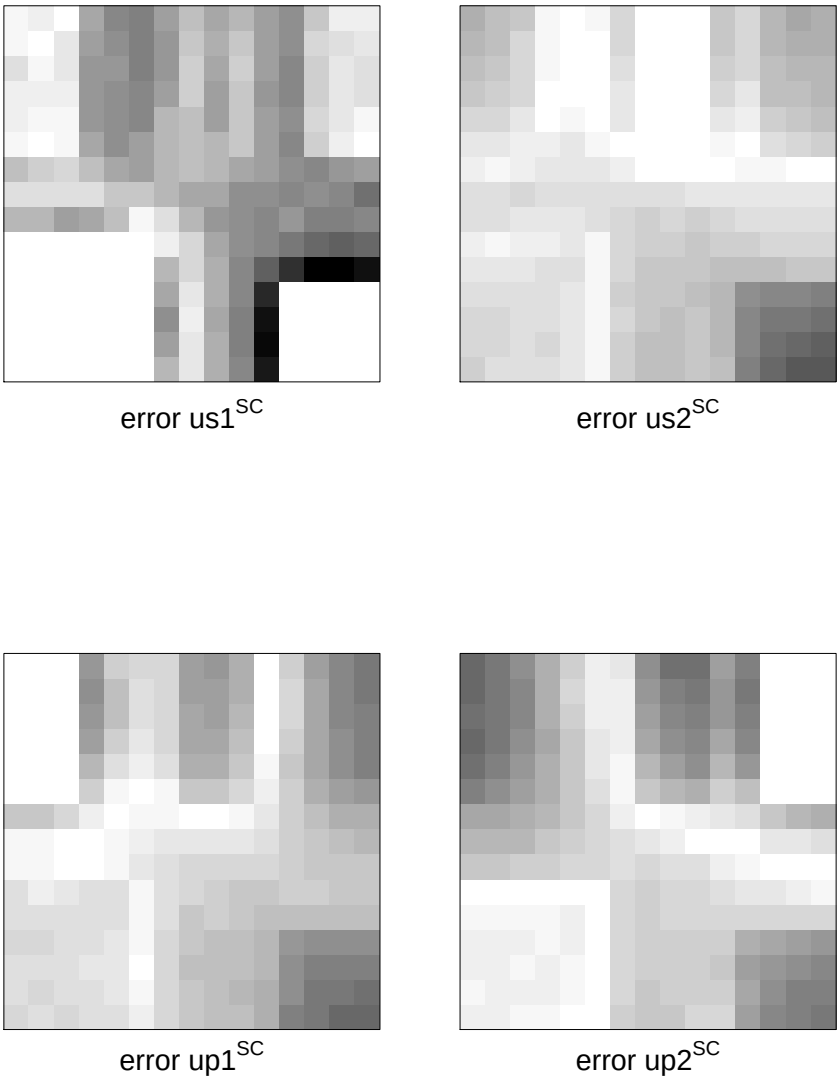
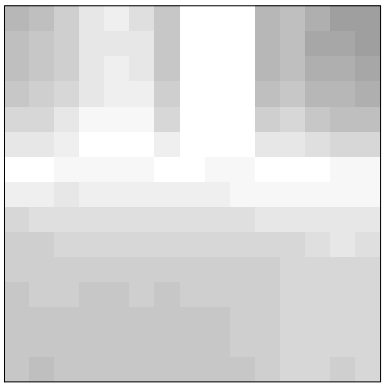
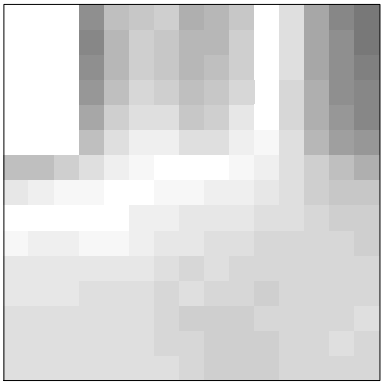


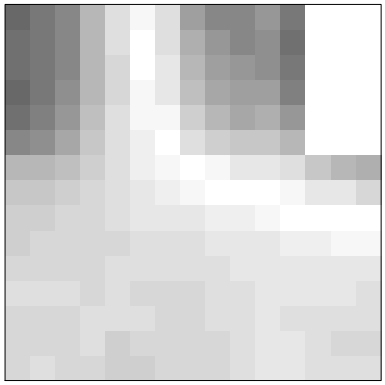
Figure 5: Relative error maps with SC



error $us2^{PL}$



error $up1^{PL}$



error $up2^{PL}$

Figure 6: Relative error maps with PL