Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 71-90

Small Area Estimation of Gender-based Violence: Rape Incidence Risks in Uttar Pradesh, India

Vicente, G.[1], **Goicoa, T.**[1,2], **Puranik, A.**[3], **and Ugarte, M.D.**[1,2,4]
[1] *Department of Statistics and O. R., Public University of Navarre, Spain*
[2] *Institute for Advanced Materials (InaMat), Public University of Navarre, Spain*
[3] *Department of Statistics, Prasanna School of Public Health, Manipal Academy of Higher Education, India*
[4] *Department of Mathematics, UNED Pamplona, Spain*

Received: November 20, 2017; Revised: December 28, 2017; Accepted: January 03, 2018

Abstract

Violence against women is considered an endemic problem in communities and countries around the world, and it has been declared an issue of epidemic proportions by the World Health Organization (WHO). In India, where the patriarchal nature of the country contributes to increasing violence against women, there has been a dramatic increase of this gender-based violence in the past decades. In this paper we focus on analyzing rape incidence risks in the most populous state of India. In particular, small area models including spatial, temporal, and spatio-temporal components are used to estimate rape incidence risks in the districts of Uttar Pradesh during the period 2001-2014. We discover interesting spatio-temporal patterns of rape incidence as well as point out districts with significant high risks.

Key words: Small area estimation; Gender-based violence; rape incidence, spatio-temporal

1 Introduction

Nowadays, public and private institutions demand information regarding different issues at very disaggregated administrative levels. This increasing demand of local statistics has promoted a huge development of small area estimation techniques. The key point in small area estimation is to produce reliable estimates for areas with very small (or even zero) sample size because they were not planned in the sample design originally conceived to produce estimates for larger regions. Classical design-based estimators are no longer applicable in small areas because they are extremely variable or even they cannot be calculated. Consequently, statistical models including auxiliary information become essential to derive precise estimates borrowing information from other areas,

Corresponding Author: María Dolores Ugarte
Email id:- lola@unavarra.es

past observations or both. Thorough reviews about the topic include Jiang and Lahiri (2006), Datta (2009), and Pfeffermann (2013). The book by Rao and Molina (2015) is a comprehensive and updated reference in this area.

A somehow different small area problem is disease mapping, where no sampling process is involved. Disease mapping deals with data from official registers and it is a small area problem because in some areas the classical estimators (such as standardized incidence or mortality ratios) could be very variable, particularly if the population is scarce or the phenomenon under study is rare, as it is the case of rape incidence in India analysed in this paper. Then models are required to improve estimates, to disentangle spatial patterns and temporal trends, and to look into the space-time interactions. Typically, disease mapping models include conditional autoregressive models for the spatial effects and random walk priors for the temporal effects. Interaction effects are also usually included and they may be completely unstructured, or spatially and/or temporally structured (Knorr-Held, 2000). Model fitting is generally carried out under a fully Bayesian setting using Markov chain Monte Carlo (McMC) techniques (Gilks, 2005). However, when using McMC great care is devoted to the tuning and monitoring convergence phases in order to find the best setting (in terms of parameterization, prior distributions, initial values, and Metropolis Hastings proposal distributions) that gives the most reliable and accurate McMC output (Blangiardo and Cameletti, 2015). In addition, computational time could be sometimes prohibitive. To avoid these problems a new technique, based on nested Laplace approximations and numerical integration (INLA), has been recently developed (Rue et al., 2009). This fitting technique has been adopted in disease mapping (see for example Ugarte et al., 2014), but, as far as we know, it has been scarcely used in more “classical” small area applications. Two recent papers are Chen at al. (2014) and Mercer et al. (2014). This paper, in honour of Professor JNK Rao, is an opportunity to show the potential of this technique when fitting spatio-temporal models. We think that this exchange of ideas may be beneficial as disease mapping and small area estimation are closely related, but they are very often considered as two different research areas with their own methodology. There are examples where “classical” small area estimation methods have been applied in disease mapping. As an example, Ugarte et al. (2008, 2009a) and Goicoa et al. (2012) derive confidence intervals for the relative risks using some predictors of the mean squared error widely used in small area estimation. On the other hand, there are a battery of models including spatio-temporal interactions in the field of disease mapping that can be extremely helpful in small area estimation, where spatio-temporal models are not so abundant.

The goal of this paper is two-fold. Firstly, we use spatio-temporal models to estimate the evolution of the geographical pattern of rape incidence in the districts of Uttar Pradesh during the period 2001-2014. Secondly, high risks districts are identified. Rape in India was sadly the focus of international attention when a 23 year old student was gang raped and beaten on a bus in 2012 and later died because of the injuries. That year, two girls were also gang raped and hanged (Mullan, 2014). Raj and McDougal (2014) provide a concise report about sexual violence in India and affirm that the reported number of rapes is increasing. They also state that although prevalence of sexual violence in India is one of the lowest in the world (8.5%), it affects 27.5 million women. It is an underreported problem (1% of women report the crime) because most sexual violence takes place in marriage and marital rape in India is not a crime. It seems that this underreported problem did not improve much in developing countries since the nineties where other authors (Vogelman and

Eagle, 1991) stated, for example, that in South Africa only one out of twenty rapes were reported, and Heise et al. (1994) indicated that sexual assault and rape were usually underreported due to the social stigma about sexual crimes. Here, we focus on Uttar Pradesh, the most populous Indian state, to gain knowledge about the evolution of the rape incidence risk in the different districts of the state.

The rest of the paper is organized as follows. Section 2 offers descriptive figures about rapes in India and Uttar Pradesh. Section 3 presents different models including spatial, temporal, and spatio-temporal components, and gives some details about model fitting with INLA. Section 4 provides the data analysis. The paper closes with a discussion in Section 5.

2 Rape Crime in Uttar Pradesh

Uttar Pradesh is located in the North of India and it is the most populated state of the country. During the period 2001-2014 the number of districts in Uttar Pradesh has increased from 70 to 75 because some original districts have split into two or more additional new districts. As the National Crime Records Bureau (NCRB) provides the data at district level, we do not have data for the new districts at the beginning of the period. Consequently, we retain the original 70 districts during the whole period. The 70 districts labeled from 1 to 70 are displayed in Figure 1, and the names of the districts matching the labels are shown in Table 1. According to 2011 census, the total population of Uttar Pradesh is 199,812,341, and the female population is 95,331,831 (47.71% of total population). The state has a total area of 240,928 km^2, and a population density of 829 people per km^2, being higher than the national Indian average of 382 people per km^2. The literacy rate is 77.28% for male and 57.18% for female, with an overall state rate of 67.68%. The population figures are available from the 2001 and 2011 censuses, and for the rest of the years in the period the female population has been computed using linear interpolation. Data on crimes in Uttar Pradesh were provided by the NCRB, who has annually published data on different types of crime since 1953. Data on rape incidence has been recorded since 1971, whereas other crime typologies such as dowry death, cruelty by husband or relatives have been published since 1995.

According to National Crime Records Bureau (2014), in 2014, the last year of the study period, 38,467 out of 337,992 crimes against women (CAW) in India, occurred in Uttar Pradesh (11.38%). CAW include crimes against women under Indian Penal Code (rape, attempt to commit rape, kidnapping and abduction of women, dowry deaths, assault on women with intent to outrage their modesty, insult to the modesty of women, cruelty by husband or his relatives, importation of girl from a foreign country, abetment of suicide of women), and crimes against women under Special and Local Laws (commission of Sati Prevention Act, Indecent Representation of Women (P) Act, the Dowry Prohibition Act, Protection of Women from Domestic Violence Act, and Immoral Traffic (Prevention) Act).

Rapes in India in 2014 represent a 10.87% of CAW (36735 registered cases), whereas in Uttar Pradesh this figure is slightly lower, a 9.01% (3467 registered cases). The total number of rapes in Uttar Pradesh in 2014 represents the 9.43% of all rapes in India. Rape incidence rates in India and Uttar Pradesh have increased during the studied period. The increase is particularly noticeable in

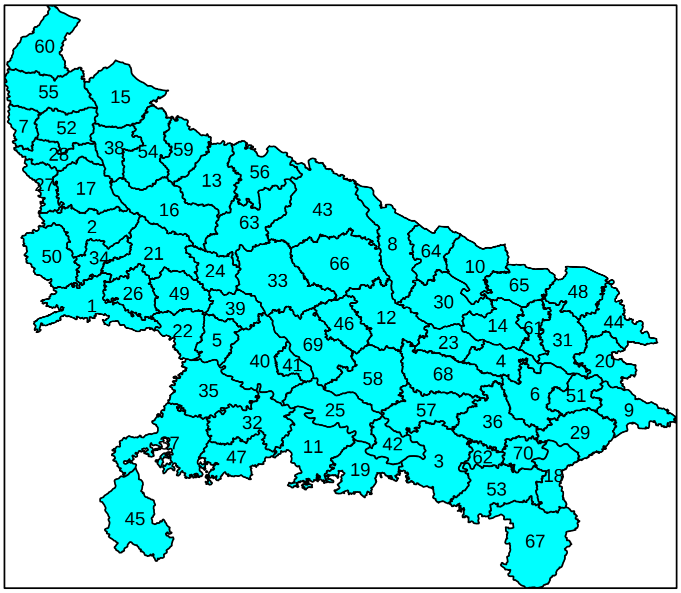

Figure 1: Districts of Uttar Pradesh.

the last year of the period (a 12.90% and a 7.02% in Uttar Pradesh and India respectively), probably due to a better support for victim disclosure (Raj and McDougal, 2014).

The growing trend that is observed in Uttar Pradesh is not the same for all its districts. Rape incidence ratios (SIR) per year, in some districts of Uttar Pradesh (Aligarh, Meerut, Sant Ravidas Nagar Bhadohi, Hardoi, Pilibhit, and Deoria) are displayed in Figure 2. We observe that some of the districts present variable SIRs, and hence models are needed to unveil the spatial and temporal evolution of rape incidence in the districts of Uttar Pradesh.

3 Spatio-temporal Models and Fitting Details

3.1 Spatio-temporal Models

In this section, different spatio-temporal models are proposed to study the evolution of rape incidence risk in the districts of Uttar Pradesh. We start with the Bernardinelli et al (1995) model including a linear temporal trend. Then we consider more flexible non-parametric spatio-temporal models (see Ugarte et al., 2014). For all model proposals, let O_{it} be the number of rapes in district i, $i = 1, \ldots, S$, and time period t, $t = 1, \ldots, T$; and let n_{it}, E_{it}, and R_{it} represent the female population, the expected number of rapes, and the relative risk corresponding to that district and period. Here E_{it} is computed as $E_{it} = n_{it} \cdot R$, where R is calculated as $R = \sum_i \sum_t O_{it} / \sum_i \sum_t n_{it}$. Then, conditional on the relative risk (R_{it}), the number of rapes (O_{it}) is assumed to follow a Poisson distribution with mean $\mu_{it} = E_{it} \cdot R_{it}$, i.e

Table 1: District identifiers (ID) of Uttar Pradesh

ID	Districts	ID	Districts	ID	Districts
1	Agra	25	Fatehpur	49	Mainpuri
2	Aligarh	26	Firozabad	50	Mathura
3	Allahabad	27	Gautam Buddha Nagar	51	Mau
4	Ambedkar Nagar	28	Ghaziabad	52	Meerut
5	Auraiya	29	Ghazipur	53	Mirzapur
6	Azamgarh	30	Gonda	54	Moradabad
7	Baghpat	31	Gorakhpur	55	Muzaffarnagar
8	Bahraich	32	Hamirpur	56	Pilibhit
9	Ballia	33	Hardoi	57	Pratapgarh
10	Balrampur	34	Hathras	58	Rae Bareli
11	Banda	35	Jalaun	59	Rampur
12	Barabanki	36	Jaunpur	60	Saharanpur
13	Bareilly	37	Jhansi	61	Sant Kabir Nagar
14	Basti	38	Jyotiba Phule Nagar	62	Sant Ravidas Nagar (Bhadohi)
15	Bijnor	39	Kannauj	63	Shahjahanpur
16	Budaun	40	Kanpur Dehat	64	Shrawasti
17	Bulandshahr	41	Kanpur Nagar	65	Siddharthnagar
18	Chandauli	42	Kaushambi	66	Sitapur
19	Chitrakoot	43	Kheri	67	Sonbhadra
20	Deoria	44	Kushinagar	68	Sultanpur
21	Etah	45	Lalitpur	69	Unnao
22	Etawah	46	Lucknow	70	Varanasi
23	Faizabad	47	Mahoba		
24	Farrukhabad	48	Mahrajganj		

$$O_{it}|R_{it} \sim Poisson(\mu_{it}),$$
$$\log(\mu_{it}) = \log(E_{it}) + \log(R_{it}),$$

where $\log(E_{it})$ is an offset. We first start with the following spatio-temporal additive model with linear temporal trend, that will be denoted as Model 1

$$\log(R_{it}) = \alpha + \beta \cdot t + \xi_i, \tag{3.1}$$

where α is the intercept, β is the slope of the temporal linear trend common to all districts, and ξ_i is a spatial random effect with a Leroux et al. (1999) prior distribution. That is,

$$\boldsymbol{\xi} \sim N\left(\mathbf{0}, \sigma_\xi^2 \boldsymbol{D}^{-1}\right),$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_S)'$ and $\boldsymbol{D} = \lambda_\xi \boldsymbol{Q}_\xi + (1-\lambda_\xi)\boldsymbol{I}_S$. $\boldsymbol{Q}_\xi$ is the neighbourhood matrix defined by contiguity (two districts are neighbours if they share a common border). The diagonal elements are equal to the number of neighbours of each district, and non-diagonal elements $(\boldsymbol{Q}_\xi)_{ij} = -1$ if districts i and j are neighbours and $(\boldsymbol{Q}_\xi)_{ij} = 0$ otherwise. The spatial smoothing parameter λ_ξ

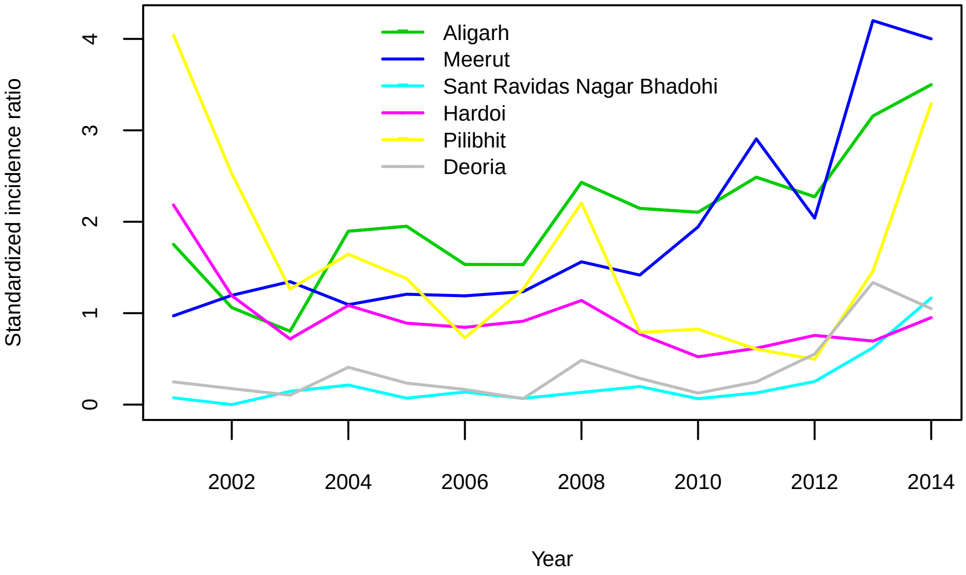

Figure 2: Standardized incidence ratio (SIR) of rapes by year in the districts Aligarh, Meerut, Sant Ravidas Nagar Bhadohi, Hardoi, Pilibhit, and Deoria.

takes values between 0 and 1, and $\boldsymbol{I}_S$ is the $S \times S$ identity matrix. Model (3.1) assumes the same linear trend for all districts, something that can be very restrictive. Hence, the model is extended considering a linear space-time interaction term (in the following we denote this model as Model 2)

$$\log(R_{it}) = \alpha + (\beta + \delta_i) \cdot t + \xi_i, \tag{3.2}$$

where $\delta_i \sim N(0, \sigma_\delta^2)$, $i = 1, \ldots, S$. Although this model allows for different time trends in each district, it is rigid as it forces all the time trends to be linear, something that is not very realistic in practice. A more flexible class of spatio-temporal models is considered next

$$\log(R_{it}) = \alpha + \xi_i + \gamma_t + \delta_{it}, \tag{3.3}$$

where γ_t is a temporal random effect common to all areas, and δ_{it} is a space-time interaction, allowing for a specific temporal evolution of each district. A random walk prior of first (RW1) or second (RW2) order is assumed for the vector of temporal random effects $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)'$, i.e.,

$$\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma_\gamma \boldsymbol{R}_\gamma^-),$$

where $\boldsymbol{R}_\gamma$ is the structure matrix of a RW1 or a RW2 (see Rue and Held, 2005, pp 95 and 110), and $^-$ denotes the Moore-Penrose generalized inverse. The spatio-temporal interaction random effect $\boldsymbol{\delta} = (\delta_{11}, \ldots \delta_{S1}, \ldots, \delta_{1T}, \ldots \delta_{ST})'$ is assumed to follow a multivariate normal distribution

$$\boldsymbol{\delta} \sim N(\boldsymbol{0}, \sigma_\delta \boldsymbol{R}_\delta^-),$$

where depending on the definition of the structure matrix $\boldsymbol{R}_{\delta}$, four different types of interactions arise (see Knorr-Held, 2000), denoted as Type I, Type II, Type III, and Type IV.

In Type I interaction, $\boldsymbol{R}_{\delta} = \boldsymbol{I}_T \otimes \boldsymbol{I}_S$, where $\boldsymbol{I}_T$ is the $T \times T$ identity matrix so that all δ_{it} are independent. In Type II interaction, $\boldsymbol{R}_{\delta} = \boldsymbol{R}_{\gamma} \otimes \boldsymbol{I}_S$, the δ_{it}'s are structured in time but unstructured in space. In Type III interaction, $\boldsymbol{R}_{\delta} = \boldsymbol{I}_T \otimes \boldsymbol{Q}_{\xi}$, and the δ_{it}'s are structured in space but unstructured in time. Finally, $\boldsymbol{R}_{\delta} = \boldsymbol{Q}_{\xi} \otimes \boldsymbol{R}_{\gamma}$ in Type IV interaction and the δ_{it}'s are structured in space and time, that is, temporal trends from neighbouring districts tend to be similar. In the following we will denote as Model 3a, an additive model with a RW1 prior for time, that is, Model (3.3) without the interaction terms δ_{it}. Model 4a, Model 5a, Model 6a, and Model 7a will refer to models with a RW1 prior for time and Type I, Type II, Type III, and Type IV interactions respectively. Analogously, we will denote by Model 3b, Model 4b, Model 5b, Model 6b, and Model 7b when a RW2 prior is used for the temporal random effects.

In all the models in this section, the spatial correlation allows to "borrow strength" from neighbouring areas, and the main temporal trend "borrows information" from close time periods. The space-time interaction models the specific behaviour of an area at a given year on top of the common spatial and temporal terms.

3.2 Model Fitting via INLA

Model fitting and inference has been carried out under a fully Bayes perspective using INLA (Rue et al., 2009). Although the INLA approach is being increasingly used in disease mapping (see Schrödle and Held, 2011a, 2011b; Goicoa et al., 2016, 2017a) and it is relatively easy to use, some caution is needed when fitting spatio-temporal models as these models are usually non-identifiable and adequate constraints need to be specified to achieve identifiability (see Goicoa et al., 2017b and Ugarte et al., 2017). In addition, changing the default hyperparameter priors is also not straightforward. Nevertheless, INLA has been shown to work well in spatial and spatio-temporal disease mapping and some comparison with McMC in this setting are available. Both fitting methods lead to practically identical results if a simplified or full Laplace strategy is adopted with INLA, (see, for example, Rue et al., 2009, and Schrödle et al., 2011). Moreover, Ugarte et al. (2014) compare INLA and PQL (penalized quasi-likelihood) and the results are nearly the same.

One of the advantages of the INLA methodology is that it is ready to use in R (R core team 2017) through the R-package R-INLA (Martino and Rue, 2009). To fit the models in R-INLA it is necessary to define a `formula` object of the type

```
formula <- response ~ f(...,model="",...) + ...
```

where `f()` is a function that R-INLA uses to implement the different terms of the models, such as the spatial, temporal, and spatio-temporal random effects. Then, the INLA algorithm is run through the function `inla()`,

```
inla(formula, family=<family>, data=<data>, ...)
```

where `family` is a string indicating the likelihood family, and `data` is a data frame containing all the variables included in the model.

A vague normal distribution with a precision close to zero was considered for the intercept (α). A Unif(0,1) prior was given to the spatial parameter λ_ξ, and Uniform prior distributions on the positive real line were given to the standard deviations σ_ξ, σ_γ, and σ_δ. INLA places a log Gamma(1, 0.00005) prior on the log precision parameter by default, but this type of prior has been criticised as it may produce uncorrect results (Carroll et al., 2015; Simpson et al., 2017). An in depth study about prior distributions for variance parameters in hierarchical models is provided in Gelman (2006). This author advocates the use of a noninformative uniform prior density on standard deviation parameters σ in hierarchical models as he expects this will generally work well. We use an improper uniform prior on the standard deviation because if something is wrong with the posterior (improper posterior), that could be veiled using vague but proper priors. We have also conducted the analysis with the default INLA log Gamma priors and the results are similar (see also Goicoa et al., 2017a).

The Deviance Information Criterion (DIC) (Spiegelhalter *et al.* 2002), the Watanabe-Akaike Information Criterion (WAIC) (Watanabe 2010), and the logarithmic score (LS) (Gneiting and Raftery 2007) were used to select the best model.

4 Data Analysis

In this section, the parametric and non-parametric models described in Section 3 are used to fit rape incidence data in Uttar Pradesh during the period 2001-2014. All models have been fitted in R-INLA using a full Laplace strategy. Table 2 displays the deviance ($\bar{D}$), the effective number of parameters (p_D), the DIC and WAIC model selection criteria, and the logarithmic score (LS) (a measure of model prediction performance). According to all criteria, the parametric models are not good options as they are rather restrictive models. Simpler (additive) models including only spatial or temporal trends have also been assessed but they do not fit the data well. Moreover, the classical standardized incidence risks (SIR) have been computed, but a significant spatial correlation has been detected using a Monte Carlo approach of the Moran test in all years indicating the need of models with spatial terms. To compute the Moran test we use function `moran.test` of the `R` package `spdep` (see Bivand and Piras, 2015 and Bivand et al., 2013). Non-parametric models including space-time interactions are clearly preferred as the non-parametric additive models do not exhibit a good performance in terms of model selection criteria. Model 5a, a non-parametric spatio-temporal model with a RW1 prior for time and a Type II spatio-temporal interaction (i.e., RW1 temporal trends that are spatially unstructured), is selected as the best candidate to analyse rape incidence in the districts of Uttar Pradesh, and it is the model we finally consider to analyze the data. The R code to fit this model can be found in the Appendix.

Figure 3 displays the posterior mean of the district specific spatial risk $\zeta_i = \exp(\xi_i)$ (top) and the posterior probability that this risk is greater than one, $P(\zeta_i > 1|\mathbf{O})$, (bottom). Districts

Table 2: Model selection criteria.

		$\overline{D}$	p_D	DIC	$WAIC$	LS
Parametric models						
Model 1	Additive	8884.59	68.91	8953.50	9225.84	4.71
Model 2	Interaction	7939.06	138.05	8077.10	8573.38	4.40
Non-parametric models						
RW1						
Model 3a	Additive	7561.71	80.94	7642.64	7845.84	4.01
Model 4a	Type I	5684.21	622.34	6306.55	6234.93	3.45
Model 5a	Type II	5716.77	459.31	**6176.08**	**6187.34**	**3.29**
Model 6a	Type III	5760.62	562.83	6323.44	6324.46	3.45
Model 7a	Type IV	5774.53	423.55	6198.09	6246.22	3.30
RW2						
Model 3b	Additive	7562.05	80.93	7642.98	7846.18	4.01
Model 4b	Type I	5981.72	325.72	6307.44	6436.75	3.37
Model 5b	Type II	5684.08	622.93	6307.01	6234.79	3.45
Model 6b	Type III	5980.30	336.62	6316.92	6447.12	3.38
Model 7b	Type IV	5761.54	562.63	6324.18	6325.78	3.45

with posterior probabilities greater than 0.9 are considered high risk districts (see Richardson et al., 2004; Ugarte et al., 2009a, 2009b, for Bayesian decision rules to detect high risk regions). A clear West-East gradient is observed in incidence rape. The districts in the west and central part of the state present high incidence risks whereas districts in the East exhibit low incidence risks. This spatial pattern ζ_i is the same along the period and can be interpreted as a basic risk associated to each district.

It is very difficult to find out the reasons of high or low spatial risks. Here we provide some information that may hypothesize about factors related to incidence rape. Most of the districts with high spatial risks present a sex-ratio (females per 1000 males) smaller than the sex-ratio in whole Uttar Pradesh, whereas most districts with low spatial risks have a sex-ratio greater than in the state. Also, the percentage of rural population is higher than the percentage of urban population in low spatial risk districts. Some studies (see Bruinsma, 2007) associate rural areas with lower levels of crime due to a higher social cohesion, yet this study is focussed on Dutch cities, and the results may not be comparable due to the big social and cultural differences. Most of the low spatial risk districts have a higher male and female literacy rate than whole Uttar Pradesh, whereas in scarcely half of the high spatial risk districts these literacy rates are over those in the overall state.

The global rape incidence temporal trend, the posterior mean of exp (γ_t), is shown in Figure 4. This temporal pattern is common to all districts and may be associated with public policies common to the entire country (or state). From 2003 onward, this temporal risk has increased, this rise being particularly noticeable since 2012. It is suspected that changes in public policies, such as support for victim disclosure, may have induced women to report rape crimes producing this increase in the temporal risk trend.

The posterior means of the interaction effect exp (δ_{it}) are displayed in Figure 5. This term

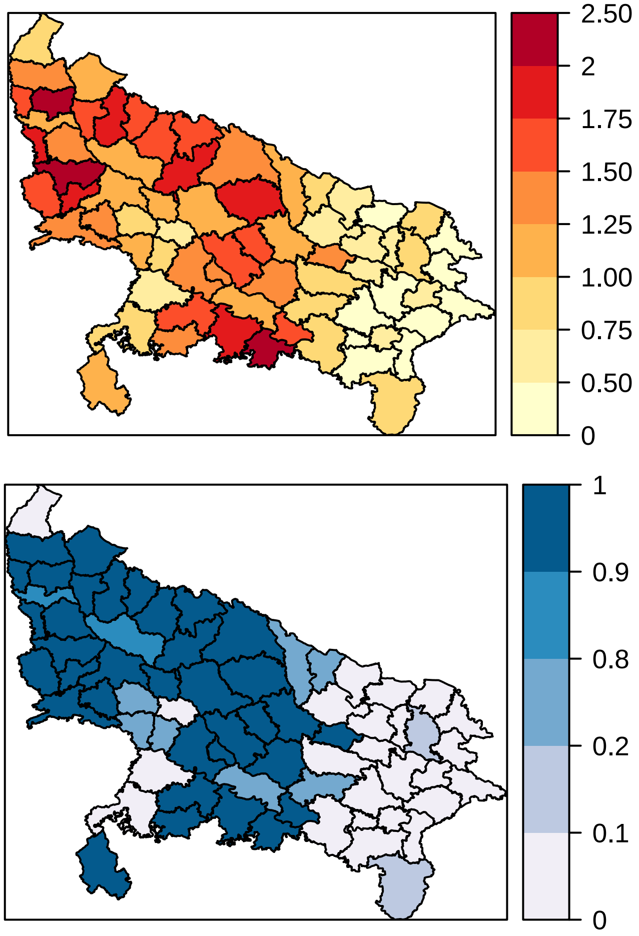

Figure 3: Posterior mean of the district-specific relative risk, $\zeta_i = \exp\left(\xi_i\right)$ (top), and posterior probabilities that the spatial risk is greater than one, P($\zeta_i > 1$| **O**) (bottom).

can be seen as a year-district specific term and can be interpreted as the deviation of each district from the basic spatial risk exp $\left(\xi_i\right)$. Those values of $\exp(\delta_{it})$ greater (lower) than one contribute to increase (decrease) the basic spatial risk. For example, districts such as Kheri, Sitapur, and Hardoi (central-north) contribute to raising the final risk in the early years of the study, but contribute to decreasing overall risks in recent years. Districts Azamgarh, Sant Kabir Nagar, and Gorakhpur exhibit the opposite behavior.

Figure 6 displays the temporal evolution of the posterior mean of the interaction effects exp $\left(\delta_{it}\right)$ (with 95% credible intervals). Clearly, these specific temporal trends are different among districts. Some districts exhibit decreasing specific temporal trends, whereas others show increasing trends. These temporal trends can be interpreted as district specific deviations of the overall temporal trend, and they may be capturing the effects of particular policies in each district along time.

Figure 7 displays maps of the posterior means of the relative risk (top) and the posterior probabilities that the relative risks are greater than one, $P(R_{it} > 1|\mathbf{O})$, (bottom), during the period

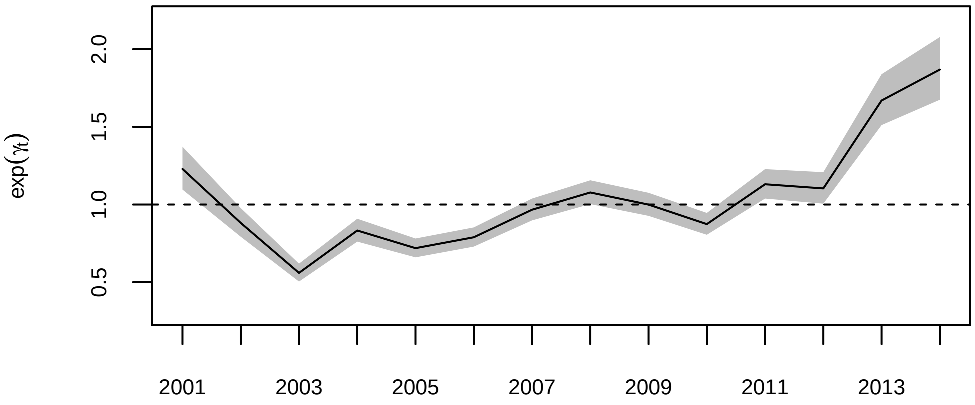

Figure 4: Global temporal pattern of rape incidence risk in Uttar Pradesh (exp (γ_t)).

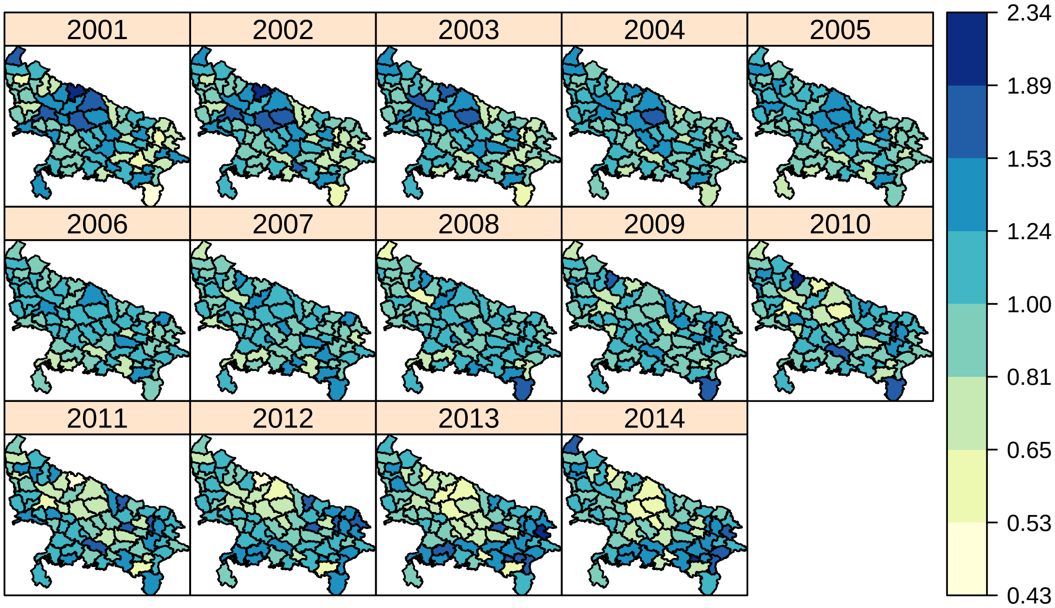

Figure 5: District-year interaction effect of rape incidence in Uttar Pradesh (exp (δ_{it})).

2001-2014. A clear increase in risks is observed from 2003 onward. This rise in risk is more noticeable in the two last year of the period, probably due to an increase of rape reports. The map

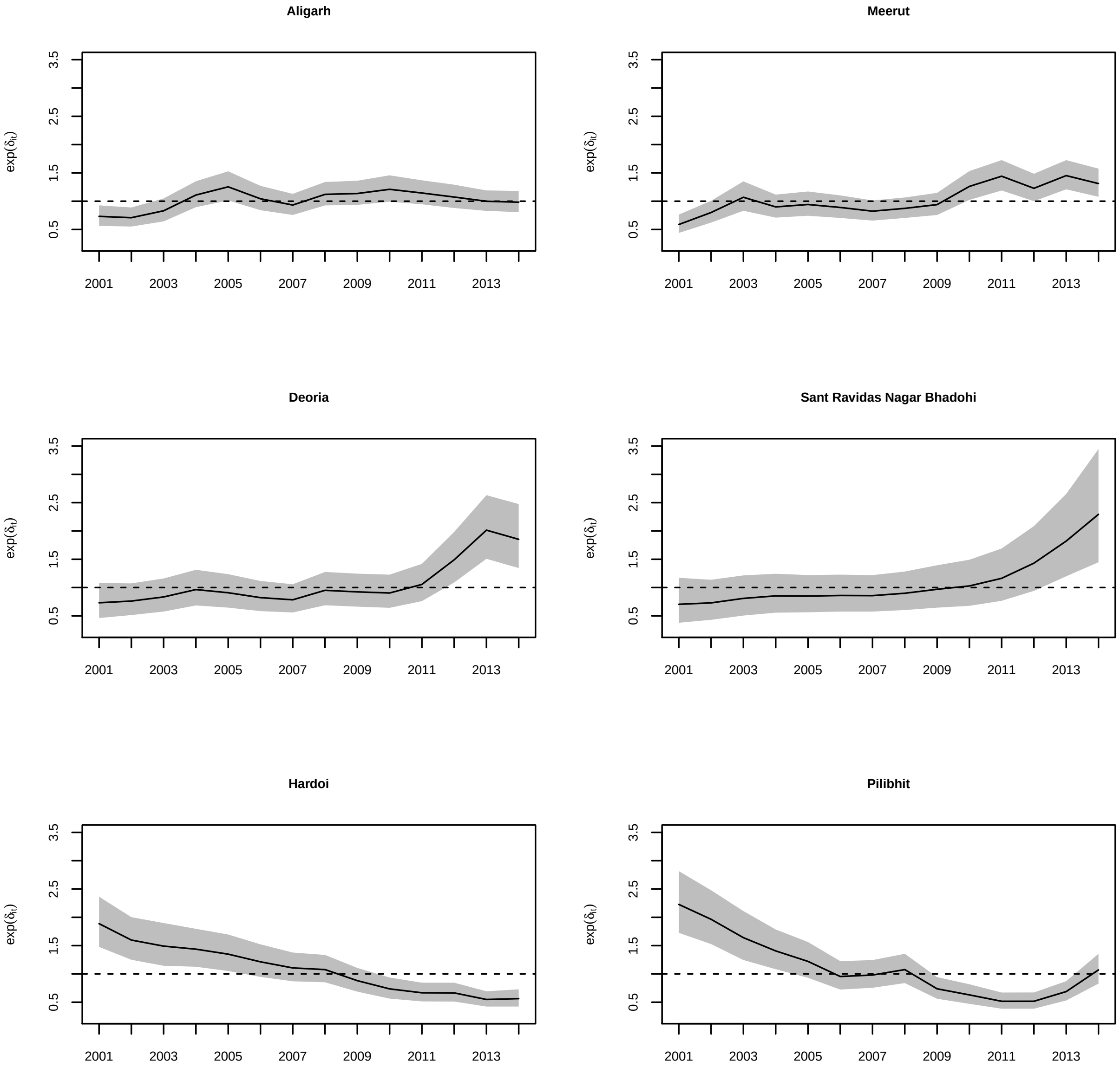

Figure 6: Specific temporal trends (exp (δ_{it})) for six selected districts: Aligarh, Meerut, Deoria, Sant Ravidas Nagar Bhadohi, Hardoi, and Pilibhit.

of posterior probabilities reveals that in these last two years, the state is divided into the west and central area where the posterior probabilities that the risks are greater than one exceed 0.9, and the most eastern part of the district where these probabilities are below 0.1. Then, the districts in the west-central part of Uttar Pradesh are high risks districts and a few regions in the east part of the state are low risk districts.

Finally, the percentages of total variability explained by the spatial, temporal, and the interac-

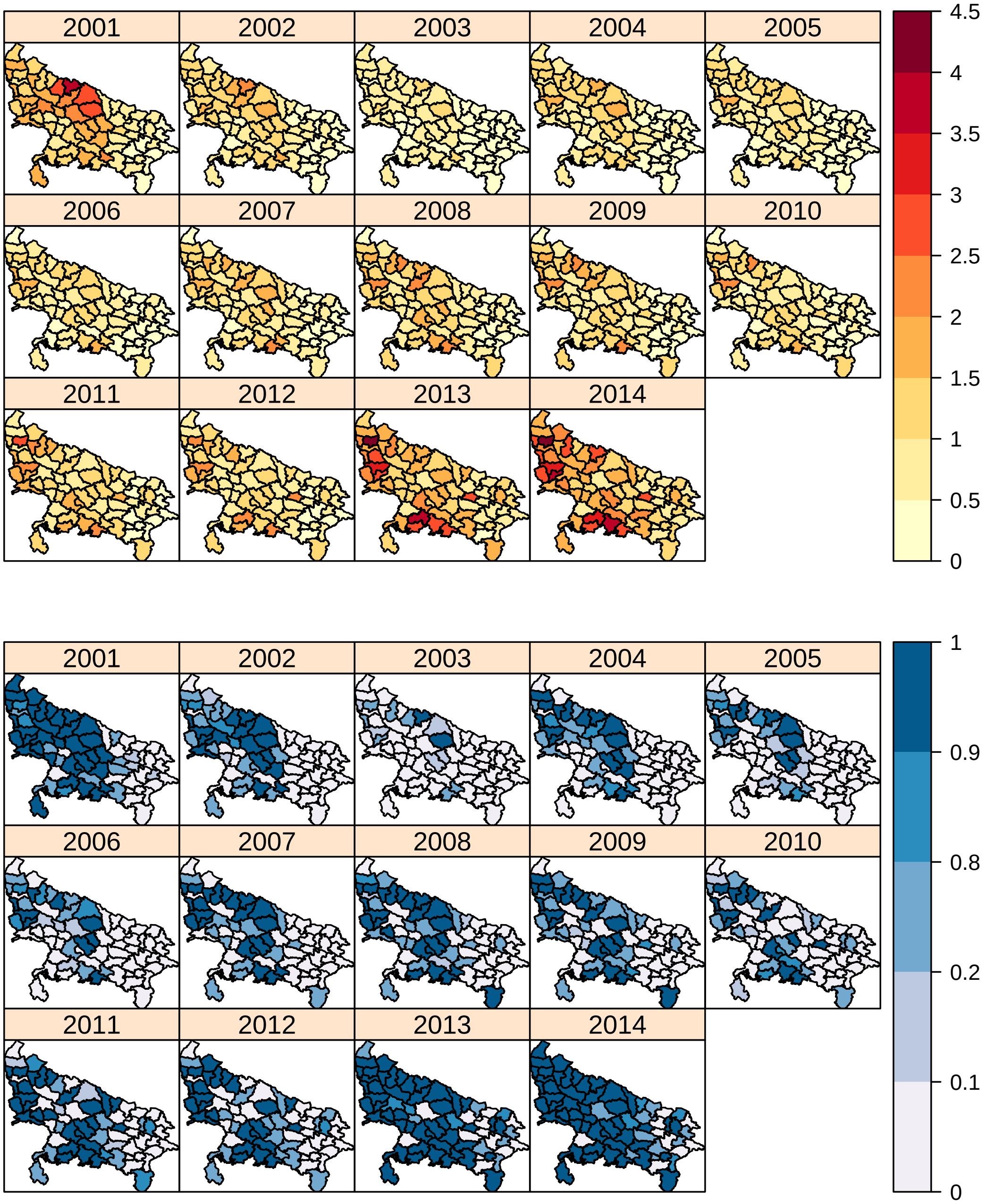

Figure 7: Map of the posterior means of rape incidence risks in Uttar Pradesh during the period 2001-2014 (top), and posterior probabilities that the relative risks are greater than one, $P(R_{it} > 1|\mathbf{O})$ (bottom).

tion terms are 64%, 22%, and 13% respectively (see Adin et al., 2017, for the decomposition of the variability). Though a large portion of variability is spatial, the temporal term plays an important role capturing the potential effects of the different policies along time. The percentage of specific space-time variability is also fairly noticeable.

5 Discussion

Disease mapping is a small area estimation problem where sampling is not involved. However, there has been a huge development of statistical techniques that may be very well applied in "classical" small area applications involving sampling processes and auxiliary information. This paper in honour of Professor JNK Rao, one of the most outstanding researchers in the field, is an excellent opportunity to see the potential of the disease mapping toolkit in general small area estimation problems. In particular, we put emphasis on the use of approximate Bayesian inference using integrated nested Laplace approximations (INLA) to speed up computations in comparison to McMC techniques in spatio-temporal disease mapping. Although disease mapping is a small area problem, computational time is an important issue when the number of areas or time periods becomes large.

In this paper we consider spatio-temporal models to analyze rape incidence in the districts of Uttar Pradesh, the most populous state in India. Prevalence of sexual violence in India is one the lowest in the world but it affects millions of women, and it is believed to be underreported due to social stigma. As far as we know this is the first study using spatio-temporal models to unveil spatio-temporal patterns of rape incidence. These models allow for identifying high risk areas and provide valuable information to make conjectures about the potential causes or risk factors related to rape incidence, which in turn makes possible to establish prevention programs. The analysis of rape incidence risks in Uttar Pradesh reveals that from 2003 onward there has been a steady increase in risk, particularly noticeable in the last years of the period. This may be caused by a rise in rape reports due to victims support policies. Our results show that in the last two years of the period, the state is divided into two groups of districts: those with a significant excess of risk (more than 150%) in comparison to the whole state in the west and central part of Uttar Pradesh, and those with low relative risk in the east side of the state (see years 2013 and 2014 in Figure 7). This spatial pattern reveals some interesting findings, yet it is not possible to clearly establish the risk factors associated to rape crimes. In general, in most of the low risk districts, male and female literacy rate and sex-ratio are greater that in the whole state. The opposite happens in most of the high risk districts. Also, most of low risks districts have a higher percentage of rural population.

Our study has however some limitations related in general to some data shortcomings apart from the crucial problem of underreported cases. Firstly, there are two districts (Farrukhabad and Kanpur Dehat) with no information on the number of incident rapes. We have imputed the data using information from neighbouring districts. Secondly, the data are not disaggregated by age group and the number of expected cases have been computed using a global rate, something that may lead to some bias if the rape incidence varies greatly among the different age groups in the state. In addition, we have used the complete female population, and again this may lead to some bias if we do not expect rape crimes in old women or very young girls. Finally, an additional limitation of our study arises from obtaining the populations using linear interpolation. In any case, and despite these limitations, we believe that our study may help to detect rape hot spots and, more importantly, to better establish prevention plans and education policies in Uttar Pradesh.

Acknowledgements

This work has been supported by Spanish Ministry of Economy and Competitiveness (project MTM2014-51992-R) [Ugarte and Goicoa], by the Spanish Ministry of Economy, Industry, and Competitiveness (project MTM2017-82553-R) jointly financed with the European Regional Development Fund (FEDER) [Ugarte, Goicoa, Vicente], and by Fundación CAN - Obra Social Caixa - UNED Pamplona 2017 [Ugarte, Goicoa]. The work by Puranik has been supported by project NRDMS/01/122/015 (Department of Science and Technology, India). On behalf of all the authors, Lola Ugarte is very grateful to the editors of this issue for inviting her to contribute with this work to felicitate his beloved friend and excellent researcher Jon Rao on his 80th birthday. They have both spent unforgettable moments together in international SAE meetings.

Appendix

The R code to fit the selected spatio-temporal Model 5a in INLA is presented below. The data frame that contains the model variables is defined as follows

```
> Data <- data.frame(O=<observed>, E=<expected>,
+                    ID.area=rep(1:n,times=t),
+                    ID.area1=rep(1:n,times=t),
+                    ID.year=rep(1:n,each=t),
+                    ID.area.year=seq(1,(n*t)))
```

where `observed` and `expected` are the vectors of observed and expected rape incidence cases respectively, and n and t are the number of small areas and time periods for which data are available.

The spatial neighborhood matrix $\boldsymbol{Q}_{\xi}$ and the structure matrix needed to implement the LCAR prior in INLA are defined as

```
> spdep::nb2INLA("uttar_pradesh_nb.graph", poly2nb(Carto))
> g <- inla.read.graph("uttar_pradesh_nb.graph")

> Q_xi <- matrix(0, g$n, g$n)
> for (i in 1:g$n){
+  Q_xi[i,i]=g$nnbs[[i]]
+  Q_xi[i,g$nbs[[i]]]=-1
+ }
> Q_Leroux <- diag(n)-Q_xi     #matrix to define the LACAR in INLA
```

where `"uttar_pradesh_nb.graph"` is an inla.graph object containing the neighbouring structure of the Uttar Pradesh districts.

We define the temporal structure matrix $\boldsymbol{Q}_\gamma$ of a random walk of first order as

```
> D1 <- diff(diag(t), differences=1)
> Q_gammaRW1 <- t(D1)%*%D1
```

We define the hyperparameter priors,

```
## Uniform distribution on the real line
> sdunif="expression:
+ logdens=-log_precision/2;
+ return(logdens)"

## Unif (0,1) or Beta(1,1)
> lunif = "expression:
+ a = 1;
+ b = 1;
+ beta = exp(theta)/(1+exp(theta));
+ logdens = lgamma(a+b)-lgamma(a)-lgamma(b)+(a-1)*log(beta)+
+ (b-1)*log(1-beta);
+ log_jacobian = log(beta*(1-beta));
+ return(logdens+log_jacobian)"
```

The formula for a Model 5a (Type II interaction and RW1 prior for time) is defined as

```
> R <- kronecker(Q_gammaRW1,diag(n))
> r_def <- n
> A_constr <- kronecker(matrix(1,1,t),diag(n))

> formula <- O ~  f(ID.area, model="generic1",
+                   Cmatrix=Q_Leroux, constr=TRUE,
+                   hyper=list(prec=list(prior=sdunif),
+                              beta=list(prior=lunif))) +
+                 f(ID.year, model="rw1", constr=TRUE,
+                   hyper=list(prec=list(prior=sdunif))) +
+                 f(ID.area.year, model="generic0", Cmatrix=R,
+                   rankdef=r_def, constr=TRUE,
+                   hyper=list(prec=list(prior=sdunif)),
+                   extraconstr=list(A=A_constr, e=rep(0,n)))
```

where R is the structure matrix $\boldsymbol{Q}_\gamma \otimes \boldsymbol{I}_n$, whose rank deficiency is specified in the `rankdef` argument. The linear constraints that makes this model identifiable are given by the `constr=TRUE` and the `extraconstr` arguments.

Finally, we run the INLA algorithm with a call to the `inla()` function as

```
> results<-inla(formula, family="poisson", data=Data, E=E,
+                    control.predictor=list(compute=TRUE, cdf=c(log(1))),
+                    control.compute=list(dic=TRUE, cpo=TRUE, waic=TRUE),
+                    control.inla=list(strategy="laplace", npoints=21))
```

where the approximation strategy to compute the marginal posteriors of the latent Gaussian field is specified with the `strategy` argument. In this case a full Laplace approximation (`"laplace"`) is used.

References

Adin, A., Martínez-Beneito, M.A., Botella-Rocamora, P., Goicoa, T. and Ugarte, M.D. (2017). Smoothing and high risk areas detection in space-time disease mapping: A comparison of p-splines, autoregressive, and moving average models. *Stochastic Environmental Research and Risk Assessment*, **31**, 403-415.

Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* **14**, 2433-2443.

Bivand, R. S., Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software***63(18)**, 1-36. URL http://www.jstatsoft.org/v63/i18/.

Bivand, R. S., Hauke, J., and Kossowski, T. (2013). Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis* **45(2)**, 150-179.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. Chichester: Wiley.

Bruinsma, G.J. (2007). Urbanization and urban crime: Dutch geographical and environmental research. *Crime and Justice* **35**, 453-502.

Chen, C., Wakefield, J. and Lumley, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-temporal Epidemiology* **11**, 33-43.

Datta, G.S. (2009). Model-based approach to small area estimation. *Handbook of Statistics* **29**, 251-288.

Gelman A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**(3), 515-534.

Gilks, W.R. (2005). Markov chain Monte Carlo. *Encyclopedia of Biostatistics*. Wiley.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359-378.

Goicoa, T., Adin, A., Etxeberria, J., Militino, A.F. and Ugarte, M.D. (2017a). Flexible bayesian p-splines for smoothing age-specific spatio-temporal mortality patterns. *Statistical Methods in Medical Research*, in press. DOI: 10.1177/0962280217726802

Goicoa, T., Adin, A., Ugarte, M.D. and Hodges, J.S. (2017b). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, in press. DOI 10.1007/s00477-017-1405-0.

Goicoa, T., Ugarte, M.D., Etxeberria, J. and Militino, A.F. (2012). Comparing car and P-spline models in spatial disease mapping. *Environmental and Ecological Statistics*, **19**, 573-599.

Goicoa, T., Ugarte, M.D., Etxeberria, J. and Militino, A.F. (2016). Age–space–time CAR models in Bayesian disease mapping. *Statistics in Medicine*, **35**, 2391-2405.

Heise, L.L., Raikes, A., Watts, C.H. and Zwi, A.B. (1994). Violence against women: A neglected public health issue in less developed countries. *Social Science & Medicine* **39**, 1165-1179.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test* **15**, 1-96.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**, 2555-2567.

Leroux, B.G., Lei, X., Breslow, N., Halloran, M. and Elizabeth, B.D. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical models in epidemiology, the environment, and clinical trials*, 179-191.

Martino, S. and Rue, H. (2009). Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*.

Mercer, L., Wakefield, J., Chen, C., and Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics* **8**, 69-85.

Mullan, Z. (2014). Gender-based violence: more research (funding) please. *The Lancet Global Health (Editorial)*, **2**, e672.

National Crime Records Bureau (2014). *Crime in India 2014 Compendium*. Ministry of Home Affairs

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* **28**, 40-68.

R Core Team (2014). R: A language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing; 2014.

Raj, A. and McDougal, L. (2014). Sexual violence and rape in India. *The Lancet (Correspondence)* **383**, 865.

Rao, J.N. and Molina, I. (2015). *Small-Area Estimation*. 2nd ed. Wiley.

Richardson, S., Thomson, A., Best, N. and Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives* **112**, 1016.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman; Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**, 319-392.

Schrödle, B. and Held, L. (2011a). A primer on disease mapping and ecological regression using. *Computational Statistics* **26**, 241-258.

Schrödle, B. and Held, L. (2011b). Spatio-temporal disease mapping using INLA. *Environmetrics* **22**, 725-734.

Schrödle, B., Held, L., Riebler, A. and Danuser, J. (2011). Using integrated nested laplace approximations for the evaluation of veterinary surveillance data from switzerland: A case-study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**, 261-279.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* **64**, 583-639.

Ugarte, M.D., Adin, A., Goicoa, T.(2017) One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: Model fitting and model identifiability. *Spatial Statistics* **22**, 451-468.

Ugarte, M.D., Adin, A., Goicoa, T. and Militino, A.F. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research* **23**, 507-530.

Ugarte, M.D., Goicoa, T. and Militino, A.F. (2009a). Empirical Bayes and fully Bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics & Data Analysis* **53**, 2938-2949.

Ugarte, M.D., Goicoa, T., Ibañez, B. and Militino, A.F. (2009b). Evaluating the performance of spatio-temporal bayesian models in disease mapping. *Environmetrics* **20**, 647-665.

Ugarte, M.D., Militino, A.F., and Goicoa, T. (2008). Prediction error estimators in Empirical Bayes disease mapping. *Environmetrics* **19**, 287-300.

Vogelman, L. and Eagle, G. (1991). Overcoming endemic violence against women in south africa. *Social Justice* **18**, 209-229.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571-3594.