

**USING INLA/SPDE APPROACH FOR ESTIMATING A SPATIAL MODEL FOR
LUNG CANCER MORTALITY IN ALGERIA 2016.**

**UTILISATION DE L'APPROCHE INLA/SPDE POUR L'ESTIMATION D'UN
MODELE SPATIALE POUR LA MORTALITE PAR CANCER DU POUMON EN
ALGERIE (2016).**

***Ayoub ASRI**

*Ecole des Hautes Etudes Commerciales (EHEC) et Ecole Supérieure de Statistiques et
Economie Appliquée (ENSSEA)*
a.asri@hec.dz, asri.ayoub@outlook.com

Rachid BENAMIROUCHE

Ecole Supérieure de Statistiques et Economie Appliquée (ENSSEA)
rbena2002@hotmail.com

Reçu le: 06/01/2021 **Accepté le :** 22/03/2021 **Publication en ligne le:** 29/06/2021

ABSTRACT: Recently, Bayesian methods have developed greatly in the field of public health. the advances of Markov Chain Monte Carlo methods (MCMC) has permitted Bayesian modelling to be vastly used by the research community. However, MCMC requires a huge amount of computational power especially with large datasets.

Now, Gaussian random fields have become increasingly useful in public health applications that are characterized with a spatial/spatio-temporal structure which is needed to be included when inferring models. An efficient alternative to MCMC has been developed using random fields. The Integrated Nested Laplace Approximation (INLA) combined with SPDE can provide a solution to large scale problems in the mentioned field.

In this paper we review the INLA approach and present an application on Lung Cancer mortality in Algeria using first an Areal model then combining both approaches to fit a Geostatistical SPDE model.

Keywords : Bayesian Modeling, Spatial modeling, Lung cancer mortality, INLA/SPDE Approach, Hierarchical models.

JEL Classification : C11 I12 C13

RESUME : Récemment, les méthodes bayésiennes se sont considérablement développées dans le domaine de la santé publique. Les progrès des méthodes Markov Chain Monte Carlo (MCMC) a permis à la modélisation bayésienne d'être largement utilisée par la communauté de recherche. Cependant, MCMC nécessite une énorme puissance de calcul, en particulier avec de grands ensembles de données.

* Auteur Correspondant.

Désormais, les champs aléatoires gaussiens sont devenus de plus en plus utiles dans les applications de santé publique qui sont caractérisées par une structure spatiale / spatio-temporelle qui doit être incluse lors de l'inférence des modèles. Une alternative efficace à MCMC a été développée en utilisant des champs aléatoires. L'approximation intégrée de Laplace (INLA) combinée avec SPDE peut fournir une solution aux problèmes à grande échelle dans le domaine mentionné.

Dans cet article, nous passons en revue l'approche INLA et présentons une application sur la mortalité par cancer du poumon en Algérie en utilisant d'abord un modèle territorial puis en combinant les deux approches pour estimer un modèle SPDE géostatistique.

Mots clés : Modélisation Bayésienne, Modélisation spatiale, mortalité par cancer du Poumon, L'approche INLA/SPDE, Modèles hiérarchiques.

1. INTRODUCTION :

Au cours des trois dernières décennies, les méthodes bayésiennes ont beaucoup évolué et sont maintenant largement établies dans de nombreux domaines de recherche, des essais cliniques (Scott, M. Berry et al., 2011), à l'évaluation économique de la santé (Baio, 2012) aux sciences sociales (Jackman, 2009), à l'épidémiologie (Greenland, 2006).

L'idée de base de l'approche bayésienne est qu'il n'existe effectivement qu'une seule forme d'incertitude, qui est décrite par des distributions de probabilité appropriées. Il n'y a donc pas de distinction fondamentale entre les données observables ou les paramètres non observables, qui sont également considérés comme des quantités aléatoires. L'incertitude quant à la valeur réalisée des paramètres compte tenu de l'état actuel de l'information (c.-à-d. avant d'observer de nouvelles données) est décrite par une distribution a priori. Le processus d'inférence combine les données a priori et le modèle de données (actuelles) pour dériver la distribution a posteriori, qui est généralement, mais pas nécessairement, l'objectif de l'inférence (Bernardo et al., 2000; Lindley, 2006)

L'approche bayésienne présente plusieurs avantages : par exemple, la spécification des distributions a priori permet l'inclusion formelle d'informations qui peuvent être obtenues par des études antérieures ou par l'avis des experts ; la probabilité (a posteriori) qu'un paramètre ne dépasse ou ne dépasse pas un certain seuil est facilement obtenue à partir de la distribution a posteriori, fournissant une quantité plus intuitive et interprétable qu'une p-value fréquentiste. En outre, dans l'approche bayésienne, il est facile de spécifier une structure hiérarchique sur les données et/ou les paramètres, qui présente l'avantage supplémentaire de faire des prévisions relativement simples pour les nouvelles observations et l'imputation des données manquantes.

Les données épidémiologiques, en termes d'une seule variable résultante et d'un ou plusieurs facteurs de risque ou facteurs de confusion, sont souvent caractérisées par une structure spatiale et/ou temporelle qui doit être pris en compte dans le processus d'inférence. Dans ces circonstances, l'approche bayésienne est particulièrement efficace (Dunson, 2001) et a été appliquée dans plusieurs applications épidémiologiques, de l'écologie (J. Clark S., 2005) aux études environnementales (J. S. Clark et al., 2006; Wikle, 2007) à des maladies infectieuses (Jewell et al., 2009). Par exemple, si les données sont constituées de comptages agrégés des résultats et des variables explicatives, il est

généralement possible de préciser la cartographie des maladies et/ou la régression écologique (Lawson, 2019). Alternativement, si les données sur les résultats ou les facteurs de risque sont observées à des endroits précis, les modèles géostatistiques sont considérées comme des représentations appropriées du problème (Diggle et al., 2007).

Les deux modèles peuvent être spécifiés dans un cadre bayésien en élargissant simplement le concept de structure hiérarchique, permettant de tenir compte des similitudes basées sur le voisinage ou sur la distance, pour les données de référence au niveau de la zone ou du point, respectivement. Toutefois, en particulier dans ces cas, le principal défi des statistiques bayésiennes réside dans les aspects de calcul. Les méthodes de Monte Carlo par Chaînes de Markov (MCMC) (Brooks et al., 2011; Robert et al., 2004) sont généralement utilisées pour le calcul bayésien, sans doute grâce à la grande popularité du logiciel BUGS (Lunn et al., 2009, 2012). Bien qu'extrêmement flexibles et capables de traiter pratiquement n'importe quel type de données et de modèles, dans tous les cas sauf les cas triviaux, les méthodes MCMC impliquent des simulations qui consomment beaucoup de temps et puissance de calculs pour obtenir la distribution a posteriori des paramètres. Par conséquent, la complexité du modèle et la dimension de la base de données demeurent souvent des questions fondamentales.

L'approche Integrated Nested Laplace Approximation (INLA ; (Rue et al., 2009)) a récemment été élaborée en tant que solution alternative efficace aux méthodes MCMC. INLA est conçu pour les modèles gaussiens latents, une classe très large et flexible de modèles allant de modèles linéaires (généralisés) à effet mixtes à des modèles spatiaux et spatio-temporels. Pour cette raison, INLA peut être utilisée avec succès dans une grande variété d'applications également grâce à la disponibilité du package du logiciel R nommé R-INLA (Lindgren et al., 2015). De plus, INLA peut être combinée à l'approche des équations aux dérivées partielles stochastiques (SPDE) proposée par (Lindgren et al., 2011) afin de mettre en œuvre des modèles spatiaux et spatio-temporels pour les données de référence ponctuelle.

L'objectif de cet article est de présenter les caractéristiques fondamentales de l'approche INLA appliquée aux données spatiales. Le document est structuré comme suit : tout d'abord dans la section 2 nous passons en revue les principales caractéristiques des données spatiales définies au niveau régional et ponctuel. Ensuite, nous donnons un aperçu de la théorie qui sous-tend INLA à la section 3. Ensuite, dans la section 4, nous passons en revue l'approche SPDE pour traiter les données géostatistiques. Puis, dans la section 5, on présente une application sur les données de mortalité par cancer du poumon en Algérie. Enfin, la section 6 traite certaines des questions et présente quelques conclusions.

2. DONNEES SPATIALES :

Les données spatiales sont définies comme la réalisation d'un processus stochastique indexé par l'espace

$$Y(s) \equiv \{y(s), s \in \mathcal{D}\} \quad (1)$$

où $\mathcal{D}$ est un sous-ensemble (fixe) de la $\mathbb{R}^d$ (ici nous considérons $d = 2$). Les données réelles peuvent alors être représentées par un ensemble d'observations $\mathbf{y} = \{y(s_1), \dots, y(s_n)\}$ où l'ensemble $(s_1, \dots, s_n)$ indique les unités spatiales dans lesquelles les mesures sont effectuées. Selon que $\mathcal{D}$ soit une surface continue ou une collection dénombrable d'unités spatiales à d dimensions, le problème peut être défini comme un processus aléatoire spatial continu ou discret, respectivement (Gelfand et al., 2010).

Par exemple, nous pouvons considérer un ensemble de mesures de polluants atmosphériques obtenues par des moniteurs situés dans l'ensemble $(s_1, \dots, s_n)$ de n points. Dans ce cas, $\mathbf{y}$ est une réalisation du processus de pollution atmosphérique qui change continuellement dans l'espace et que nous appelons généralement des données géostatistiques ou ponctuelles. Alternativement, nous pouvons être intéressés à étudier le modèle spatial d'un certain état de santé observé dans un ensemble $(s_1, \dots, s_n)$ de n régions (au lieu de points) définies par exemple par des régions administratives ou des unités de sondages. Dans ce cas, $\mathbf{y}$ peut représenter une agrégation appropriée, par exemple, le nombre de cas observés dans chaque région.

La première étape de la définition d'un modèle spatial dans le cadre bayésien consiste à identifier une distribution de probabilité pour les données observées. Habituellement nous choisissons une distribution de la famille Exponentielle, indexé par un ensemble de paramètres θ en tenant compte de la corrélation spatiale — notez que par souci de simplicité nous abusons légèrement de la notation et indexons le point ou la zone géographique générique en utilisant juste l'indice i , plutôt que l'indicateur s_i , ci-après.

Dans le cas des données géostatistiques, les paramètres sont définis comme un Champ aléatoire Gaussien stationnaire (GF) latent, fonction de certains hyperparamètres ψ associés à une distribution a priori appropriée $p(\psi)$. Ceci équivaut à supposer que θ a une distribution Normale Multivariée avec une moyenne de $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ et une matrice de covariance spatialement structurée Σ , dont l'élément générique est $\Sigma_{ij} = \text{Cov}(\theta_i, \theta_j) = \sigma_c^2 \mathcal{C}(\Delta_{ij})$. Ici σ_c^2 est la variance et pour $i, j = 1, 2, \dots, n$

$$\mathcal{C}(\Delta_{ij}) = \frac{1}{\Gamma(\lambda)2^{\lambda-1}} (\kappa \Delta_{ij})^\lambda K_\lambda(\kappa \Delta_{ij}) \quad (2)$$

est la fonction de covariance spatiale (isotrope) du Matérn¹ (Cressie, 1993) en fonction de la distance euclidienne entre les endroits $\Delta_{ij} = \|s_i - s_j\|$. Ici, K_λ désigne la fonction de Bessel modifiée du second type et de l'ordre $\lambda > 0$, qui mesure le degré de lissage du processus et est généralement maintenu fixe. À l'inverse, $\kappa > 0$ est un paramètre d'échelle lié à la portée r c-à-d. la distance à laquelle la corrélation spatiale devient presque nulle. Généralement, la définition empirique $r = \frac{\sqrt{8\lambda}}{\kappa}$ est utilisé (voir la section 2 dans (Lindgren et al., 2011)), avec

¹ Dans la littérature spatiale, il existe plusieurs fonctions de covariances isotropes utilisées pour modéliser les données géostatistiques. Mais, on se concentre seulement sur les fonctions de covariance de Matérn, qui vont nous permettre de définir l'approche SPDE ultérieurement.

r correspondant à la distance à laquelle la corrélation spatiale est proche de 0,1, pour chaque κ .

Dans le cas des données au niveau de la zone (données régionales), il est possible de reformuler le problème en fonction de la structure de voisinage. En vertu de la propriété de Markov qui stipule que l'élément générique du vecteur de paramètres θ_i est indépendant de tout autre élément, compte tenu de l'ensemble de ses voisins $\mathcal{N}(i)$

$$\theta_i \perp \theta_{-i} \mid \theta_{\mathcal{N}(i)} \quad (3)$$

(θ_{-i} indique tous les éléments de θ sauf le i ème élément), la matrice de précision $\mathbf{Q} = \Sigma^{-1}$ est creuse « sparse », ce qui produit de grands avantages de calcul. En d'autres termes, pour toute paire d'éléments (i, j)

$$\theta_i \perp \theta_j \mid \theta_{-ij} \Leftrightarrow \mathbf{Q}_{ij} = 0 \quad (4)$$

c.-à-d. que la structure non nulle dans la matrice de précision est donnée exclusivement par la structure de voisinage du processus. Ainsi, $\mathbf{Q}_{ij} \neq 0$ seulement si $j \in \{i, \mathcal{N}(i)\}$. Cette spécification est connue sous le nom de Champ aléatoire de Markov Gaussien « Gaussian Markov Random Field » (GMRF ; (Rue et al., 2005)).

3. LA METHODE INLA :

Souvent, dans une analyse statistique, l'intérêt est d'estimer l'effet d'un ensemble de variables explicatives pertinentes sur certaines fonctions (généralement la moyenne) des données observées, tout en tenant compte de la corrélation spatiale ou spatio-temporelle implicite dans le modèle.

Une façon très générale de préciser ce problème consiste à modéliser la moyenne de la i ème unité au moyen d'un prédicteur linéaire additif, défini sur une échelle appropriée (par exemple, échelle logistique pour les données binomiales)

$$\eta_i = \alpha + \sum_{m=1}^M \beta_m x_{mi} + \sum_{l=1}^L f_l(z_{li}) \quad (5)$$

ici α est un scalaire représentant la constante ; les coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)$ quantifient l'effet de certaines variables explicatives $\mathbf{x} = (x_1, \dots, x_M)$ sur la variable à expliquer, et $\mathbf{f} = \{f_1(\cdot), \dots, f_L(\cdot)\}$ est un ensemble de fonctions définies en termes d'un ensemble d'autres variables explicatives $\mathbf{z} = (z_1, \dots, z_L)$. En faisant varier la forme des fonctions $f_l(\cdot)$, cette formulation peut s'adapter à un large éventail de modèles, allant de la régression standard et hiérarchique aux modèles spatiaux et spatio-temporels (Rue et al., 2009).

Compte tenu de la spécification de (2), le vecteur des paramètres est représenté par $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\beta}, \mathbf{f}\}$. Conformément à la discussion de la section 2, nous pouvons supposer une distribution a priori GMRF pour $\boldsymbol{\theta}$, avec une moyenne $\mathbf{0}$ et une matrice de précision $\mathbf{Q}$. En outre, en raison des relations d'indépendance conditionnelle qu'implique le GMRF, le vecteur des K hyperparamètres $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)$ va avoir une dimension de l'ordre de $(1 + L)$.

Comme toute méthode d'inférence bayésienne, l'objectif est d'estimer les distributions marginales a posteriori pour chaque composante du vecteur des paramètres

$$p(\theta_i | \mathbf{y}) = \int p(\theta_i, \boldsymbol{\psi} | \mathbf{y}) d\boldsymbol{\psi} = \int p(\theta_i | \boldsymbol{\psi}, \mathbf{y}) p(\boldsymbol{\psi} | \mathbf{y}) d\boldsymbol{\psi} \quad (6)$$

et (éventuellement) la distribution a posteriori pour chaque élément du vecteur des hyperparamètres

$$p(\psi_k|\mathbf{y}) = \int p(\boldsymbol{\psi}|\mathbf{y})d\psi_{-k} \quad (7)$$

Alors, en général, il faut :

- i. Calculer $p(\boldsymbol{\psi}|\mathbf{y})$ qui nous permet de déterminer les distributions marginales $p(\psi_k|\mathbf{y})$ pour chaque élément.
- ii. Calculer $p(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ qui sera utilisée pour calculer la distribution a posteriori marginale $p(\theta_i|\mathbf{y})$

La méthode INLA est basée essentiellement sur l'utilisation de l'approximation de Laplace pour essayer d'effectuer les deux tâches (i) et (ii).

La première tâche (i) consiste à calculer une approximation de la distribution marginale a posteriori des hyperparamètres. On a :

$$p(\boldsymbol{\psi}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}{p(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})} \propto \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})} \quad (8)$$

En utilisant l'approximation de Laplace $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ de $p(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$, on trouve l'approximation $\tilde{p}(\boldsymbol{\psi}|\mathbf{y})$ donnée par :

$$\tilde{p}(\boldsymbol{\psi}|\mathbf{y}) \propto \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})} \quad (9)$$

où $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ est l'approximation Gaussienne ((Rue et al., 2009), Section 2.2) de $p(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ et $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ est son mode pour un $\boldsymbol{\psi}$ donnée. Tel que décrit dans Rue et al. (2009), l'équation (3) est équivalente à l'approximation de Laplace d'une densité a posteriori marginale proposée dans (Tierney et al., 1986).

La deuxième tâche (ii) est légèrement plus complexe, car en général il y aura plus d'éléments dans $\boldsymbol{\theta}$ que dans $\boldsymbol{\psi}$ et donc ce calcul est plus coûteux. Une alternative plus facile est d'approximer les distributions conditionnelles a posteriori $p(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ directement comme les distributions marginales de $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$, c.-à-d. en utilisant une distribution normale, où la matrice de précision est basée sur la décomposition de Cholesky de la matrice de précision $\mathbf{Q}$ (Rue et Martino, 2007). Bien que cela soit très rapide, l'approximation n'est généralement pas très bonne. Alternativement, il est possible de réécrire le vecteur de paramètres comme $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$, alors la distribution s'écrit :

$$p(\theta_i|\boldsymbol{\psi}, \mathbf{y}) = \frac{p((\theta_i, \boldsymbol{\theta}_{-i})|\boldsymbol{\psi}, \mathbf{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})} \propto \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}{p(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})} \quad (10)$$

puis utiliser une deuxième approximation de Laplace $\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})$ de $p(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})$ pour trouver :

$$\tilde{p}_{LA}(\theta_i|\boldsymbol{\psi}, \mathbf{y}) \propto \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})}{\tilde{p}(\boldsymbol{\theta}_{-i}|\theta_i, \boldsymbol{\psi}, \mathbf{y})} \Big|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}_{-i}^*(\theta_i, \boldsymbol{\psi})} \quad (11)$$

Où $\theta_{-i}^*(\theta_i, \psi)$ est le mode de $p(\theta_{-i} | \theta_i, \psi, \mathbf{y})$. Étant donné que les variables aléatoires $(\theta_{-i} | \theta_i, \psi, \mathbf{y})$ sont en général raisonnablement normales, l'approximation fournie par (4) fonctionne généralement très bien.

Cette stratégie, cependant, peut être très coûteuse en termes de calcul. Par conséquent, l'algorithme le plus efficace est l'algorithme de « L'approximation de Laplace Simplifié », qui est basé sur un développement en séries de Taylor de l'approximation de Laplace $\tilde{p}(\theta_i | \psi, \mathbf{y})$ dans l'équation (4). Il s'agit habituellement de « corriger » en ajoutant un terme de lissage (par exemple., Spline) pour augmenter l'ajustement à la distribution requise. La précision de cette approximation est suffisante dans de nombreux cas et le temps nécessaire pour les calculs est beaucoup plus court.

Sur le plan opérationnel, INLA procède d'abord par l'exploration de la distribution marginale a posteriori conjointe pour les hyperparamètres $\tilde{p}(\boldsymbol{\theta} | \psi, \mathbf{y})$ afin de localiser le mode ; une recherche en grille est ensuite effectuée et produit un ensemble de $\mathcal{G}$ de points « pertinentes » $\{\psi^*\}$ avec un ensemble de poids $\{w_{\psi^*}\}$ correspondants, pour donner l'approximation de cette distribution. Chaque distribution marginale a posteriori $\tilde{p}(\psi^* | \mathbf{y})$ peut être obtenu en utilisant l'interpolation basée sur les valeurs calculées et en corrigeant l'asymétrie (probable). Pour chaque ψ^* , les distributions a posteriori conditionnelles $\tilde{p}(\theta_i | \psi, \mathbf{y})$ sont ensuite évalués sur une grille de valeurs sélectionnées pour θ_i et les distributions a posteriori marginales $\tilde{p}(\theta_i | \mathbf{y})$ sont obtenus par intégration numérique

$$\tilde{p}(\theta_i | \mathbf{y}) \approx \sum_{\psi^* \in \mathcal{G}} \tilde{p}(\theta_i | \psi^*, \mathbf{y}) \tilde{p}(\psi^* | \mathbf{y}) w_{\psi^*} \quad (12)$$

4. L'APPROCHE SPDE :

Les données ponctuelles peuvent être traitées avec l'approche des équations différentielles partielles stochastiques (SPDE) proposée par (Lindgren et al., 2011). Il s'agit de représenter un processus spatial continu, par exemple, un Champs aléatoire Gaussien (GF) avec une fonction de covariance de Matérn définie en (1), comme un processus aléatoire spatial discret (par exemple un GMRF). Cela produit à son tour d'importants avantages de calcul. En fait, les champs aléatoires Gaussien spatiaux sont affectés par « le problème de BIG n » (Banerjee et al., 2015; Lasinio et al., 2013), qui est attribuable aux coûts de calcul de l'ordre $\mathcal{O}(n^3)$ pour effectuer des opérations d'algèbre matricielle avec n matrices de covariance dense (dont la dimension est donnée par le nombre d'observations à tous les endroits spatiaux).

En revanche, comme on l'a vu à la section 2, les GMRF se caractérisent par des matrices de précision creuses, ce qui permet de mettre en œuvre des méthodes numériques efficaces sur le plan du calcul, en particulier pour la factorisation rapide des matrices (Rue et al., 2005). Pour un modèle GMRF en $\mathbb{R}^2$, le coût de calcul est typiquement de $\mathcal{O}(n^{3/2})$, ce qui est une amélioration significative par rapport à $\mathcal{O}(n^3)$ des GF. De plus, l'inférence bayésienne impliquant des GMRF spatiaux peut être effectuée en utilisant l'approche INLA introduite dans la section 3.

Dans cette section, nous introduisant brièvement les bases de l'approche SPDE et nous nous référons à (Lindgren et al., 2011) pour une description complète et les preuves des résultats. Les applications du SPDE pour les données géostatistiques se trouvent dans (Bolin, 2012; Cameletti et al., 2012; Simpson et al., 2012, 2016).

Envisageons un cadre simple pour les données géostatistiques dans lequel l'observation y_i est modélisée comme

$$y_i \sim \text{Normal}(\eta_i, \sigma_e^2) \quad (13)$$

où σ_e^2 est la variance de l'erreur de mesure centré e_i qui est supposé être indépendante des e_j pour $i \neq j$. La moyenne est donnée par :

$$\eta_i = \alpha + \sum_{m=1}^M \beta_m x_{mi} + \xi_i \quad (14)$$

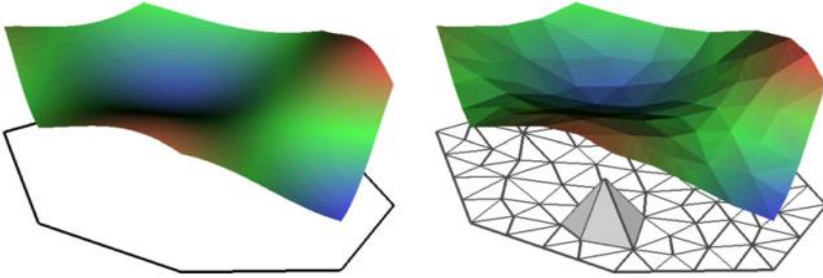
où ξ_i est la i ème réalisation du champs aléatoire Gaussien latent $\xi(s)$ avec une fonction de covariance spatiale définie en (1). Dans la littérature géostatistique, le terme $\alpha + \sum_{m=1}^M \beta_m x_{mi}$ est souvent appelé, composante à grande échelle, tandis que la variance d'erreur de mesure σ_e^2 est appelée l'effet de pépite (Cressie, 1993). En comparaison avec le prédicteur linéaire introduit dans (2), la fonction $f_i(\cdot)$ est représentée par le terme spatialement structuré ξ_i . L'idée principale de l'approche SPDE consiste à représenter un champs aléatoire Gaussien continu de Matérn $\xi(s)$ comme un GMRF discret au moyen d'une représentation de fonctions de base définie sur une triangulation du domaine $\mathcal{D}$.

$$\xi(s) = \sum_{g=1}^G \varphi_g(s) \tilde{\xi}_g \quad (15)$$

Ici G est le nombre total de vertex dans la triangulation, $\{\varphi_g\}$ est l'ensemble des fonctions de base et $\{\tilde{\xi}_g\}$ sont des poids Normalement distribué. Pour des raisons de calcul, comme expliqué dans (Lindgren et al., 2011), les fonctions de base sont choisies pour être linéaires par tranches sur chaque triangle, c.-à-d. φ_g est 1 au vertex g et 0 par ailleurs. L'utilisation de la notation formelle $\xi(s)$ dans l'équation (15) est justifié par le fait que les SPDE fournissent une représentation de l'ensemble du processus spatial (défini pour n'importe quel point s) qui varie d'une manière continue dans le domaine $\mathcal{D}$ considéré.

Une illustration de l'approche SPDE est donnée à la figure 1 qui affiche un champ spatial aléatoire continu et la représentation en éléments finis correspondante avec des fonctions de base linéaire par tranches sur une triangulation « mesh ». (Lindgren et al., 2011) montrent que le vecteur des poids de base $\tilde{\xi} = (\tilde{\xi}_1, \dots, \tilde{\xi}_N)'$ est un GMRF avec une matrice de précision creuse $\mathbf{Q}_{\tilde{\xi}}$ en fonction du paramètre de la fonction de covariance de Matérn κ et la variance σ_C^2 , pour un $\alpha = 1, 2, 3, \dots$ où $\alpha = \lambda + 1$ (il faut rappeler que λ est le paramètre de lissage).

Figure N°1 : Exemple d'un champ aléatoire continu et la représentation en fonctions de base correspondante selon l'équation 15



Source : (Simpson et al., 2016)

Compte tenu de la représentation GF fournie par la méthode SPDE, nous pouvons réécrire le prédicteur linéaire en (10) comme

$$\eta_i = \alpha + \sum_{m=1}^M \beta_m x_{mi} + \sum_{g=1}^G \tilde{A}_{ig} \tilde{\xi} \quad (16)$$

Où $\tilde{A}$ est la matrice creuse de taille $n \times G$ qui retrace le GMRF $\tilde{\xi}$ à partir des n emplacements d'observation aux nœuds de triangulation G .

5. APPLICATION SUR LES DONNEES DU CANCER DU POUMON :

5.1. Les données :

Pour illustrer les concepts cités, on va présenter une application sur la mortalité par Cancer du poumon en Algérie. Les données utilisées sont la mortalité par cancer du poumon par wilaya (découpage administrative en 48 wilayas) durant la période 2012-2016 comme une variable à expliquer (y) et une estimation sur le nombre de cigarettes consommées par wilaya (en millions d'unités) sur la même période. Les données proviennent des registres du cancer régis par ministère de la santé, de la population et de la réforme hospitalière en Algérie.

5.2. Le modèle des données territoriales

5.2.1. Spécification du modèle :

En premier lieu, on va utiliser les données agrégées (par wilaya) pour définir un modèle bayésien hiérarchique pour modéliser la mortalité par cancer du poumon en prenant en compte la variation spatiale.

Pour la i -ème wilaya, la mortalité par cancer du poumon, est modéliser comme :

$$y_i \sim \text{Poisson}(\lambda_i) \quad (17)$$

où la moyenne λ_i est définie : $\lambda_i = \rho_i E_i$, où ρ_i définie le risque spécifique pour chaque zone et E_i est la mortalité espérée pour chaque zone, calculée en utilisant la méthode de standardisation indirecte. Le prédicteur linéaire est alors défini sur l'échelle logarithmique :

$$\eta_i = \log(\rho_i) = \alpha + \beta x + v_i + v_i \quad (18)$$

α est la constante qui défini la moyenne de la mortalité sur toute les 48 wilayas, $v_i = f_1(i)$ et $v_i = f_2(i)$ sont deux effets spécifiques pour chaque wilaya.

v_i est un résidu spatialement structuré et qui est modélisé par un modèle BYM (Besag-York-Mollie, Besag et al. 1991) avec une structure autorégressive conditionnelle intrinsèque (ICAR)

$$v_i | v_{j \neq i} \sim N(m_i, s_i^2) \tag{19}$$
$$m_i = \frac{\sum_{j \in \mathcal{N}(i)} v_j}{\#\mathcal{N}(i)} \quad \text{et} \quad s_i^2 = \frac{\sigma_v^2}{\#\mathcal{N}(i)}$$

où $\#\mathcal{N}(i)$ est le nombre de zones dans le voisinage de la zone i , le voisinage de la zone est défini par le nombre de zones ayant une frontière commune avec i (Banerjee et al., 2015). Enfin, v_i représente les résidus non structurés et qui soient supposés suivre une distribution a priori non échangeable Gaussienne $v_i \sim N(0, \sigma_v^2)$.

Après plusieurs choix, on a opté pour prendre des distributions a priori peu informatives pour les hyperparamètres, $\sigma_v^2 \sim \log \text{Gamma}(1, 0.005)$ et $\sigma_v^2 \sim \log \text{Gamma}(0.1, 0.5)$. Généralement, dans ce contexte, il est préférable de modéliser $\zeta_i = v_i + v_i$. Le modèle est ainsi estimé par la méthode INLA en prenant le vecteur de paramètres $\theta = \{\alpha, \beta, \zeta, v\}$ et le vecteur des hyperparamètres contient les précisions $\psi = \{\tau_v^2 = \frac{1}{\sigma_v^2}, \tau_v^2 = \frac{1}{\sigma_v^2}\}$.

5.2.2. Résultats du modèle

Dans le tableau 1 on trouve la comparaison du modèle choisi avec le modèle avec les hyperparamètres $\sigma_v^2 \sim \log \text{Gamma}(0.1, 0.01)$ et $\sigma_v^2 \sim \log \text{Gamma}(0.1, 0.1)$. Pour comparer on peut utiliser le critère d'information de la déviance (DIC), le critère d'information d'Akaike-Watanabe ($WAIC$) qu'ils doivent avoir des valeurs faibles pour le meilleur modèle et le nombre effectif de variables dans le modèle (P_D) qu'il faut maximiser.

D'après le tableau 1, le modèle choisi a la (DIC) et ($WAIC$) la plus faible et (P_D) le plus élevé parmi tous les modèles proposés.

Dans toute analyse bayésienne on peut définir les ordonnées prédictives conditionnelles (CPO). CPO est la distribution de probabilité d'une seule valeur observée basée sur l'estimation de toutes les autres. Les valeurs de CPO représentent la probabilité d'une valeur attendue, ce qui implique que les grandes valeurs sont privilégiées.

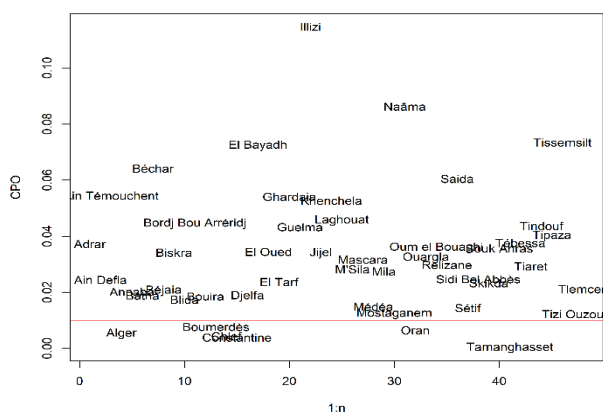
Table N° 1 : Comparaison entre les deux modèles

	Modèle choisi	Modèle proposé
DIC	339.7792	339.822
P_D	29.89674	28.05
$WAIC$	337.0815	338.8572

Source : élaboré par nos soins

La figure 2, présente les différentes valeurs de la CPO pour chaque wilaya. On remarque que 6 wilayas présentent des résultats inattendus (des valeurs de $CPO < 0.01$). Les wilayas concernées sont : Alger (la wilaya avec la plus grande mortalité), Oran et Constantine (les capitaux de l'est et de l'ouest et qui ont des valeurs assez élevées de mortalité) et Boumerdès, Chlef et Tamnaghasset.

Figure N°2 : Les valeurs de CPO



Source : élaboré par nos soins

Après le diagnostic du modèle, on présente les résultats du modèle. Les informations sur les effets fixes du modèle choisi peuvent être résumées dans le tableau 2.

Table N° 2 : Comparaison entre les deux modèles

	moyenne	écart – type	2.5%	50%	97.5%
α	-0.03091	0.06159	-0.1516	-0.0311	0.0907
β	0.004052	0.007927	-0.0117	0.0041	0.0195

Source : élaboré par nos soins

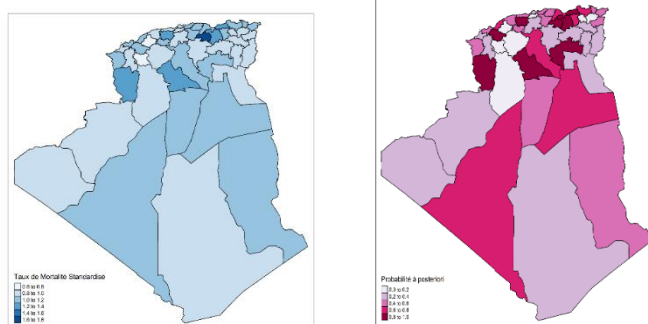
La constante exponentielle α est la valeur du taux de mortalité moyen sur toute la région du pays (3.04%) avec un intervalle de crédibilité de 95% qui varie entre 0.14% et 9.5%. Le coefficient β peut être interprété en tant qu'un facteur de risque en exponentielle $\exp(\beta) = 1.00406$, une augmentation d'une unité (100 Millions) en consommation de cigarettes augmente le risque de mortalité par cancer du poumon par 0.406%.

La figure 3(a) montre la moyenne a posteriori du risque relatif de mortalité par cancer du poumon pour chaque wilaya $\zeta = \exp(\zeta)$ comparé à tout le pays. On s'intéresse aussi à l'excès du risque, en calculant la probabilité a posteriori que le risque dépasse 1, $P(\zeta_i > 1 \mid \mathbf{y})$. La probabilité est présentée dans la figure 3(b).

5.3. Le modèle de données ponctuelles (Géostatistiques)

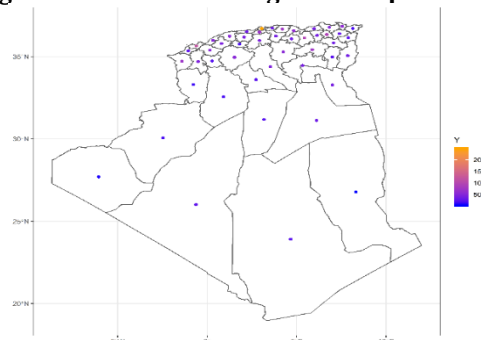
Pour utiliser un modèle géostatistique, on va supposer que chaque wilaya est représentée par un seul point (le centroïde de la surface de la wilaya) et contenant le même nombre de mortalité par cancer du poumon déjà utilisé dans le modèle régional. Les centroïdes et les données correspondantes sont présentés dans la figure 4.

Figure N°3 : Le risque relatif de chaque wilaya et la probabilité a posteriori
 (a) Distribution du risque relative de chaque wilaya $\zeta = v_i + v_i$ (b) distribution de la probabilité a posteriori de chaque wilaya $P(\zeta_i > 1 | y)$



Source : élaboré par nos soins

Figure N°4 : Les données géostatistiques utilisées



Source : élaboré par nos soins

5.3.1. Spécification du modèle :

L'une des objectives de la modélisation géostatistique est la prédiction des valeurs du phénomène (mortalité par cancer du poumon) à des endroits non observés (où le reste des

endroits dans le domaine d'étude $\mathcal{D}$). On va illustrer cette prédiction en utilisant l'approche SPDE déjà présenté en section 4.

Le modèle utilisé est une extension des équations 13 et 14 pour une distribution de Poisson. La forme matricielle de ces équations est donnée par :

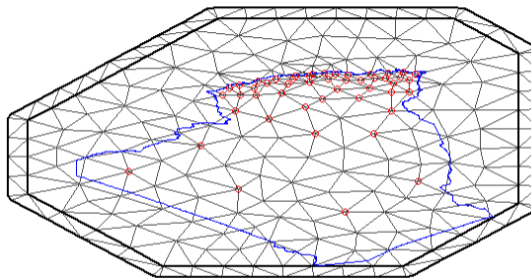
$$\begin{aligned} \mathbf{y} &\sim \text{Poisson}(\boldsymbol{\rho}) \\ \boldsymbol{\eta} = \log(\boldsymbol{\rho}) &= \mathbf{1}\alpha + \tilde{\mathbf{A}} \boldsymbol{\xi} \end{aligned} \quad (21)$$

où $\mathbf{y} = (y_1, \dots, y_n)'$ est le vecteur des observations. Suivant la même notation introduite pour un modèle INLA, le vecteur des paramètres est $\boldsymbol{\theta} = \{\boldsymbol{\xi}, \alpha\}$ et le vecteur des hyperparamètres est $\boldsymbol{\psi} = \{\kappa, \sigma_c^2\}$.

5.3.2. Résultats du modèle :

La première étape avant l'estimation du modèle consiste à définir la triangulation « mesh » appropriée. La triangulation utilisée dans pour estimer u modèle SPDE est une triangulation de Delaunay contraint. Cette triangulation est construite d'une manière à que chaque vertex est superposé avec une location des données et puis le reste sont ajoutées d'une manière à satisfaire les contraintes de qualité (Lindgren et al. 2011). Pour satisfaire ces contraintes la triangulation utilisée contient $G = 248$ vertex (voir figure 5).

Figure N°5 : La triangulation « mesh »
Constrained refined Delaunay triangulation



Source : élaboré par nos soins

Les résultats d'estimation pour les effets fixes (α , la portée $\log \tau$ et $\log \kappa$) sont présentés dans le tableau 3. Les deux derniers paramètres sont utilisés pour estimer les valeurs de la portée et la variance minimale. La moyenne de la portée est d'environ 70 Km, ce qui indique que l'autocorrélation spatiale est vraiment faible pour ces données, car le domaine d'étude est vraiment étendu (la distance maximale entre deux endroits est de plus de 1590 Km).

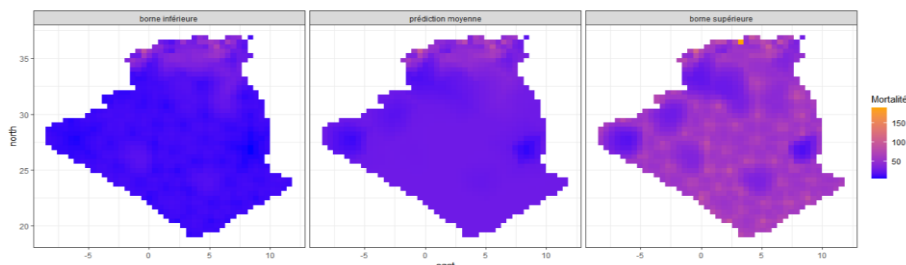
Table N° 3 : Les statistiques des effets fixes et aléatoires pour le modèle SPDE

	moyenne	écart – type	2.5%	50%	97.5%
α	3.29	0.127	3.02	3.29	3.53
$\log \tau$	-2.07	0.3	-2.704	-2.059	-1.52
$\log \kappa$	0.992	0.239	0.554	0.98	1.493
σ^2	0.713	0.178	0.437	0.686	1.133
ρ	1.077	0.252	0.637	1.061	1.618

Source : élaboré par nos soins

On peut se concentrer maintenant sur la prédiction de la mortalité sur l'ensemble du pays. La prédiction (Krigage) est basée sur la triangulation utilisée et le modèle choisi. La prédiction sur tous les autres endroits du domaine d'étude est montrée en figure 6.

Figure N°6 : La mortalité moyenne prédite et la borne supérieure et inférieure de l'intervalle de crédibilité pour la mortalité



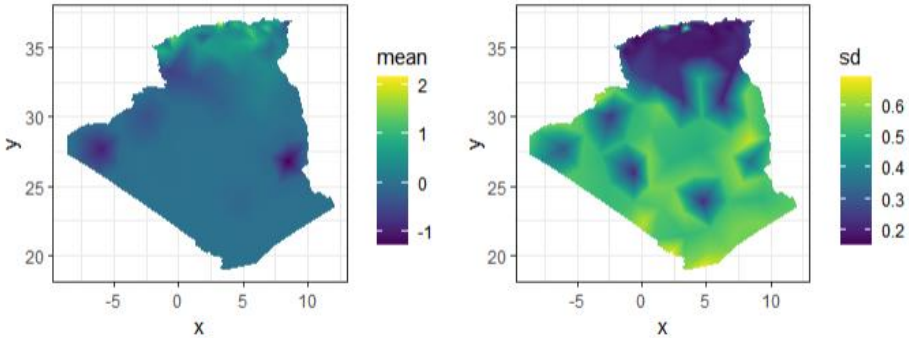
Source : élaboré par nos soins

On remarque qu'à cause de la faible corrélation entre les endroits, une estimation moyenne (parfois faible) est constatée sur tout le territoire national. Sur le nord du pays, on remarque quelques valeurs élevées et cela est due que le nord contenait plus de mortalité par rapport au sud. L'intervalle de confiance de la mortalité est souvent modéré.

Dernièrement, on peut utiliser le modèle géostatistique pour générer le processus générateur de la mortalité. Il est possible de projeter les champs aléatoires sur les vertex de chaque triangle dans le « mesh » sur une grille contenant tous les endroits du pays. Cette projection peut servir à mieux comprendre la dispersion du phénomène sur tous les pays.

On remarque que la mortalité par Cancer est plus fréquente à l'Est et l'Ouest du pays, moins fréquente au sud et pratiquement non fréquente aux extrêmes est et Ouest du sud du pays. On remarque que l'écart-type est faible et constant sur l'ensemble du Nord du pays (sauf Alger et environs) et élevé au Sud, ce qui est expliquée par la distance élevée entre les endroits utilisées dans le modèles, ce qui entraine une variation plus au moins considérable dans des endroits éloignés des emplacements utilisés dans le modèle.

Figure N°7 : Moyenne et écart-type a posteriori du champ aléatoire



Source : élaboré par nos soins

5.4. Résultats

Une analyse approfondie des résultats des deux modèles utilisés, peut nous mener à établir quelques conclusions. Les grandes villes du pays (Alger, Oran, Constantine, ...) montrent des signes très inquiétants sur la situation de la mortalité par Cancer du poumon. Les autorités sont conseillées d'intervenir pour trouver la cause du problème, qui est probablement dû aux déplacements des malades entre les différentes wilayas.

Les deux modèles suggèrent que la mortalité est essentiellement élevée dans les wilayas avec une grande population et surtout dans les wilayas qui disposent des Centres Anti Cancer. L'hypothèse de l'existence d'un phénomène environnemental influant sur la mortalité par Cancer du poumon en Algérie est infirmé pour les deux modèles, mais cela n'empêche pas l'utilisation d'autres variables explicatives environnementale (géostatistiques) pour avoir une idée plus claire.

6. DISCUSSIONS :

Dans cet article, on a présenté une application sur l'utilisation de la méthode INLA pour un modèle spatial. Malgré que ces modèles soient largement utilisés, surtout en épidémiologie, mais leur complexité reste un problème essentiel pour lors de l'implémentation. L'approche INLA fournit généralement des estimations fiables avec un temps de calcul considérablement faible comparé aux estimations basées sur les MCMC.

L'utilisation de la méthode SPDE pour des données géostatistiques, qui soient généralement des données de taille massive, permet aussi de gagner en temps de calcul, mais augmente la complexité du modèle considérablement, surtout avec le nombre de paramètres à estimer qui soit plus au moins élevé avec l'utilisation de la matrice de covariance de Matérn. INLA et SPDE sont souvent utilisés dans des problèmes de changement de support, comme présenté dans cet article, où on a passé des données au niveau de la wilaya vers des données ponctuelles.

L'application dans cet article peut inciter à l'utilisation des approches INLA/SPDE pour répondre à une large gamme de problèmes, surtout avec l'utilisation de base de données géostatistique de taille plus importante pour des problèmes de santé publique ou d'épidémiologie.

BIBLIOGRAPHIE:

1. Baio, G. (2012). *Bayesian Methods in Health Economics*. Boca Raton: CRC Chapman and Hall.
2. Banerjee, S., Carlin, B., P., & Gelfand, A., E. (2015). *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed.). Boca Raton: CRC Press.
3. Bernardo, J., M., & Smith, A., F. (2000). *Bayesian Theory*. Chichester: John Wiley & Sons.
4. Bolin, D. (2012). *Models and methods for random fields in spatial statistics with computational efficiency from Markov properties* [PhD dissertation]. Lund University.
5. Brooks, S., Gelman, A., Jones, G., L., & Meng, X.-L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. Boca Raton: Chapman & Hall / CRC Press.
6. Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2012). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Advances in Statistical Analysis*, 97(2), 109–131. doi: <https://doi.org/10.1007/s10182-012-0196-3>
7. Clark, J., S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, 8(1), 2–14. doi: <https://doi.org/10.1111/j.1461-0248.2004.00702.x>.
8. Clark, J. S., & Gelfand, A. E. (Eds.). (2006). *Hierarchical Modelling for the Environmental Sciences : Statistical Methods and Applications*. Oxford: Oxford University Press.
9. Cressie, N. (1993). *Statistics for Spatial Data*. Hoboken: John Wiley & Sons.
10. Diggle, P., J., & Riberio, P. J. Jr. (2007). *Model-based Geostatistics*. New York: Springer Science & Business Media.
11. Dunson, D., B. (2001). Commentary: Practical Advantages of Bayesian Analysis of Epidemiologic Data. *American Journal of Epidemiology*, 153(12), 1222–1226. doi: <https://doi.org/10.1093/aje/153.12.1222>
12. Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (2010). *Handbook of Spatial Statistics*. Boca Raton: CRC Press.
13. Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, 35(3), 765–775. doi: <https://doi.org/10.1093/ije/dyi312>
14. Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Chichester: John Wiley & Sons.
15. Jewell, C. P., Kypraios, T., Neal, P., & Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(4), 465–496. doi: <https://doi.org/10.1214/09-BA417>
16. Lasinio, J. G., Mastrantonio, G., & Pollice, A. (2013). Discussing the “big n problem”. *Statistical Methods & Applications*, 22, 97–112. doi: <https://doi.org/10.1007/s10260-012-0207-2>

17. Lawson, A. (2019). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology* (3rd ed.). Boca Raton: CRC Press.
18. Lindgren, F., & Rue, H. (2015). Bayesian Spatial Modeling with R-INLA. *Journal of Statistical Software*, 63(19), 1–25. doi: <https://doi.org/10.18637/jss.v063.i19>
19. Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4), 423–498. doi: <https://doi.org/10.1111/j.1467-9868.2011.00777.x>.
20. Lindley, D., V. (2006). *Understanding Uncertainty*. Hoboken: John Wiley & Sons.
21. Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton: CRC Press.
22. Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049–3067. doi: <https://doi.org/10.1002/sim.3680>.
23. Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd ed.). New York: Springer Science & Business Media.
24. Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: CRC Press.
25. Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392. doi: <https://doi.org/10.1111/j.1467-9868.2008.00700.x>
26. Scott, M. Berry, Carlin, B., P., Lee, J., Jack, & Muller, P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton: CRC Chapman and Hall.
27. Simpson, D., Illian, J., Lindgren, F., Sørbye, S., H., & Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1), 49–70. doi: <https://doi.org/10.1093/biomet/asv064>
28. Simpson, D., Lindgren, F., & Rue, H. (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1, 16–29. doi: <https://doi.org/10.1016/j.spasta.2012.02.003>
29. Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), 82–86. doi: <https://doi.org/10.1080/01621459.1986.10478240>
30. Wikle, C. K. (2007). Hierarchical Models in Environmental Science. *International Statistical Review*, 71(2), 181–199. doi: <https://doi.org/10.1111/j.1751-5823.2003.tb00192.x>