UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Representativeness of Lyme surveillance in Belgium

Terence Achuo Tem

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
dr. Pietro COLETTI

SUPERVISOR :
Dr. Tinne LERNOUT

2018
2019

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Representativeness of Lyme surveillance in Belgium

Terence Achuo Tem

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Pietro COLETTI

SUPERVISOR :

Dr. Tinne LERNOUT

Abstract

Background: Epidemiological surveillance is the systematic collection, analysis, interpretation and timely publication of health data for planning, utilization and evaluation of public health programs. In Belgium, a surveillance system for communicable diseases was established over 30 years ago based on a voluntary sentinel network of microbiological laboratories (SNL) and sentinel network of general practitioners (SGP) both coordinated by the scientific institute of public health (Sciensano). The aim of this study is to evaluate the aspect of representativeness of the SNL and the SGP on the occurrence of Lyme borreliosis (LB) amongst other infectious diseases that are under surveillance in Belgium.

Method: Test coverage of the SNL was estimated using the ratio of reimbursed tests performed by sentinel laboratories to the total number of tests performed between 2010 and 2015 for positive serology tests of LB.High risk areas for LB was estimated at the provincial level using cumulative incidence based on the positive serology tests on LB between 2008 and 2016 performed by participating laboratories of the network. The Integrated Nested Laplace Approximation (INLA) approach using the Besag York and Mollie (BYM) specification was used to estimate the relative risks for LB at the district level taking the spatial structure of the data from SGP into account. Correlation tests were conducted to find the relationship between the occurrence of erythema migrans reported by SGP and the number of tick bites from the whole of Belgium.

Results: It was found that test coverage was greater than 50% at the national and regional levels but there was much variability at the provincial level. Some provinces had low test coverage while others had very high test coverage. The high risks areas of lyme at the district level estimated by the BYM model for all data sources didn't have much difference. All the disease mapping models showed similar patterns of risk of LB. There was a positive correlation between the risk of LB and risk of tick bite.

Conclusion: The representativeness of the sentinel network system for surveillance on the occurrence of LB based on test coverage results suggests that the SNL is capable to describe trend and can monitor changes of LB occurrence over time at the national and regional level. Areas at high risk of LB are sufficiently covered by the network. The number of tick bites was used as a proxy to estimate the risk of LB in Belgium. The representativeness of the SGP on the occurrence of EM as estimated by the correlation tests shows that SGP is representative at the national and district levels. This method of reprensentativeness was not able to identify trends in changes in LB.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview of Epidemiological surveillance of infectious diseases in Belgium

Epidemiological surveillance can be defined as the systematic collection, analysis, interpretation and timely publication of health data for planning, utilization and evaluation of public health programs. A surveillance cycle in public health is being completed by the application of these data to disease prevention and health promotion programs [1]. Based on the quality as well as usefulness and cost, established surveillance systems should be reviewed periodically. The quality of a surveillance system can be assessed by reviewing the following seven attributes; sensitivity, specificity, representativeness, timeliness, simplicity, flexibility and acceptability [2]. The main interest of this research is the attribute on representativeness. A representative surveillance system accurately observes both the occurrence of a health event over time and the distribution of that event in the population by person and place at any point in time. Representativeness can be measured by comparing surveillance data covering part of a population to a sample assumed to be complete. A surveillance system that collects reports on essentially all occurrences of a health event is said to be representative by definition and further assessment of this attribute is not necessary [3].

Even though there is a rising contribution of noncommunicable diseases to the global burden of disease, communicable diseases still continue to threaten population health especially in developed countries. In Belgium there are two main surveillance systems that contribute to routine surveillance of Lyme borreliosis (LB): a network of sentinel laboratories (SNL) that performs laboratory surveillance by weekly reporting the number of positive serological tests for *Borrelia burgdorferi* s.l. and a sentinel network of general practitioners (SGP) that reports the weekly number of patients consulting with an asymptomatic tick bite or an erythema migrans (EM), in prospective studies. These two surveillance systems have been existing for more than 30 years now [4]. A previous report on sensitivity and representativeness of the surveillance system on infectious diseases based on test coverage suggests that the SNL is capable to describe trends and to monitor changes in the infectious disease pathogens studied, both at national and regional levels. This previous SNL representativeness study included *B. burgdorferi* s.l., for the period 2007-2012. The results showed a stable coverage of the network over time for the three regions of Belgium, but with a great coverage variability between provinces; the network seems less likely to capture local changes over time in some provinces such as Namur, Walloon Brabant, Liege and Limburg [5]. Three of these provinces include areas with a higher risk of tick bites and LB. Also, the Belgian SGP has proven to be a reliable surveillance system for a wide range of health-related data, for infectious and non-infectious diseases [5,6]. However, unlike for other infectious pathogens, the occurrence of LB is very region dependant, since it is influenced by landscape, presence of ticks, presence of animal hosts for ticks and human behaviour. Thus, the geographical representativeness of the surveillance might be different and insufficient.

1.2 Lyme Borreliosis

LB is the most common vector-borne infectious disease in North America and some European countries with moderate climates. LB is a multistage disease caused by a group of related spirochaetes of the *B. burgdorferi* sensu lato species complex, which are transmitted by ticks.The tick *Ixodes ricinus* is the main vector of LB in Europe. The disease is transmitted to humans through injection of tick saliva during blood feeding of an infected tick [8]. Early clinical manifestation is a localized infection known as erythema migrans (EM), a red expanding skin lesion at the site of the tick bite. If it is untreated, disseminated infection can occur in an early or late stage of the disease. This can cause more severe manifestations of which multiple EM, Lyme neuroborreliosis (LNB) and Lyme arthritis (LA) are the most frequent ones. Other less frequent early or late manifestations are borrelial lymphocytoma, Lyme carditis, ocular manifestations and acrodermatitis chronica atrophicans (ACA) [9, 10]. EM is the most common of the various objective clinical presentations in Europe. About 89% in one case series of patients with LB had EM by itself [11]. Usually, typical EM is sufficiently distinctive to allow a clinical diagnosis in the absence of a supporting laboratory test. Laboratory diagnosis for non-erythema migrans presentations of LB are necessary [8]. Laboratory diagnosis of LB in Europe follows a is a two step procedure. The first stage is a sensitive enzyme linked immunosorbent assay (ELISA). If ELISA is positive then the second step is conducted; a separate IgM and IgG immunoblots on the same serum sample is done. This two step procedure was adopted because of the complexity of the antigenic composition of *B. bugdorferi* sl and the temporal appearance of antibodies to different antigens at successive time intervals after infection.This

means that the developement of a serological test with high sensitivity and specificity is a big challenge [10]. ELISA alone cannot detect the presence of LB because of it's very low specificity. The disease burden of LB are influenced by many factors. Some of the most important factors include; factors that influence the human-tick encounter such as changes in human recreational behaviour including changes due to altered climatic conditions, factors influencing vector and reservoir animal abundance such as climate change which directly affects the survival and development of ticks and has an indirect effect on tick abundance and pathogen transmission and finally, factors that affect society's adaptive capability to change such as presence of surveillance networks and monitoring [9].

1.3 Research Question

This thesis is based on the following research questions

- To what geographical level and to what extent (general trends analysis only) are the results of the SNL representative for the occurrence of LB in Belgium. Are areas at risk for Lyme sufficiently covered by the network?

- Are the results of the SGP representative for the occurrence of an EM in Belgium? Can the results of the sentinel GPs be extrapolated in the same way as for other (infectious) diseases or conditions, or is the geographical distribution of Lyme disease too specific.

2 Methodology

2.1 Data Description

Four different data sources are provided to answer the above mentioned researched questions.

2.1.1 Reimbursement tests for *B. burgdorferi* s.l. serology

The data were obtained from the Belgian National Institute for Health and Disability Insurance (INAMI-RIZIV) reported yearly from 2010 to 2015. INAMI-RIZIV only provided the aggregated number of tests conducted by district (containing data from one or more laboratories) not the number of tests performed for each laboratory due to privacy reasons. The variables included in the data sets are: the type of anti-borrelia tests conducted(IgG and IgM both ELISA and western blot tests), the district from where the laboratory tests were conducted, the total number of reimbursed tests and an indicator if the laboratory or laboratories in the district belong to the sentinel network or not. Due to the complexity of *B. burgdorferi* s.l. and its many clinical manifestations making it very difficult to detect the disease using ELISA screening, only confirmatory serology tests by Western Blot were considered. The variables of interest are the number of reimbursed tests for *B. burgdorferi* s.l conducted by the laboratories belonging to the sentinel network (SNL) and non-sentinel network (NSNL) for each region, province and district and the year. All the microbiology tests for which laboratories have claimed reimbursement to the compulsory national social security system are contained in the INAMI-RIZIV database. With the unlikely exception of tests that were performed without being reimbursed, the database is therefore said to be virtually exhaustive both over time and place.

2.1.2 Data on Confirmed positive serology Test

The data set contains patients who are confirmed for positive serology test on Lyme by the SNL. Each participating laboratory reports number of confirmed Lyme cases to Sciensano on a weekly basis. The place of residence, age, gender, date of birth, test date of the patients, test method, week, month and year of diagnosis are recorded alongside.The variables of interests are the aggregated number of cases (confirmed serology tests by Western Blot) at the level of the region, province and district from 2008 to 2015. Within each year all duplicate values where checked and removed if present. The response variable is the aggregated counts.

2.1.3 Consultations on Erythema migrans (EM)

For each consultation concerning a tick bite or a presumption of lyme disease, the SGP records the presence or absence of erythema migrans and the number of tick bites. Also some basic patient characteristics such as age, sex, geographic location of tick bite, patient place of residence, diagnostic methods and treatments patient received are recorded by the SGP. The variables of interests are the aggregated number of cases (erythema migrans) and aggregated number of tick bites at the district level,as well as the total number of SGP doctors for each region, province and district for two time periods, 2008-2009 and 2015-2017.

2.1.4 Data from TekenNet

TekenNet is a web application that is used by the general public to report tick bites in Belgium. Data include the number of bites, the location (post code) and the date of the bite. Variables of interest are the aggregated number of tick bites at the district level between July 2015 and June 2017.

2.2 Descriptive Data Analysis

By looking at the different levels of test coverage, the data on the reimbursement test from INAMI-RIZIV are used to asses the aspect of representativeness of the SNL. Due to the limits of the surveillance data (absence of the information on the true number of diagnosed cases in Belgium), the reimbursement data was used as a way of alternative case coverage. The test coverage is the ratio between the number of reimbursed tests performed by the SNL and the total number of tests reimbursed in Belgium. High values of test coverage indicates that most of the tests that are performed and the positive results are captured by the surveillance system. In contrast, a low value of test coverage indicates that most of the tests are performed by laboratories that do not take part in the surveillance system and thus positive results are not reported. The test coverage of the SNL of lyme disease is examined at the national level, its variations by region and province and the stability of the coverage over time between 2010 to 2015. Since the study is based on the total number of tests reimbursed in Belgium for lyme disease, only descriptive statistics is reported without statistical inference made. Graphical representation of coverage values will allow to visually identify variations over time and place.

2.3 Disease Mapping in Spatial Data

In epidemiological studies, disease mapping has a long history with one major goal in investigating the geographical distribution of disease burden [14]. Spatial data also known as geospatial data or geographical information are defined as the realizations of a stochastic process measured by space.

$$Y(S) \equiv \{y(s), s\epsilon D\}$$

D is a fixed subset of $\mathbb{R}^d$ we consider $d = 2$. The data can be represented by a collection of observations $\mathbf{y} = \{y(s_1), ..., y(s_n)\}$, where the set $(s_1, ..., s_n)$ indicates the spatial units (areas) where the measurements are taken. The problem can be specified as spatially continuous or discrete random process depending if D is a continuous surface or a countable collection of d-dimensional spatial units [15].

2.3.1 Bayesian hierarchical models

Bayesian models are very attractive models since they provide a unified approach to data analysis [16]. The last three decades has seen a great development of Bayesian methods which are now widely established in many research areas, from clinical trials [17], to health economic assessment [18], social sciences [19], and epidemiology [20]. The main idea behind a Bayesian approach is that only one form of uncertainty exists. This uncertainty is described by suitable probability distributions. There is no difference between observable data and un-observable parameters because the parameters are also considered random quantities. A *prior* distribution describes the uncertainty about the realized value of the parameters given the current state of information. To derive the *posterior* distribution, the inferential process combines the prior and the current data model. Epidemiological data are usually characterized by a spatial and/or temporal structure which needs to be taken into account when doing inferences. Bayesian approach is generally particularly effective under these circumstances and has been applied to several epidemiological applications from ecology to environmental studies to infectious diseases[21]. Given data consisting of aggregated counts of outcome and some covariates, diseases mapping and/or ecological regression can be specified [21]. This model can be specified in a Bayesian framework by extending the concept of hierarchical structure which gives the allowance to account for similarities based on the neighborhood at the area-level. The main challenge in Bayesian statistics in this case resides in the computational aspects. Bayesian computations normally use Markov Chain Monte Carlo (MCMC) methods [22], this is arguably thanks to the wide popularity of the *BUGS* software [23].In trivial cases MCMC methods involve computational and time intensive simulations to obtain the posterior distribution for the parameters even though it is extremely flexible and able to deal with virtually any type of data and model. Thus the complexity of the model and the database dimension will often remain fundamental issues. Recently the Integrated Nested Laplace Approximation (INLA) [24] approach has been developed as a computationally efficient alternative to MCMC. Designed for *latent Gaussian models*, INLA can be successfully used in a great variety of applications [25]. This is because latent Gaussian models are a very wide and flexible class of models ranging from generalized linear mixed models to spatial and spatio-temporal models.This is also thanks to the R package called R-INLA [26].
Bayesian model is usually characterized by 3 stages of observations and parameters. The first stage consists of distributional assumptions for the observations. Given that disease counts $y_i(i = 1, ...I)$ observed

for I geographic regions within a time period that is pre-specified, y_i can be assumed to follow a poisson distribution with rate parameter λ_i.This is the relative risk of disease case in region i. It is assumed that the y_is are conditionally independent given the λ_is. The second stage is defined by a prior model for the λ_is or more often a specific transformation of them. It is most common to use $log(\lambda_i) = \eta_i$. The variable η_i is usually an additive term of unknown random components, called the linear predictor [27]. Assigning Gaussian priors to all components of the linear predictor can lead to high flexibility. Such models are also referred to as latent Gaussian models [28]. The final stage consist of prior distributions of unknown hyperparameters $\psi_1, ..., \psi_R$, which are typically the variances or correlations for random effects within η.This setting is very appealing for modeling spatial and spatio-temporal data [29].

Gaussian Markov random fields and conditional independence A vector x can be formed consisting of the linear predictor η^T and all its additive components. Since the $\eta_i's$ are on the first I position in the vector x, each observation y_i depends only on its corresponding ith element x_i in the vector x. Also, since the Gaussian priors are assigned to all components of x as mentioned earlier, the vector x is also Gaussian and forms a so called Gaussian Markov random field(GMRF) [30]. Sometimes GMRFs are called conditional autoregressions since they fulfill conditional independence or the so called Markov properties [31]. Let $\pi(\cdot|\cdot)$ denote a conditional density of its arguments and given two variables x_i and x_j where $(i \neq j)$, x_i and x_j are said to be conditionally independent given x_{-ij} if $\pi(x_i, x_j|x_{-ij}) = \pi(x_i|x_{-ij}) \cdot \pi(x_j|x_{-ij})$. This can also be written as $x_i \perp x_j|x_{-ij}$. x_{-ij} contains all components of x except for the ith and jth components. The conditional independence between two components x_i and x_j of a GMRF can be read off from its precision matrix $\mathbf{Q}$ and it holds that

$$x_i \perp x_j|x_{-ij} \Longleftrightarrow Q_{ij} = 0$$

. Formally, a random vector $\mathbf{x} = (x_1, ..., x_n)^T$ can be defined as a GMRF with mean μ and a positive definite precision matrix Q , if it has a density of the form

$$\pi(x) = (2\pi)^{\frac{-n}{2}} |Q|^{\frac{1}{2}} exp\Big(-\frac{1}{2}(x-\mu)^T Q(x-\mu)\Big)$$

. The inverse of the precision matrix is the covariance matrix of the GMRF $\Sigma = Q^{-1}$. The GMRF at the second stage within a hierarchical model provides a flexible tool to model the dependence between latent effects this implicitly models the dependence between the observed data [24]. The key to the fast computation of integrated nested laplace approximation is due to non zero pattern (sparseness) of the precision matrix Q because of the Markov properties. Generally, numerical methods for sparse matrices are much quicker than dense matrices [30].

2.3.2 Integrated Nested Laplace Approximation (INLA)

The interest in statistical analysis is often to estimate th effect of a set of covariates on a function of the observed data taking into account the spatial correlation implied in the model. The problem can be specified by modeling the ith unit by means of an additive linear predictor which is defined on a suitable scale for example; logistic for binomial data or poison for count data.

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_1(z_{li}) \tag{1}$$

β_0 is defined as a scalar which represents the intercept, the parameters β quantifies the effects of some covariates x where x=$(x_1, ..., x_M)$ on the response variable.The set of covariates $\mathbf{z} = (z_1, ..., z_l)$ are defined by the collection of functions $\mathbf{f} = \{f_1(.), ..., f_L(.)\}$. When the form of the functions are varied, the formulation can accommodate a wide range of models from hierarchical regression to spatial and spatio-temporal models [24].
The vector of parameters is represented by

$$\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \mathbf{f}\}$$

given the above specification. A GRMF prior on θ can be assumed with mean $\mathbf{0}$ and a precision matrix $\mathbf{Q}$.

For each of the elements of the vector of parameters their marginal posterior distributions are computed and also if possible, the computation for each element of the hyper parameters vector.

$$\pi(\theta_i|\mathbf{y}) = \int_{\boldsymbol{\psi}} \pi(\theta_i|\boldsymbol{\psi}, \mathbf{y})\pi(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \tag{2}$$

The integral in 2 can be approximated by the finite sum

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_k \tilde{\pi}(\theta_i|\psi_k, y)\tilde{\pi}(\psi_k|\mathbf{y})\ \Delta_k, \tag{3}$$

The approximations of $\pi(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ and $\pi(\boldsymbol{\psi}|\mathbf{y})$ are given by $\tilde{\pi}(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ and $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{y})$ respectively. The sum is finally evaluated at support points $\boldsymbol{\theta_k}$ by numerical integration using appropriate weights Δ_k.
Two computations needs to be done;

- The approximation of $\pi(\boldsymbol{\psi}|\mathbf{y})$, where all the relevant marginals of the hyper-parameters $\pi(\psi_k|\mathbf{y})$ can also be obtained. From the factorization of $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{y}) = \pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y})$ it follows that $\pi(\boldsymbol{\psi}|\mathbf{y})$ can be approximated by

  $$\tilde{\pi}(\psi|\mathbf{y}) \propto \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{y})}{\tilde{\pi_G}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})}\ |_{\theta=\theta^*(\psi)} \tag{4}$$

  The denominator $\tilde{\pi}_G(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ denotes the Gaussian approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$ with $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ the mode for a given $\boldsymbol{\psi}$. This ratio which corresponds to the Laplace approximation is introduced by Tierney and Kadane and is mainly needed to integrate out the uncertainty with respect to $\boldsymbol{\psi}$ through equation 3 [32]. This task consist of the computation of an approximation to the posterior marginal distribution of the hyper-parameters.

- $\pi(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ is needed to compute the marginal posterior for the parameters. Three strategies can be employed: A Gaussian, a full Laplace and a simplified Laplace approximation. Gaussian approximation is computationally most convenient where each marginal $\tilde{\pi}_G(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ can simply be derived from $\tilde{\pi}_G(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y})$, i.e. $\tilde{\pi}_G(\theta_i|\boldsymbol{\psi}, \mathbf{y}) = N(\theta_i; \mu_i(\boldsymbol{\psi}), \sigma_i^2(\boldsymbol{\psi}))$ where the mean and marginal variance of the Gaussian approximation are $\mu_i(\boldsymbol{\psi})$ and $\sigma_i^2(\boldsymbol{\psi})$ respectively [26]. However, errors due to location of the posterior marginals or errors due to lack of skewness or both can occur [26]. A typically very exact alternative but time consuming is to use another Laplace approximation for $\pi(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ similar to equation 4.

  $$\tilde{\pi}(\theta_i|\boldsymbol{\psi}, \mathbf{y}) \propto \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{y})}{\tilde{\pi_G}(\boldsymbol{\theta_{-i}}|\theta_i, \boldsymbol{\psi}, \mathbf{y})}\ |_{\theta_{-i}=\theta^*_{-i}(\theta_i,\psi)} \tag{5}$$

  Since the denominator must be recomputed for each θ_i and $\boldsymbol{\psi}$, the evaluation of this approximation is computationally very demanding. The so-called simplified Laplace approximation $\pi_{SLA}(\theta_i|\boldsymbol{\psi}, \mathbf{y})$ is less expensive than the full Laplace approximation with only a slight loss of accuracy. A series expansion of the Laplace approximation in equation 5 performed around $\theta_i = \mu_i(\boldsymbol{\psi})$ which allows to correct the Gaussian approximation for location and skewness. More details on this method on how to compute the two steps to get the necessary marginal posteriors of the parameters and hyper-parameters can be found in the papers [24, 33, 34].

2.3.3 Besag-York-Mollie (BYM) Model

In small area estimation studies, disease mapping is commonly used to assess the pattern of a particular disease or condition and to identify the areas that are characterized by high or low relative risks. Since the number of cases are count data, for the ith area, the number of cases y_i is modeled as

$$y_i \sim Poisson(\lambda_i)$$

where the mean λ_i is defined in terms of the rate ρ_i and the expected number of cases e_i.

$$\lambda_i = \rho_i e_i$$

.

$$e_i = p_i \left(\frac{\sum_i y_i}{\sum_i p_i} \right)$$

y_i is the number of cases of disease observed in area i and p_i is the total population at risk of disease in the ith area. A general model formulation is to assume that the log risk $\eta_i = log(\rho_i)$ has the decomposition:

$$\eta_i = \alpha + \mathbf{z}_i^T \beta + b_i$$

. The overall risk level is denoted by α, $\mathbf{z}_i^T = (z_{i1}, ..., z_{ip})^T$ is a set of p covariates with corresponding parameters $\beta = (\beta_1, ..., \beta_p)$ and b_i a random effect. The extra-Poison variation or spatial correlation due to latent or unmeasured risk factors are accounted for by the random effects $\mathbf{b}= (b_1, ..., b_n)^T$. Generally, it seems reasonable to assume that areas that are close in space show more similar disease burden than areas that are not close. The meaning of "close" can be defined by setting up a neighbourhood structure. Areas i and j are assumed to be neighbours if they share a common border here denoted as $i \sim j$. Furthermore, the set of neighbours of region i is denoted by δi and the size by $n_{\delta i}$.
Intrinsic Gaussian Markov random field here referred to as the Besag model is one of the most popular approaches to model spatial correlation [31]. The conditional distribution for b_i is

$$b_i \mid \mathbf{b}_{-i}, \tau_b \sim N\left(\frac{1}{n_{\delta i}} \sum_{j \in \delta_i} b_j, \frac{1}{n_{\delta i} \tau_b}\right),$$

where τ_b is a precision parameter and $\mathbf{b}_i = (b_1, ..., b_{i-1}, b_{i+1}, ...b_n)^T$. The mean of the effect over all neighbours equals the mean of b_i and the precision is proportional to the number of neighbours. The joint distribution for $\mathbf{b}$ is given by

$$\pi(\mathbf{b} \mid \tau_b) \propto exp\left(-\frac{\tau_b}{2} \sum_{i \sim j} (b_i - b_j)^2\right) \propto exp(-\frac{\tau_b}{2} \mathbf{b}^T \mathbf{Q} \mathbf{b}),$$

where the precision matrix has the entries

$$Q_{ij} = \begin{cases} n_{\delta i} & i = j \\ -1 & i \sim j \\ 0 & else \end{cases}$$

For the Besag model, the rank deficiency is equal to the number of connected subgraphs. If all regions are connected the rank deficiency is equal to one, hence the density for the Besag model is

$$\pi(\mathbf{b} \mid \tau_b) = K \tau_b^{\frac{n-I}{2}} exp\left(-\frac{\tau_b}{2} \mathbf{b}^T \mathbf{Q} \mathbf{b}\right),$$

the number of connected subgraphs is denoted by I and K is a constant. A sum-to-zero constraints are imposed on each connected subgraph so as to prevent confounding with the intercept.
The Besag model cannot take the limiting form that allows for no spatially structured variability since it only assumes a spatially structured component. Unstructured random errors or pure overdispersion within area i will be modelled as spatial correlation. This will give misleading parameter estimates. The issue is addressed by Besag-York-Mollie (BYM) model [31] which decomposes the regional spatial effect $\mathbf{b}$ into a sum of an unstructured and structured spatial components. The linear predictor is then defined as

$$\eta_i = log(\rho_i) = \alpha + \nu_i + v_i \tag{6}$$

The two area specific effects are $\nu_i = f_1(i)$ and $v_i = f_2(i)$. In the BYM specification, ν_i is called the spatially structured residual and v_i represents the unstructured residual. The spatially structured component ν_i is modeled using an intrinsic conditional autoregressive structure (the Besag model) $\nu_i \sim N(\mathbf{0}, \tau_\nu^{-1} \mathbf{Q}^-)$. And v_i which accounts for pure overdispersion is modelled as $v_i \sim N(\mathbf{0}, \tau_v^{-1} \mathbf{I})$ using the exchangeable prior, whereby $\mathbf{Q}^-$ denotes the generalised inverse of $\mathbf{Q}$. The resulting covariance matrix of $\mathbf{b}$ from the BYM model is given by

$$Var(\mathbf{b} \mid \tau_v, \tau_\nu) = \tau_v^{-1} \mathbf{I} + \tau_\nu^{-1} \mathbf{Q}^-$$

R-INLA parameterizes the two components into $\xi_i = \nu_i + v_i$ and ν_i.

The proportion of variance explained by the structured spatial component can be evaluated. Since σ_ν^2 is the variance of the conditional autoregressive specification while σ_v^2 is the variance of the marginal unstructured component, the two are not directly comparable thus the posterior marginal variance of the structured effect is gotten empirically through

$$s_\nu^2 = \frac{\sum_{i=1}^{n} (\nu_i - \bar{\nu})^2}{n-1} \tag{7}$$

where $\bar{\nu}$ is the average of ν. This is then compared to the posterior marginal variance for the unstructured effect provided by σ_v^2.

$$frac_{spatial} = \frac{s_\nu^2}{s_\nu^2 + \sigma_v^2}$$

.

As in any Bayesian analysis, the choice of the prior may have considerable effects on the results. It is thus necessary to consider what prior to use and to conduct sensitivity analyses to assess the influence of the prior on the estimations [34]. By default, minimally informative priors are specified on the log of the unstructured effects precision $log\tau_v \sim logGamma(1, 0.0001)$ and on the log of the structured effect precision $log\tau_\nu \sim logGamma(1, 0.0001)$.

We want to investigate the risk of tick exposure between 2015 to 2017 using the aggregated number of tick bites reported through TekenNet and to investigate the risk of having LB using the aggregated number of cases of EM in the period 2008-2009 and 2015-2017 from patient consultations at sentinel GPs. The risk of having a positive serology results of LB using the confirmed positive serology results between 2008-2016 from the SNL will also be estimated. These relative risks compared to the whole of Belgium will be estimated at the district level. In this study, the risk of exposure to ticks was used as a proxy for the risk of exposure to LB.

3 Results

3.1 Exploratory Data Analysis

Given the four different datasets, exploratory data analyses where done at different levels.

3.1.1 Data on Reimbursement tests for *B. burgdorferi* s.l. serology

This data contains the number of Western Blot serology tests reimbursed for LB between 2010-2015 period. The number of tests where aggregated at the national, regional and provincial levels. Test coverages were calculated at each level.

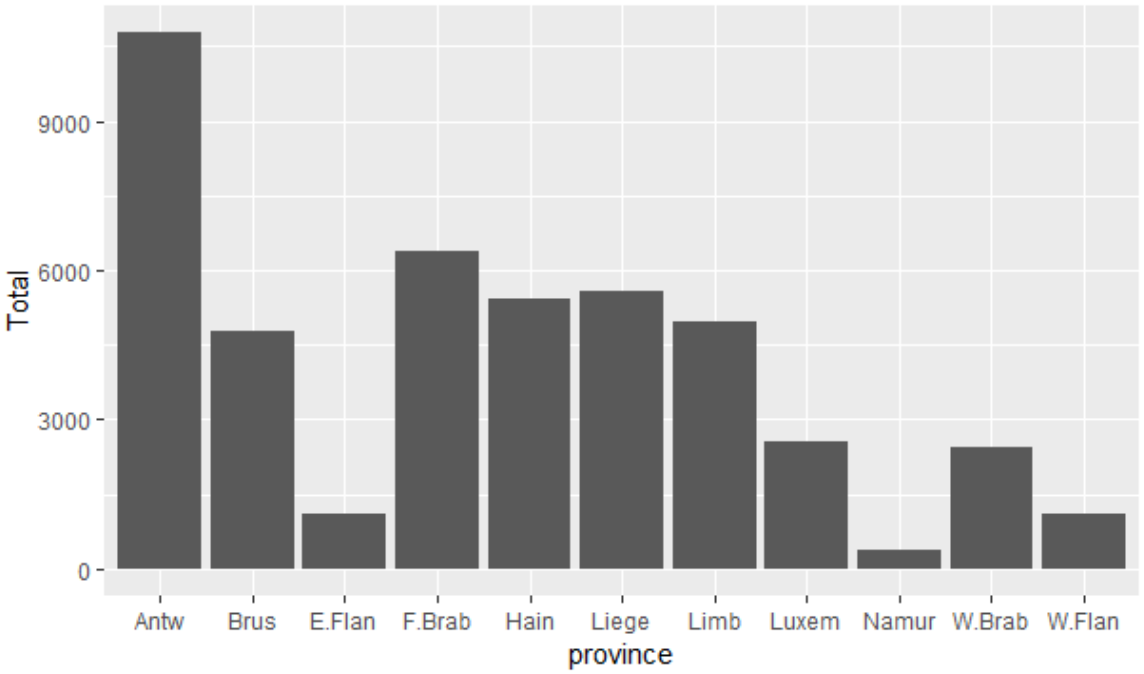


Figure 1: Total number of serology tests reimbursed between 2010 and 2015 by province

Figure 1 shows the total number of Western Blot serology tests on LB reimbursed by INAMI-RIZIV between 2010 and 2015 by province in Belgium. Antwerpen shows the highest number of serology tests reimbursed within this period followed by Flemish Brabant, Hainaut, Liege, Limburg and Brussels.

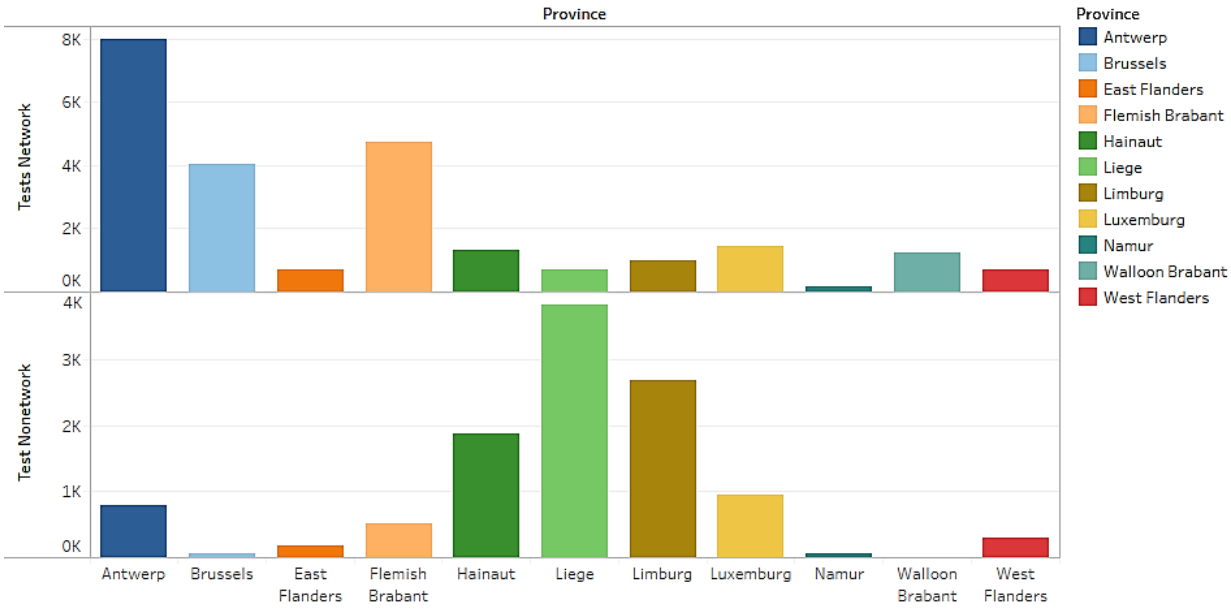


Figure 2: Total number serology tests reimbursed per network (SNL versus NSNL) between 2010 and 2015 by province

The total number of reimbursed serology tests conducted by the SNL were highest in the provinces of Antwerpen, Flemish Brabant and Brussels while Liege, Limburg and Namur had the highest total number of reimbursed serology tests conducted by non sentinel network of laboratories. This is shown in figure 2 above.

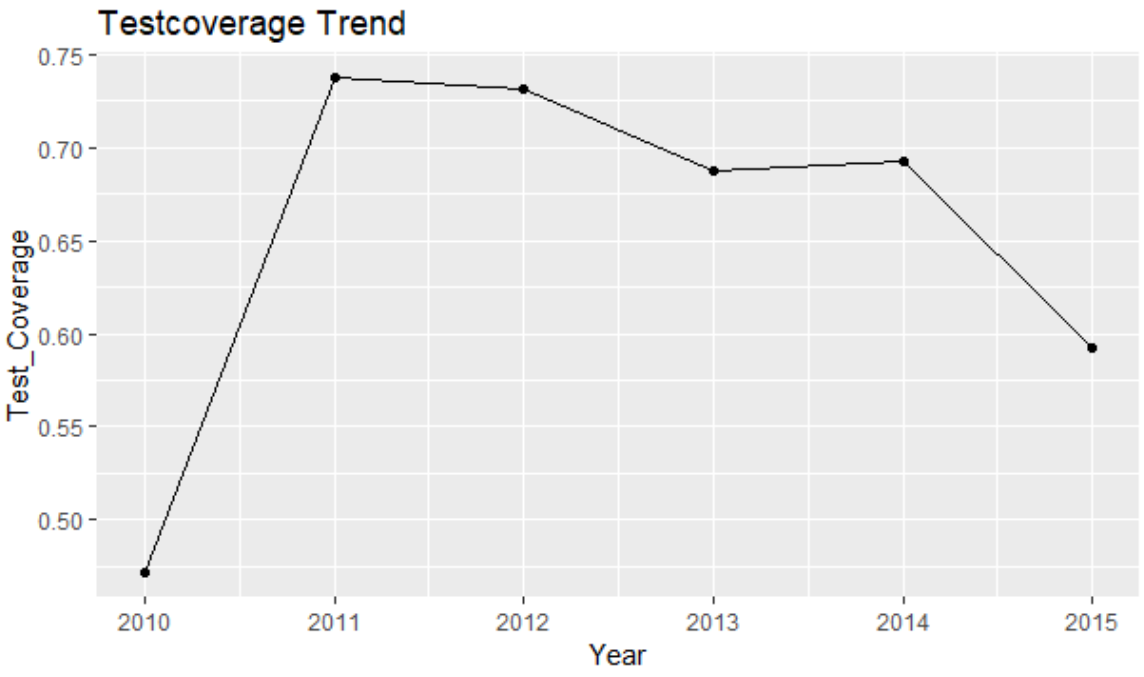


Figure 3: Evolution of test coverage of Sentinel Laboratory Network of lyme in Belgium (2008-2015)

At the national level, the SNL performed the majority of the reimbursed tests in 2011 with a test coverage of 73%. The test coverage was just under 50% in 2010 but rapidly increased in 2011 and later drops slowly to above 55% in 2015 with a range between approximately 50 to 70%. The overall mean test coverage at the national level was 65%. This is represented in figure 3

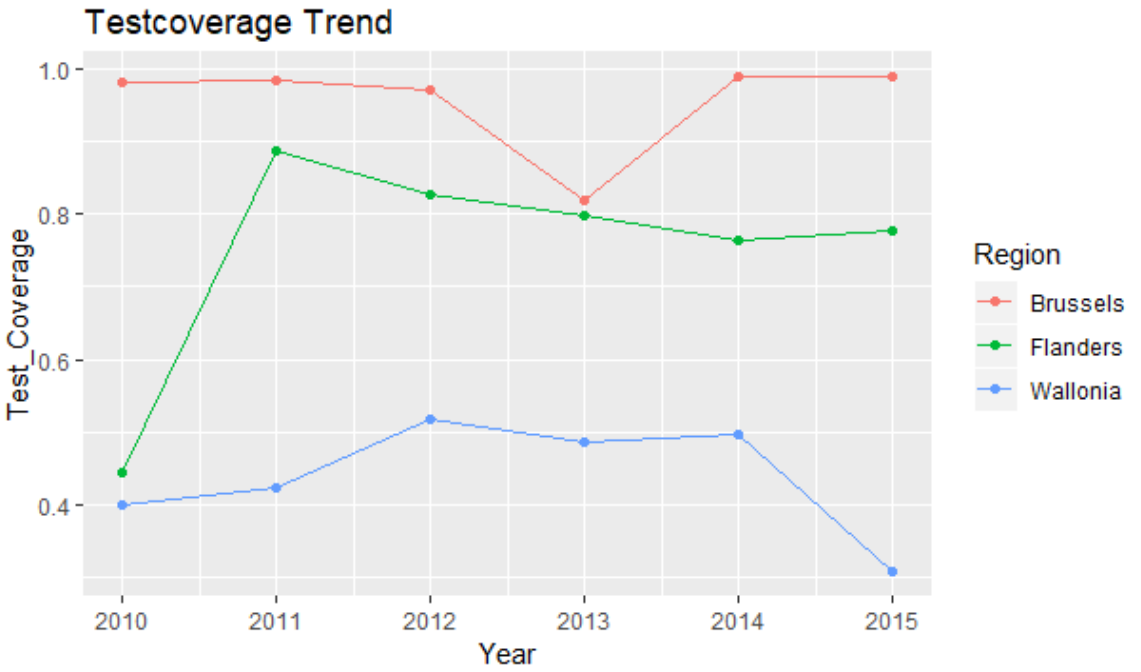


Figure 4: Evolution of the test coverage of Sentinel Laboratory Network of lyme by region in Belgium (2010-2015)

Regional test coverage shows that coverage was generally highest in Brussels with an average of over 95% and lowest in Wallonia with average of 44%. Flanders had an average coverage of 75% above that at the national level (65%). Figure 4 shows the changes in coverage between the regions. Coverage was stable over time in Brussels but for a sudden drop in 2013. The box plot in figure 6a shows the different average levels of coverage for the three regions.

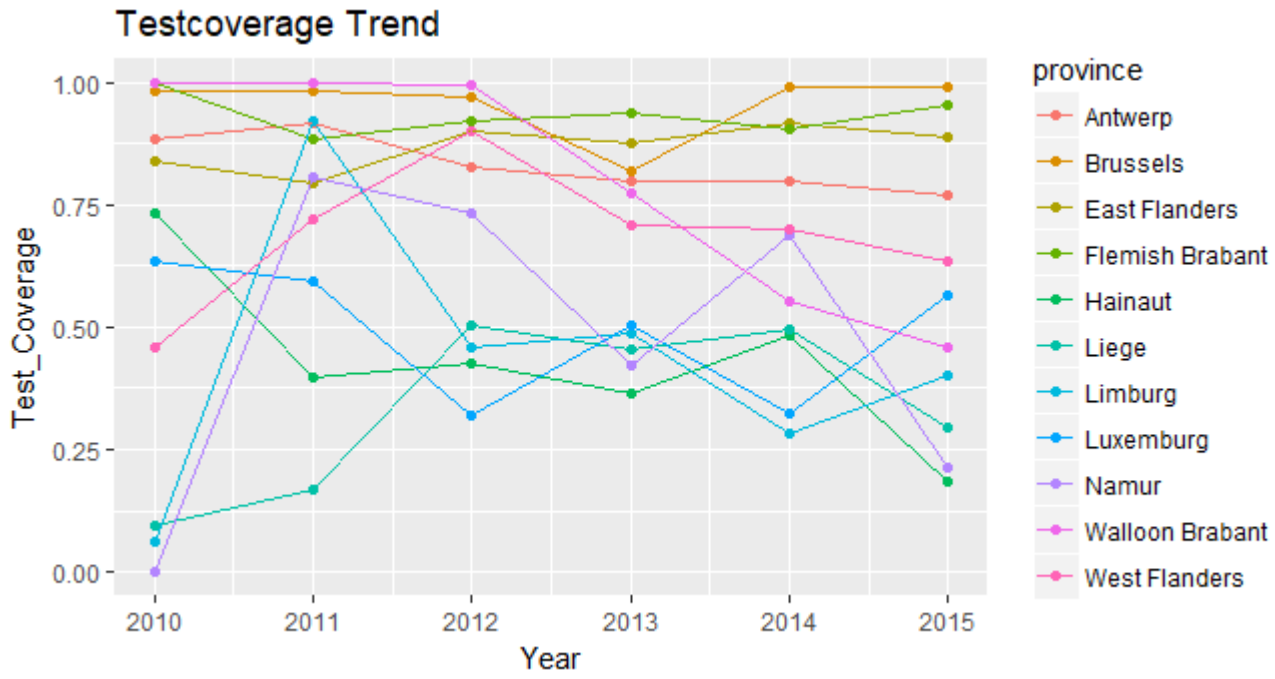
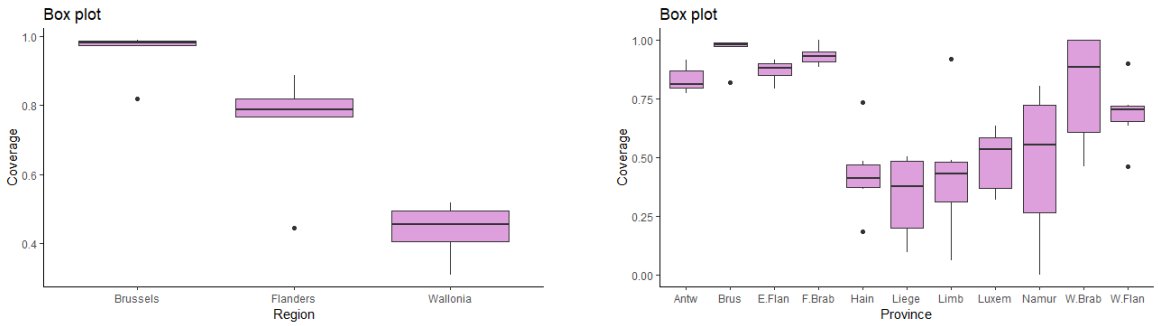


Figure 5: Evolution of the test coverage of Sentinel Laboratory Network of lyme by province in Belgium (2010-2015)

At the provincial level, variation in coverage was larger than at the regional level. Much variability is seen in the provinces of Namur, Walloon Brabant and Liege, with global average coverage of below 50% while Antwerpen, Brussels, East Flanders, west Flanders and Flemish Brabant had much lesser variability in coverage with an average coverage of above 70% as seen in figure (6b). In 2010, Namur had the lowest coverage of 0% closely followed by Limburg and Liege while all the other provinces had coverages of 50% and above. Coverage rapidly increased to above 75% for the provinces of Namur and Limburg in 2011 and then continues to vary till 2015.



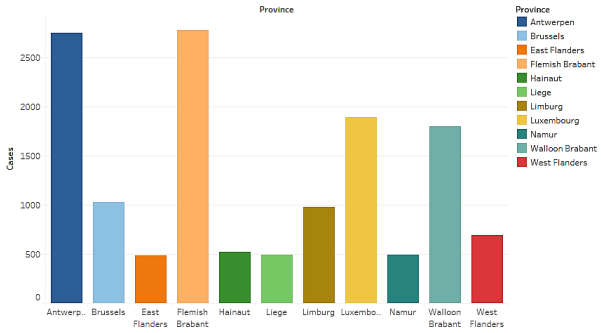
(a) Box plots for regional coverage

(b) Box plots for provincial coverage

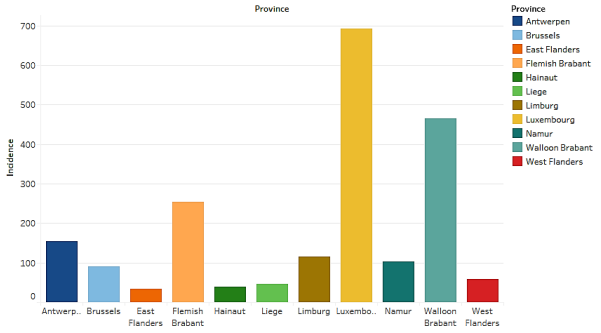
Figure 6: Regional and provincial box plots for serology test coverage of SNL

3.1.2 Data on the confirmed positive serology tests for *B. burgdorferi* s.l

This dataset contains all the positive Western Blot serology results reported by SNL between 2008 and 2016.



(a) Aggregated number of positive serology tests for LB reported by SNL by province (2008-2016)



(b) Reported cumulative incidence of positive serology tests for LB per 100,000 inhabitants by province (2008-2016)

Figure 7: Number of positive serology tests on LB between 2008 - 2016

Cumulative incidence of positive serology results for LB per 100,000 inhabitants at the provincial level between 2008 and 2016 was highest in the provinces of Luxemburg and Walloon Brabant. This was followed by Flemish Brabant and Antwerpen which are approximately three times lower. The other provinces had much lower cumulative incidence as shown in figure 7b. Variability in incidence over time was also much higher in the provinces of Luxemburg and Walloon Brabant.Flemish Brabant had a gradual decrease in incidence from 2008 until 2012 where it starts increasing and reaches it's peak in 2014. All the other provinces had constant incidence from 2008 until 2012 where they start increasing and also peaked in 2014. This is represented in figure 8

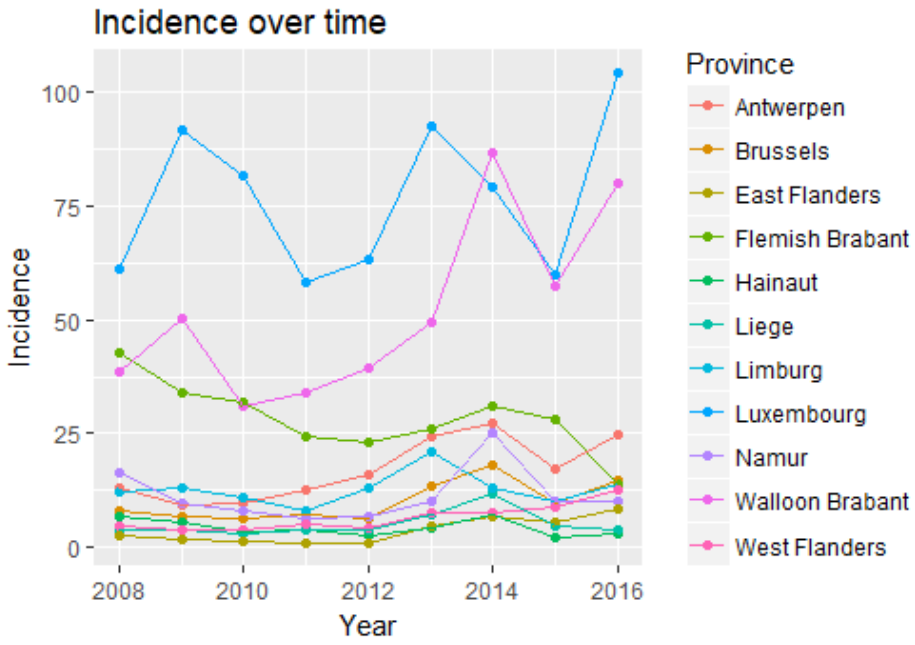


Figure 8: Evolution of incidence of positive serology results tests for LB per 100,000 inhabitants by province

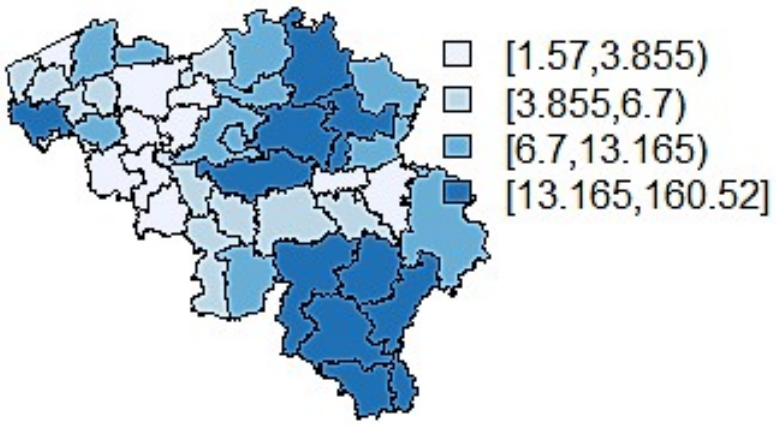


Figure 9: Map of the cumulative incidence of positive serology tests for LB per 10,000 inhabitants by district

3.1.3 Consultations from SGP

This data contains all the confirmed cases of EM reported by general practitioners under the sentinel network between 2008 - 2009 and 2015 - 2017 at the district level. Cumulative incidence of having an EM was calculated per 10,000 inhabitants at the district level.

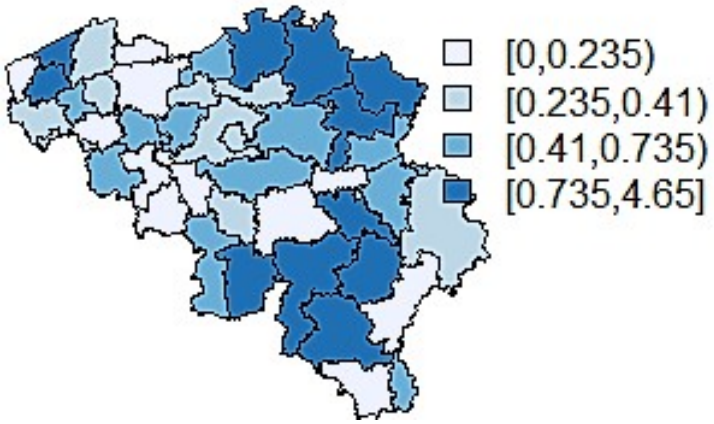


Figure 10: Map of the cumulative incidence of the cases of EM per 10,000 inhabitants by district

3.1.4 Data on Tick bites (source TekenNet)

The data contains all the number of tick bites and location of bites in Belgium. Cumulative incidence of tick bites was calculated per 10,000 inhabitants at the district level.

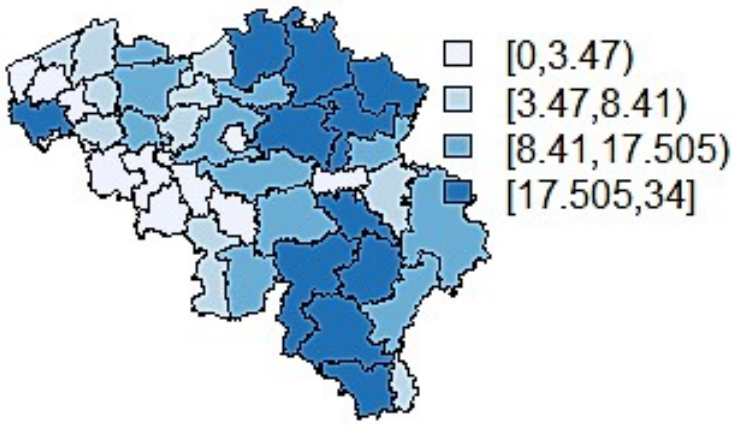


Figure 11: Map of the cumulative incidence of tick bites per 10,000 inhabitants by district

Generally, most of the districts in the province of Luxembourg, Limburg, Antwerpen and Flemish Brabant showed very high cumulative incidence of tick bites while the districts in the provinces of Hainaut, East Flanders and west Flanders has very low cumulative incidence of tick bites as shown in figure 11. This result is very similar to the results of the cumulative incidence of positive serology results for LB (figure 9). The results of the cumulative incidence of the cases of EM also have similar patterns but for a few districts in the province of Luxembourg that showed very low cumulative incidence of the EM cases (figure 10).

3.2 Disease mapping by BYM Specification

The INLA model with BYM specification was used to analyse three data sets; the aggregated number of EM cases reported by the SGP (mod1) in 2008-2009 and 2015-2015, the aggregated number of tick bites recorded on TekenNet website between 2015 to 2017 (mod2) and finally the aggregated number of positive serology results for LB reported by the SNL between 2008-2015 (mod3).
The parameters estimated by INLA are represented by $\boldsymbol{\theta} = \{\alpha, \xi, \nu\}$ and the hyper-parameters are given by the precisions $\boldsymbol{\psi} = \{\tau_\nu^2, \tau_v^2\}$.

The posterior means and the standard deviation together with the 95% credible interval for the intercept (α) are reported on table 1 below. α quantifies the average rate of LB in all the 43 districts of Belgium. Sensitivity analysis was conducted with three different log gamma priors. The parameter estimates using the different priors are also reported in table 1. The random effects are (ξ, ν). ξ gives information on the area specific residuals, the primary interests in disease mapping and ν presents information on the spatially structured residuals only. The posterior means of the district-specific relative risks of LB compared to the whole of Belgium are given by $\zeta = exp(\xi)$. The uncertainty (posterior probabilities) associated with the posterior means are also reported. This provide useful information since interest lies in excess risk.The posterior probability $p(\zeta > 1|y)$ is visualized on a map (in appendix A). The relative risks for mod1, mod2 and mod3 are plotted in figures 12, 13 and 14 respectively.

Table 1: Fixed effects from BYM models

Parameter	Prior	mean	sd	95% CI
α_{mod1}	$log\Gamma$(1,0.001)	-0.4037	0.0688	(-0.5432, -0.2728)
	$log\Gamma$(1,0.0001)	-0.4032	0.0677	(-0.5404, -0.2700)
	$log\Gamma$(1,0.1)	-0.4097	0.0819	(-0.5744, -0.2519)
α_{mod2}	$log\Gamma$(1,0.001)	-0.4648	0.098	(-0.6577, -0.2698)
	$log\Gamma$(1,0.0001)	-0.4649	0.0979	(-0.6576, -0.2698)
	$log\Gamma$(1,0.1)	-0.4609	0.0753	(-0.608, -0.307)
α_{mod3}	$log\Gamma$(1,0.001)	-0.4824	0.0223	(-0.5256, -0.4395)
	$log\Gamma$(1,0.0001)	-0.4824	0.0183	(-0.5187, -0.4469)
	$log\Gamma$(1,0.1)	-0.4831	0.0559	(-0.5963, -0.3698)

The proportion of spatial variance for mod1, mod2 and mod3 are 0.64, 0.73 and 0.71 respectively. This means that like in mod1, 64% of the variability is explained by the spatial structure. The posterior means of the exponentiated intercepts α for all the 3 models have relative risks less than 1. This implies an overall reduction of risk across the whole of Belgium. Given mod1 with $exp(\alpha) = exp(-0.4037) = 0.67$ implies a risk reduction by a factor of 0.67 or a 33% lower risk of EM across the whole of Belgium. The same analogy follows for mod2 with 37% lower risk of tick bites and for mod3 with 38% lower risk of positive serology results across the whole of Belgium respectively. All the models have approximately the same % of risk reduction across the whole country.
The results of the sensitivity analysis carried out by using the different priors showed that the priors didn't have any effects on the parameter estimates of the models. This is because the estimates of the fixed effect (α's) were stable after different priors were applied to the models. This is shown in table 1.

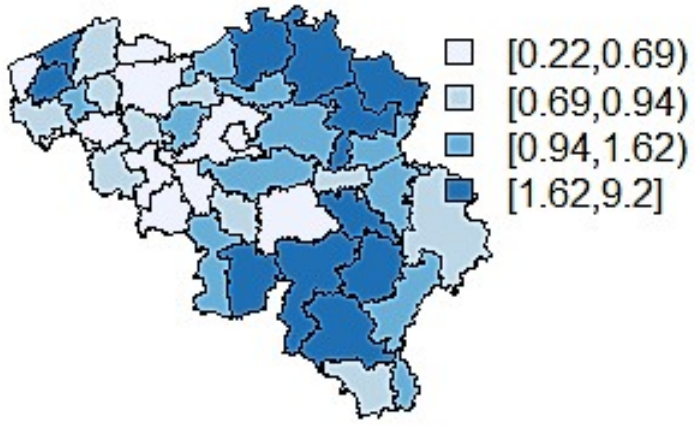


Figure 12: Distribution of district-specific relative risks of EM compared to the whole of Belgium

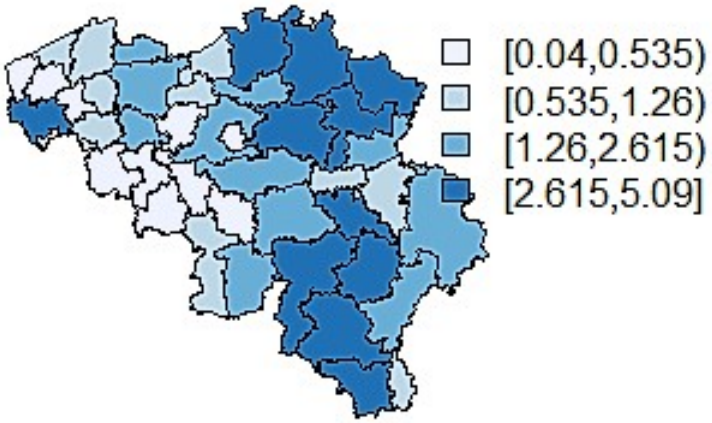


Figure 13: Distribution of district-specific relative risks of tick exposure compared to the whole of Belgium

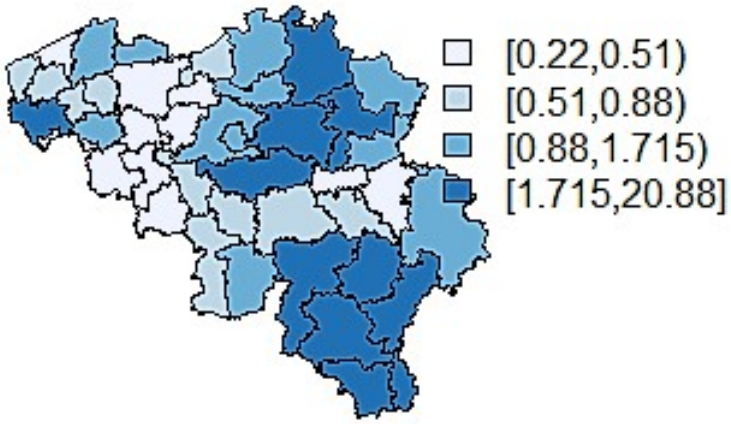


Figure 14: Distribution of district-specific relative risks of positive serology for LB

The results from the disease mapping models (figures 12, 13, 14) showed that the relative risks from all the models have similar patterns. The high risk districts are mostly from the provinces of Luxembourg, Antwerpen, Limburg and Flemish Brabant. These results showed similar patterns to those of the high risk districts estimated by the cumulative incidence per 10,000 inhabitants (figures 9, 10, 11). The random effects for each district from the models and the length of their confident intervals are given in the tables below. Comparing the length of the confidence intervals from the models, the model with the overall shortest length of confidence interval is mod3 (positive serology tests for LB) as shown in figure 17 followed by the confidence intervals from mod2 except for a few districts (figure 16). The length of the confidence intervals of the random effects estimated from the cases of EM were the longest overall (figure 15).

District-specific random effects and length of confidence intervals

Table 2: District specific random effects of EM (mod1)

	mean	sd	lower	upper	length
Brussel-Hoofdstad	-0.48	0.18	-0.85	-0.14	0.71
Antwerpen	0.65	0.12	0.41	0.89	0.48
Mechelen	-0.09	0.25	-0.60	0.37	0.98
Turnhout	1.69	0.12	1.46	1.92	0.45
Hasselt	1.69	0.12	1.46	1.92	0.46
Maaseik	1.88	0.13	1.61	2.14	0.53
Tongeren	0.17	0.28	-0.40	0.69	1.10
Aalts	-0.01	0.26	-0.56	0.48	1.03
Dendermonde	-0.17	0.32	-0.83	0.43	1.26
Eeklo	-1.01	0.66	-2.43	0.16	2.59
Gent	-0.95	0.28	-1.53	-0.44	1.09
Oudenaarde	-0.31	0.39	-1.12	0.41	1.53
Sint Niklaas	-0.11	0.29	-0.72	0.43	1.15
Halle Vilvoorde	-0.56	0.23	-1.04	-0.13	0.91
Leuven	0.42	0.17	0.07	0.75	0.68
Brugge	-0.18	0.28	-0.75	0.33	1.09
Diksmuide	0.64	0.40	-0.18	1.37	1.55
Leper	-0.37	0.56	-1.55	0.67	2.22
Kotrijk	-1.03	0.36	-1.80	-0.37	1.43
Oostende	0.48	0.28	-0.11	1.00	1.11
Roeselare	-0.05	0.34	-0.75	0.58	1.33
Tielt	-0.41	0.47	-1.40	0.45	1.84
Veurne	-0.44	0.57	-1.66	0.58	2.24
Nijvel	0.09	0.22	-0.35	0.50	0.86
Aat	-0.75	0.53	-1.86	0.22	2.08

Charleroi	-0.20	0.24	-0.69	0.25	0.93
Bergen	-1.65	0.54	-2.81	-0.70	2.11
Moeskroen	-0.55	0.55	-1.71	0.44	2.15
Zinnik	-1.08	0.41	-1.94	-0.33	1.61
Thuin	-0.09	0.35	-0.82	0.56	1.38
Doornik	-0.24	0.38	-1.02	0.46	1.48
Hoei	0.84	0.27	0.28	1.34	1.05
Luik	0.18	0.18	-0.17	0.52	0.69
Verviers	-0.27	0.29	-0.88	0.26	1.14
Borgworm	-0.46	0.46	-1.44	0.36	1.80
Arlen	0.13	0.47	-0.87	0.98	1.84
Bastenaken	0.03	0.51	-1.06	0.94	2.01
Marche-en-Famenne	2.19	0.21	1.75	2.59	0.84
Neufchateau	1.23	0.29	0.62	1.77	1.15
Virton	-0.31	0.61	-1.64	0.77	2.40
Dinant	1.13	0.24	0.63	1.59	0.96
Namen	-0.75	0.34	-1.46	-0.14	1.32
Philippeville	0.91	0.33	0.22	1.53	1.31

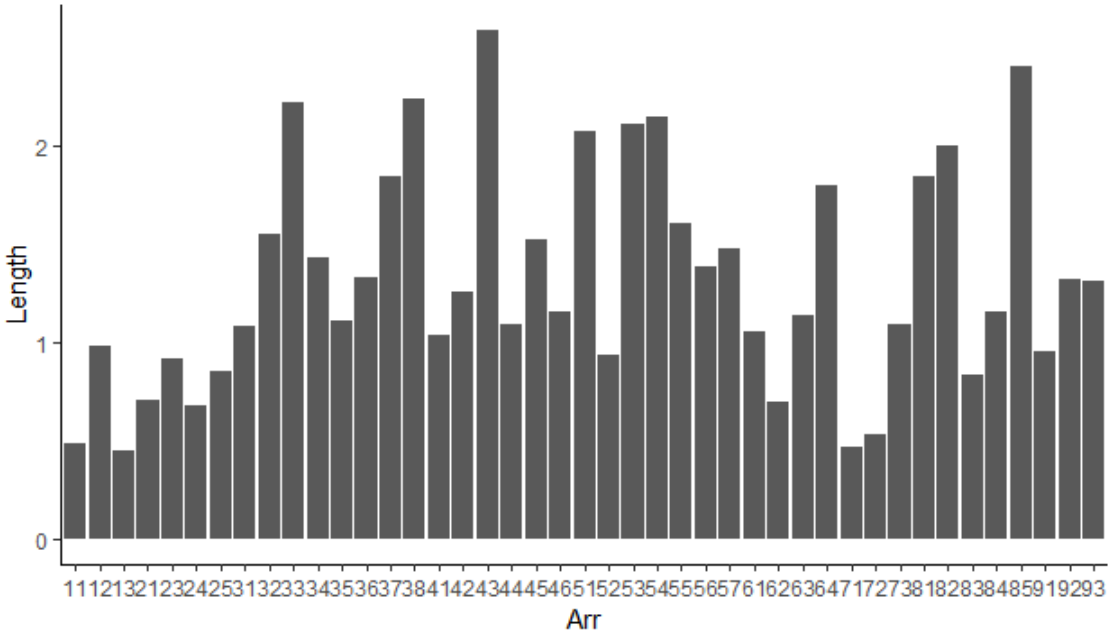


Figure 15: Length of the confidence intervals of the random effects of EM estimated by BYM (mod1)

Table 3: District specific random effects of exposure to tick bites (mod2)

District	mean	sd	lower	upper	length
Brussel-Hoofdstad	-1.99	0.14	-2.26	-1.73	0.53
Antwerpen	0.99	0.10	0.79	1.19	0.40
Mechelen	0.79	0.11	0.58	1.00	0.42
Turnhout	1.31	0.10	1.11	1.51	0.40
Hasselt	1.46	0.10	1.26	1.66	0.40
Maaseik	1.28	0.11	1.07	1.49	0.42
Tongeren	0.75	0.11	0.52	0.97	0.45
Aalts	-0.64	0.14	-0.91	-0.37	0.54
Dendermonde	0.02	0.13	-0.24	0.27	0.51
Eeklo	0.34	0.15	0.04	0.62	0.58
Gent	0.24	0.11	0.03	0.45	0.43
Oudenaarde	0.92	0.12	0.68	1.15	0.47
Sint Niklass	-0.38	0.13	-0.64	-0.12	0.53
Halle Vilvoorde	0.22	0.11	0.01	0.43	0.42
Leuven	1.02	0.10	0.81	1.22	0.41
Brugge	-0.21	0.13	-0.46	0.04	0.50
Diksmuide	-0.80	0.26	-1.34	-0.31	1.02
Leper	1.48	0.14	1.20	1.74	0.54
Kotrijk	-0.26	0.13	-0.52	-0.02	0.50
Oostende	-0.40	0.15	-0.70	-0.10	0.60
Roeselare	-0.75	0.17	-1.10	-0.42	0.67
Tielt	0.07	0.15	-0.23	0.37	0.61
Veurne	-1.11	0.28	-1.68	-0.59	1.09
Nijvel	0.89	0.11	0.69	1.10	0.42
Aat	-2.29	0.39	-3.12	-1.58	1.54
Charleroi	-1.43	0.15	-1.74	-1.14	0.60
Bergen	-1.39	0.18	-1.75	-1.05	0.70
Moeskroen	-3.40	0.68	-4.88	-2.23	2.65
Zinnik	-1.87	0.24	-2.36	-1.43	0.93
Thuin	-0.24	0.15	-0.54	0.04	0.58
Doornik	-2.41	0.33	-3.10	-1.80	1.29
Hoei	1.30	0.11	1.07	1.52	0.45
Luik	-0.08	0.11	-0.30	0.14	0.43
Verviers	0.29	0.12	0.06	0.52	0.45
Borgworm	-0.63	0.20	-1.05	-0.25	0.81
Arlen	0.16	0.17	-0.18	0.49	0.67
Bastenaken	0.51	0.17	0.18	0.83	0.65
Marche-en-Famenne	1.27	0.13	1.01	1.52	0.51
Neufchateau	1.00	0.13	0.73	1.26	0.53
Virton	1.62	0.12	1.38	1.86	0.48
Dinant	1.55	0.11	1.33	1.77	0.44
Namen	0.36	0.11	0.14	0.58	0.45
Philippeville	0.72	0.14	0.44	0.99	0.56

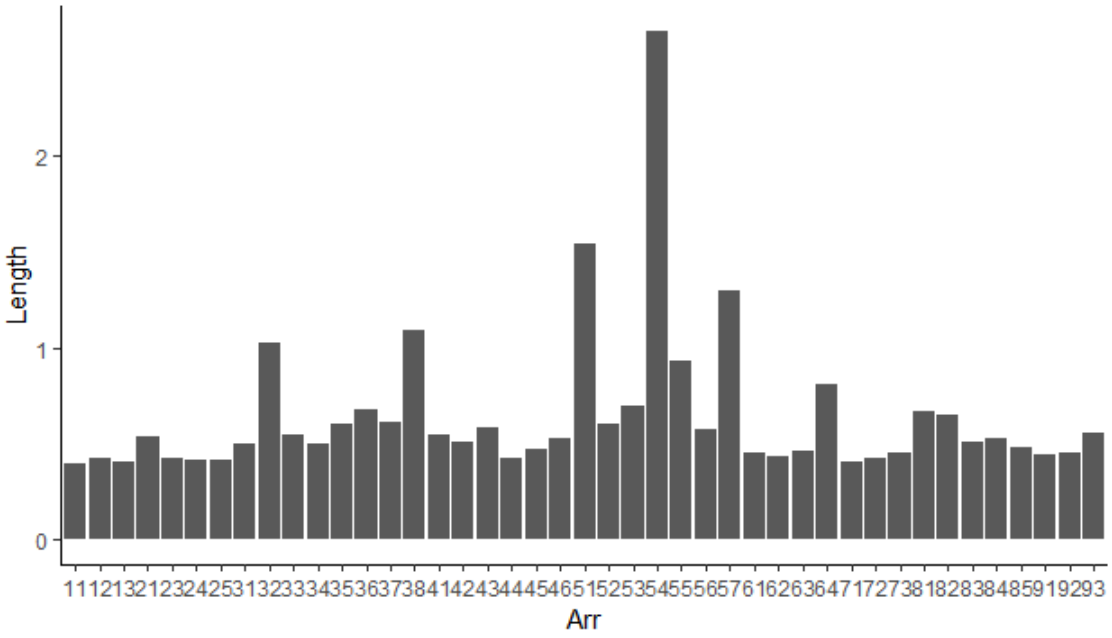


Figure 16: Length of the confidence intervals of the random effects of the exposure to tick bites estimated by BYM (mod2)

Table 4: District specific random effects positive serology tests for LB (mod1)

	mean	sd	lower	upper	length
Brussel-Hoofdstad	0.12	0.04	0.05	0.20	0.15
Antwerpen	0.53	0.03	0.46	0.59	0.14
Mechelen	0.11	0.06	-0.01	0.23	0.24
Turnhot	1.17	0.04	1.10	1.24	0.14
Hasselt	0.55	0.05	0.46	0.65	0.18
Maaseik	0.23	0.07	0.10	0.36	0.27
Tongeren	0.14	0.08	-0.01	0.28	0.30
Aalts	-1.49	0.14	-1.77	-1.22	0.54
Dendermonde	-1.32	0.15	-1.63	-1.04	0.59
Eeklo	-0.14	0.13	-0.40	0.11	0.52
Gent	-0.70	0.07	-0.84	-0.57	0.28
Oudernaarde	-1.48	0.20	-1.90	-1.11	0.79
Sink Niklaas	-0.52	0.09	-0.71	-0.34	0.37
Halle Vilvoorde	0.22	0.05	0.12	0.30	0.18
Leuven	1.74	0.03	1.68	1.80	0.12
Brugge	0.06	0.07	-0.08	0.19	0.27
Diksmuide	-0.33	0.18	-0.70	0.00	0.70
Leper	0.54	0.14	0.25	0.81	0.56
Kotrijk	-0.05	0.07	-0.19	0.09	0.28
Oostende	-1.07	0.15	-1.38	-0.78	0.60
Roeselare	-0.47	0.12	-0.70	-0.25	0.45
Tielt	-0.66	0.16	-0.98	-0.36	0.63
Veurne	-0.54	0.18	-0.92	-0.19	0.72
Nijvel	1.78	0.03	1.71	1.84	0.13
Aat	-1.06	0.20	-1.47	-0.69	0.78

Charleroi	-0.55	0.07	-0.70	-0.41	0.29
Bergen	-1.13	0.12	-1.38	-0.90	0.48
Moeskroen	-0.94	0.20	-1.35	-0.56	0.79
Zinnik	-0.36	0.10	-0.55	-0.17	0.38
Thuin	-0.16	0.10	-0.36	0.03	0.39
Doornik	-1.54	0.19	-1.94	-1.18	0.76
Hoei	-0.62	0.14	-0.91	-0.36	0.56
Luik	-1.32	0.09	-1.50	-1.15	0.35
Verviers	0.33	0.06	0.21	0.45	0.24
Borgworm	-1.04	0.20	-1.46	-0.67	0.79
Arlen	1.77	0.06	1.64	1.89	0.25
Bastenaken	1.19	0.09	1.01	1.37	0.36
Marche-ene Famenne	0.76	0.10	0.55	0.95	0.41
Neufchateau	3.04	0.04	2.96	3.11	0.15
Virton	2.29	0.05	2.19	2.40	0.21
Dinant	0.96	0.07	0.82	1.10	0.27
Namen	-0.22	0.07	-0.37	-0.08	0.29
Philippeville	0.45	0.11	0.23	0.66	0.43

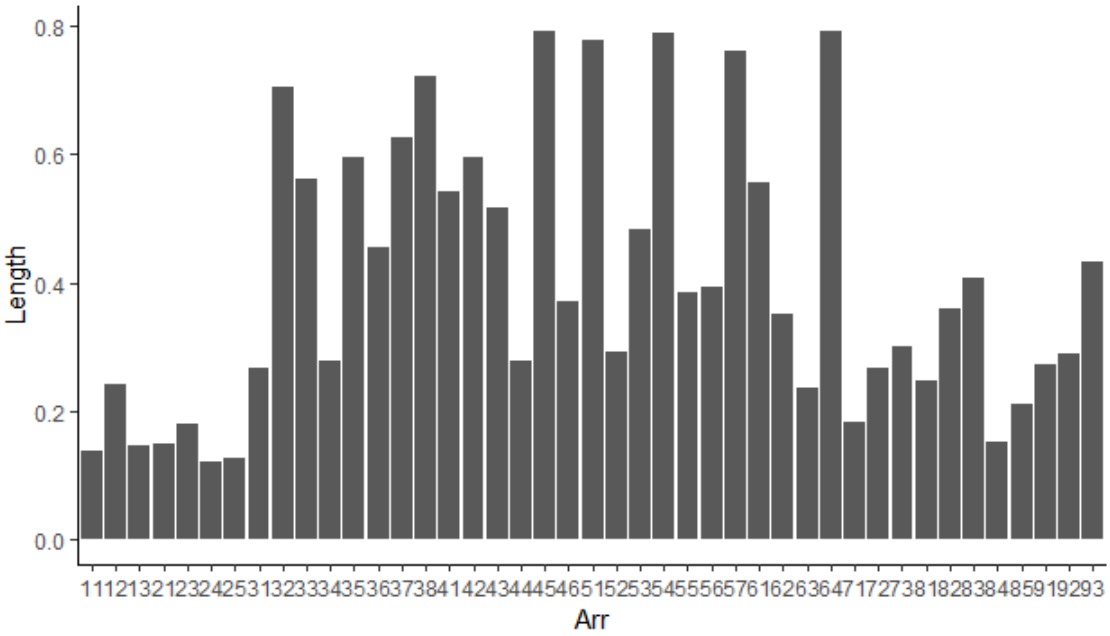


Figure 17: Length of the confidence intervals of the random effects of positive serology tests for LB estimated by BYM (mod3)

3.3 Relationship between tick bite and EM

A pre-analysis was conducted using the dataset on the SGP consultations for EM to estimate the dependence of having an EM on the number of tick bites using mixed effects poisson regression taking the effects of the districts into account. Because the time points of data collection of the cases of EM and number of tick bites are not consistent (2008-2009 and 2015-2017), time was considered as a categorical variable instead of a continuous numeric variable. The generalized linear mixed model is the most frequently used random-effects model for discrete outcomes. Conditionally on random effects b_i, it assumes that the elements Y_{it} of Y_i are independent [35]. This model also accounts for the within-subject association among the repeated measurements [36]. With generalized linear mixed effects models the joint distribution of both the vector of responses and the vector of the random effects are fully specified. As a result we can base estimation and inference on the likelihood function [36]. Based on the the data at hand with the number of EM cases as the response variable, the following GLMM model was formulated: Let Y_{it} denote count response variable/observation at time t for district i with

$$\log(\mu_{it}) = \mathbf{x}_{it}^{'}\boldsymbol{\beta} + \mathbf{z}_{it}\mathbf{b}_i,$$

where μ_{it} is the expected value of Y_{it}, the random effects $\mathbf{b}_i$ are region-specific and has a multivariate normal distribution. $\mathbf{x}_{it}$ is the column vector of values of covariates for fixed effect model parameters $\boldsymbol{\beta}$. The main interest for this analysis is the estimates of the fixed effects exp(tick)=exp(0.033052)=1.033. The results shows that a unit increase in the number of tick bites increases the expected number EM by 3.3% (controlling for the period and the random district effects). The result also shows that there was an insignificant effect of time on the cases of EM.

```
   AIC      BIC   logLik deviance df.resid
   705.3    727.4   -345.7    691.3      165
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.422536   0.128290   3.294 0.000989 ***
tick         0.033052   0.004157   7.951 1.85e-15 ***
year2009     0.143294   0.116159   1.234 0.217354
year2015    -0.029266   0.121055  -0.242 0.808965
year2016     0.025072   0.120683   0.208 0.835421
year2017     0.051077   0.129597   0.394 0.693492
```

3.4 Correlation tests

The relatives risks from the three disease mapping models were extracted and a dataset with three variables was formed. Let x = relative risks of EM cases from mod1, y =relative risks of exposure to tick bites from mod2 and z = the relative risks of positive serology results for LB from mod3. All variables maintained the same order from the areal graph. To answer the question on the representativeness of the sentinel net work of general practitioners, a correlation analysis was done between the distributions of the relative risks from the disease mapping models ($Corr_{xy}$). This was considered because of the high dependence that EM has on the number of tick bites. A correlation analysis was also done between the distributions z and y ($Corr_{zy}$). Scatter plots xy and zy were made to visually check the data as shown in figures 18 and 19. To carry out a pearson correlation test, the variables x, y and z were first checked for normality. Visual inspection of data normality using Q-Q plots (quantile-quantile plots) was also done. Q-Q plot draws the correlation between a given sample and the normal distribution. The normality plots are reported in the appendix (figures 23a, 23b and 23c shows Q-Q plots for x, y and z respectively). Tests for normality using Shapiro-Wilk normality test gave p-values less than 0.05 for all the three variables. All the variables were not normally distributed. Spearman rank correlation test was then used to assess the correlation between the relative risks. Spearmans correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. The p-value of the correlation test $Corr_xy$ was 0.0001 with correlation coefficient of 0.56. And the the p-value of the correlation test $Corr_{zy}$ was 1.029e-06 with correlation coefficient of 0.67. Both tests have p-values less than the significance level alpha = 0.05. We can conclude that relative risk of EM cases (SGP) and relative risk of the exposure to tick bites (TekenNet) are significantly correlated and the relative risk of positive serology results for LB (SNL) and relative risk of the exposure to tick bites (TekenNet) are also significantly correlated.
The scatter plots also showed some extreme cases of outlying values. Analysis of outliers was not consider

because of the nature of the data. Each pair of points in the scatter plot represent a district in Belgium and the value of these points compares the risks of Lyme for that district modelled from two different data sources. Deleting outliers to test for improvement of correlation was not necessary in this case.

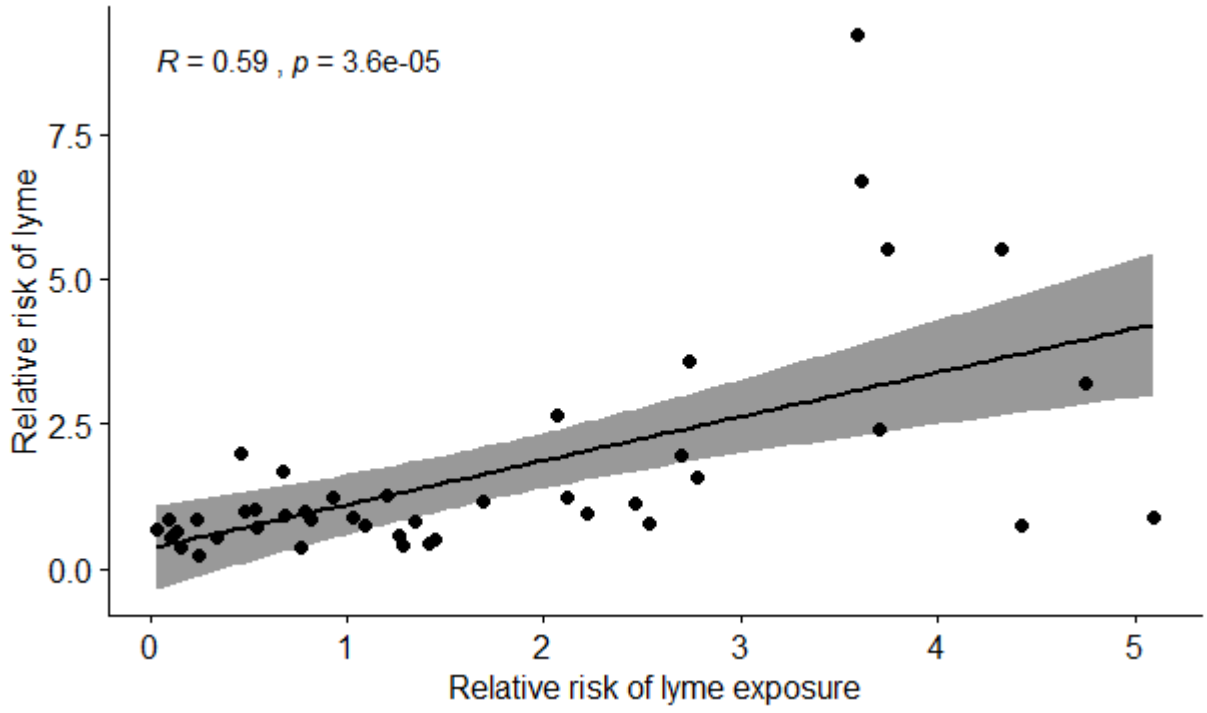


Figure 18: Correlation plot between the relative risk of EM and the relative risk of exposure to tick bites

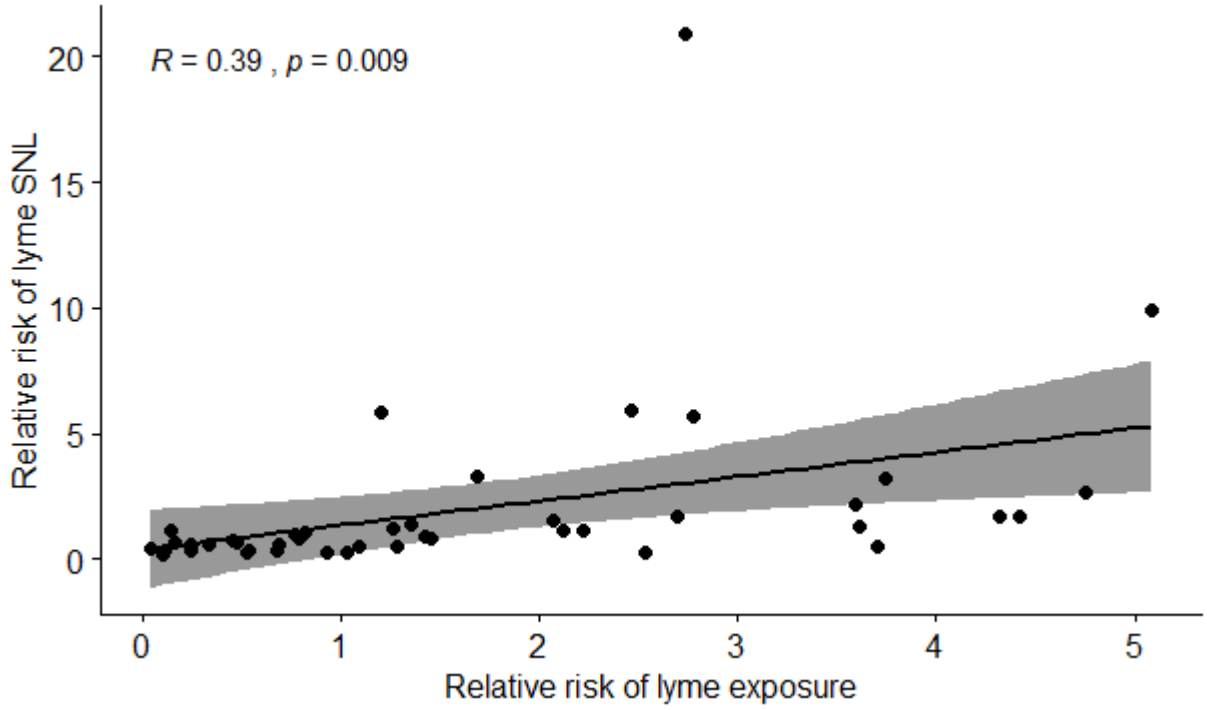


Figure 19: Correlation plot between the relative risks of positive serology test for LB and the relative of exposure to tick bites

4 Discussion

Using data from INAMI-RIZIV, it was found that test coverage of the SNL was stable between 2011 and 2015 with an average coverage of over 60%. At the regional level, Wallonia had the lowest average test coverage of 44%, Flanders had a test coverage of 75% and Brussels had a test coverage of over 95%. At the provincial level, there was larger variation in coverage over time. In 2010, the lowest number of reimbursement tests was conducted in provinces like Namur, Liege and Limburg all having a coverage of below 10%. High risk areas of lyme where identified by using the data on positive serology tests conducted by the SNL at the provincial level. Luxemburg and Walloon Brabant had the highest cumulative incidence followed by Flemish Brabant and Antwerpen. BYM model was used to locate high risks areas of LB at the district level. All the districts of Luxemburg, Walloon Brabant, Flemish Brabant and Antwwerpen were identified as high risk areas by the BYM model. Except for Walloon Brabant, all the other high risk areas had coverage of above 50%. Brussels with a very high test coverage was identified as an area with low risk of LB. Also most of the districts in west Flanders had very low risks of LB disease. Looking at the high risk areas identified by both the cumulative incidence of positive serology results and the corresponding relative risks from BYM model, it can be said that the SNL is representative in these areas since tests coverage was on average greater than 50%. Seasonal variations, outbreaks and all sorts of variations in the occurrence of LB are very likely to be captured by the surveillance system if they occur at the national or regional level. In the provinces of Namur Walloon Brabant, Liege and Limburg, the SNL is less likely to capture changes over time.

The relative risks of the occurrence of LB at the district level compared to the whole of Belgium using data from counts of positive serology results for LB, EM cases and the number of tick bites were estimated using the BYM model. The parameter estimates of the random effects from the models are also reported. Representativeness of SGP on the occurrence of EM was assessed by comparing the distributions of the relative risks of the exposure to tick bites and relative risk of of EM. This was done because TekenNet is assumed to be used by all those at risk of having Lyme (it is used as a proxy to estimate the risk of LB) while SGP data is restricted to all those who made consultations to general practitioners belonging to the sentinel network (only a part of those at risk of LB). Thus if there is a similar pattern in the relative risks from both distributions, it can be concluded that the SGP is representative for the occurrence of LB. This was done at the district level (43 districts) because it can easily be modelled using INLA. There was positive correlation between the relative risks from the model with data from TekenNet and model with data from SGP. Based on a correlation between the two distributions, it can therefore be concluded that the SGP is overall representative for the occurrence of LB. LB is a geographically specific disease with very high risks in some regions and low risks in other regions. The results showed that the SGP is representative for both high and low risk areas of LB.

The variability of the test coverage of SNL observed over time in some provinces could be a result of the changes in the participation status of laboratories and/or non-similarity between available accreditation numbers of laboratories and those used in the INAMI-RIZIV reimbursement database. This for example can likely be a reason for the sharp decrease in test coverage in Namur and generally lower test coverage in Wallonia in 2010. The results of test coverage at the regional and provincial levels are comparable to a previous study on assessing the sensitivity and representativeness of the Belgian Sentinel Network of Laboratories using test reimbursement data [37].

This study used the Integrated Nested Laplace Approximation for the spatial (disease mapping) models. Even though this model is very popular in applied research especially in epidemiology, its general complexity remains a fundamental issue for the implementation particularly in the Bayesian approach. Generally, the INLA approach is able to provide reliable estimations in lower computational time than its corresponding MCMC-based estimation. One of the main difference between MCMC and INLA methods is that MCMC provide (asymptotically) exact inference, while INLA by definition give an approximation to the relevant posterior distribution. INLA performs just as well as it MCMC counterparts in many applied cases especially when MCMC are considered in their standard implementations [34]. Since INLA is a recent approach, it is less established than MCMC. Consequently, there is still ongoing development particularly with respect to some more advanced features. However, it is important to notice at the same time that the increasing popularity of INLA is generating a number of contributed add-ons that is able to extend the build-in facilities of the package R-INLA. Given these characteristics, INLA is considered as a valuable addition to the Bayesian statistician toolkit.

This study has several limitations both at the level of data collections and the assumptions made. To begin with, part of the study relies on the quality and exhaustively of the reimbursement data, without being able to evaluate it. Laboratories where identified using certification number for reimbursement

with multiple laboratories associated or belonging to the same group that may share the same laboratory code. Also not all laboratories from the group participate in the surveillance system. It was assumed in this study that if one laboratory participated in the SNL, then all laboratories in the same group with the same identification code were also participating. In some cases, individuals might perform multiple test to diagnose LB. If these repeated tests are not randomly distributed between laboratories belonging to the network and other laboratories, it might lead to an overestimation of the representativeness of the SNL. However, the SNL is able to identify duplicates amongst the reported cases which may solve the problem of multiple diagnosis by a patient. Even though the SNL is considered to be stable and was established over 30 years ago, there is still problem of interpretation of positive results since LB is very complex. SNL gives a partial picture of the incidence since laboratory tests are not recommended for patients with an EM. Finally laboratory tests only confirm the presence of anti-Borrelia antibodies which does not necessarily mean that the patient is suffering from LB: this could also be due to a previous symptomatic or asymptomatic Borrelia infection.

Limitations based on the SGP concerns the estimation of EM itself. The SGP in Belgium is based on a voluntary participation of GPs who have to declare cases actively. Underreporting is possible, especially in districts that are less affected by LB due to lower awareness [38]. The estimation of the risk of EM in this study is based on data for a very short period of time. Also the data were not for a continuous time period and yearly fluctuations in tick bites due to climate change can lead to wrong estimations of the risk of EM.

For the tick bites data set (TekenNet), notifications depends only on the awareness and knowledge of the existence of the website. Data was available only from July 2015 to June 2017, results in insufficient data to estimate the risk of exposure to LB. The tick bites data was used as a proxy to estimated the risk of exposure of LB even though the data source is not exhaustive.

This study evaluated just a restricted aspect of the surveillance system; representativeness and it indicate that by design, the SNL has sufficient precision and reflects without any important systematic bias the Belgian situation. The SNL has many other characteristics that has an influence on the capacity on the surveillance system to effectively monitor infectious diseases. In order to have a better understanding of the actual quality of the data reported by the SNL, a wider evaluation of the surveillance system is recommended [5]. The impact of changes in the network characteristics should be systematic documented and continuously assessed because of the evolving nature of the SNL. It is thus necessary to obtain reimbursement data by laboratory not by district or province as was the case of this study.

This study has illustrated the importance of evaluating the representativeness of a surveillance system and it showed that the available reimbursement data could serve as a tool for improving representativeness of surveillance systems in the absence of other data sources.

5 Conclusion

Based on results of test coverage from SNL, this suggests that it is representative and capable of describing epidemiological situations of LB at the national and regional levels. There is much variability at the provincial level, thus not conclusive for some provinces like Namur, Liege, Walloon Brabant and with very low test coverage and much variability over time. The risk areas of LB as identified by the cumulative incidence at the provincial level and relative risks by BYM model at the district level shows that some areas at high risk are sufficiently covered by the SNL while other areas at low risks has low coverage. But some high risk areas identified are not sufficiently covered by the SNL (the case of Wallon Brabant). However, the SNL should be reinforced in three out of the eleven province (Namur, Walloon Brabant Liege) so as to be used as a sensitivity alert system at the provincial level.
For the SGP, the results from the BYM models and correlation tests suggests that the occurrence of EM is representative at the district level and in the whole of Belgium. Based on the distributions of the relative risks of LB and their posterior probabilities, the distribution of LB is geographically too specific and is not evenly distributed in the whole of Belgium. The results from the data sources (SGP and TekenNet) showed that the surveillance system is representative both at the high and low risk areas of LB in Belgium. Thus to conclude, the occurrence of EM and LB in Belgium are on average sufficiently represented by both the sentinel network of laboratories and the sentinel general practitioners.

Acknowledgement

Firstly, I want to thank the university of Hasselt and the administration for giving me the opportunity to study MSc Biostatistics. I want to appreciate the endless efforts of my former and new internal supervisors; Yannick Vandendijck and Pietro Coletti respectively, who has always been there to direct, correct and advise at every level during the process of doing my thesis. Much thanks goes to my external supervisor Tinne Lernout from Sciensano who didn't only provided the thesis project and all the data used but was ever present to discuss, give highlights and correct me from the start to the end of this thesis. I also want to give an immense thanks to my family: my partner Ann and my two sons Sander and Axel who during this period of my study has being deprived on several occasions of my presence and commitments. To all my friends, mates and professors who helped me in one way or the other to see that I go through my studies, I say thank you. Above all, I want to give a big thank you to God who gave me the strength, protection and guidance to see me through this tough period.

References

1 Langmuir, A.D. The surveillance of communicable diseases of national importance. *New England journal of medicine,* 268: 182-192 (1963).

2 Thacker, S. B. et al. The surveillance of infectious diseases. *Journal of the American Medical Association,* 249: 1181-1185 (1983)

3 Stephen B. et al. A method for evaluating systems of epidemiological surveillance (1988).

4 Kimball, A. M. et al. Shigella surveillance in a large metropolitan area: assessment of a positive reporting system. *American journal of epidemiology*, **70**: 164-166 (1980)

5 Walckiers D, Stroobant A, Yourassowsky E, Lion J, Cornelis R. A sentinel network of microbiological laboratories as a tool for surveillance of infectious diseases in Belgium. *Epidemiol Infect.* 106(2):287-303 (1991).

6 BleyenheuftC, Lernout T, Berger N, Rebolledo J, Leroy M, Robert A, Quoilin S. Epidemiological situation of lyme borreliosis in Belgium, 3003 to 3012. *Arch Public Health.* 73(1):33 (2015).

7 European Center for Disease Prevention and Control. Data quality monitoring and surveillance system evaluation. *A handbook of methods and applications. Stockholm*: ECDC (2014)

8 Gerold S, Gary P, Jeremy G, Franc S. Lyme borreliosis. *Seminar* (2012).

9 Estrada- Pena A. Tick-borne pathogens, transmission rates and climate change. *Front Biosci.* 14:2674-87 (2009)

10 Hubalek Z. Epidemiology of Lyme borreliosis. *Curr Probl Dermatol.* 37: 31-50 (2009).

11 Keesing F, Brunner LK, Duerr S, Killilea M, Logiudice K, Schmidt K, et al. Hosts as ecological traps for the vector of Lyme disease. *Proc Biol Sci.* 276(1675:3911-9 (2009).

12 Keesing F, Beldan LK, Daszak P, Dobson A, Harvell CD, Holt RD, et al. Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature.* 468(7324):647-52 (2010).

13 Stanek G, Strle F. Lyme borreliosis: a European perspective on diagnosis and clinical management. *Curr Opin Infect Dis.* 22(5):341-60 (2009).

14 Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics* 8(2): 158-183 (2007).

15 Gelfrand A, Diggle P, Fuentes M, Guttorp P, editors. Handbook of spatial statistics. *Chapman and Hall* (2010).

16 Banerjee S, Carlin B, Gelfand A. Hierarchical modelling and analysis for spatial data. Monographs on statistics and applied probability. *New York: Chapman and Hall* (2004).

17 Berry S, Carlin B, Lee J, Muller P. Bayesian adaptive methods for clinical trials. *CRC Chapman and Hall* (2011).

18 Baio G. Bayesian methods in health economics *CRC Chapman and Hall* (2012).

19 Jackman S. Bayesian analysis for the social sciences. *Wiley-Blackwell.*

20 Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 35:765-75 (2006).

21 Lawson A. Bayesian disease mapping. Hierarchical modelling in spatial epidemiology *CRC Press* (2009).

22 Brooks S, Gelman A, Jones G, Meng X, editors. Handbook of Markov chain Monte Carlo. *CRC Press, Taylor and Francis Group* (2011).

23 Lund D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. *Stat Med*, 28(25):3049-67 (2009).

24 Rue H, Martino S. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *J R Stat Soc B* 71(2):1-35 (2009).

25 Li Y, Brown P, Rue H, al Maini M, Fortin P. Spatial modelling of lupus incidence over 40 years with changes in census areas. *J R Stat Soc C (Appl Stat)*, 61(1):99-115 (2012).

26 Martino S, Rue H. Implementing approximate Bayesian using integrated nested laplace approximation: a manual for *INLA* program; Available from: http//www.math.ntnu.no/hrue/GMRFsim/manual.pdf (2010).

27 Fahrmeir L, Lang S. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C* 50(2): 201-220 (2001).

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

29 Banerjee S, Carlin B, Gelfand A. Hierarchical Modelling and Analysis for Spatial Data. *Chapman and Hall CRC, London* (2004).

30 Rue H, Held L. Gaussian Markov Random Fields. *Chapman and Hall CRC, London* (2005).

31 Besag J, York J, Mollie A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1): 1-20 (1991).

32 Tierney L, Kadane J. Accurate approximations for posterior moments and marginal densities. *J Am Stat Assoc* 393(81):82-6 (1986).

33 Martins G, Simpson D Lindgren F, Rue H. Bayesian computation with INLA: new features. *Norwegian University of Science and Technology Report* (2012).

34 Blangiardo M, Cameletti M. Bayesian Spatio and spatio-temporal models with R-INLA. *Wiley* (2013).

35 Molenberghs, G. and Verbeke, G. (2005), Models for Discrete Longitudinal Data, New York: Springer.

36 Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). Applied Longitudinal Analysis. Hoboken, NJ: Wiley.

37 Nicolas B, Gaetan M, Yves D and Sophie Q. Assessing the sensitivity and representativeness of the Belgia Sentinel Network of Laboratories using test reimbursement data, *BioMed Central* (2016).

Appendix A: Posterior probabilities in the Disease mapping models

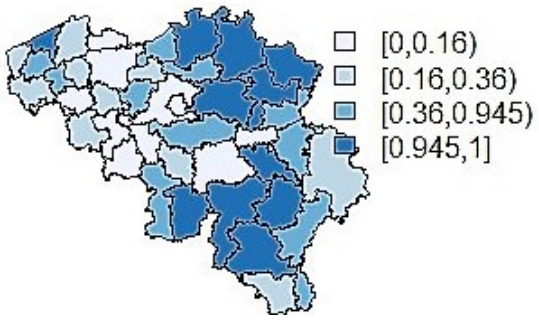


Figure 20: Distribution of the district specific posterior probability in mod1

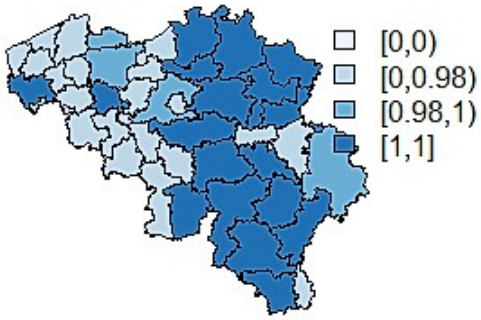


Figure 21: Distribution of the district specific posterior probability in mod2

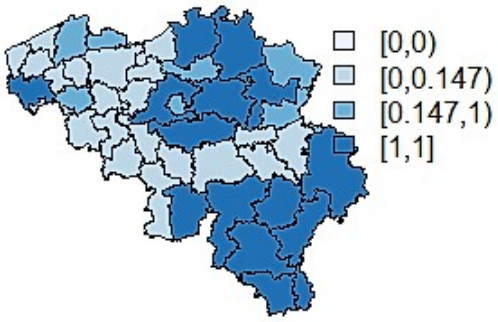


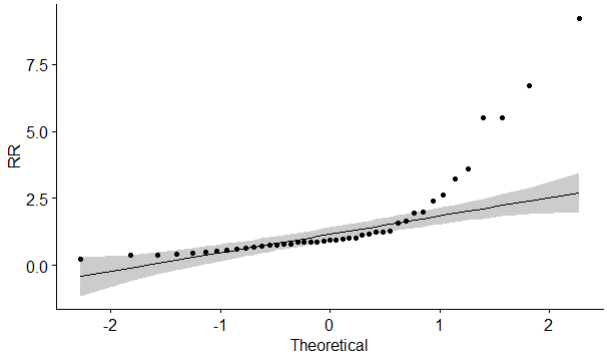
Figure 22: Distribution of the district specific posterior probability in mod3

Appendix B: Belgium districts and their codes

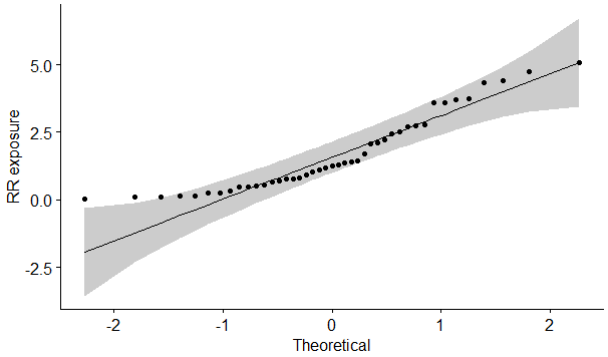
Table 5: Belgium Districts and their codes

District	District code
Brussel-Hoofdstad	21
Antwerpen	11
Mechelen	12
Turnhot	13
Hasselt	71
Maaseik	72
Tongeren	73
Aalts	41
Dendermonde	42
Eeklo	43
Gent	44
Oudernaarde	45
Sink Niklaas	46
Halle Vilvoorde	23
Leuven	24
Brugge	31
Diksmuide	32
Leper	33
Kotrijk	34
Oostende	35
Roeselare	36
Tielt	37
Veurne	38
Nijvel	25
Aat	51
Charleroi	52
Bergen	53
Moeskroen	54
Zinnik	55
Thuin	56
Doornik	57
Hoei	61
Luik	62
Verviers	63
Borgworm	64
Arlen	81
Bastenaken	82
Marche-ene Famenne	83
Neufchateau	84
Virton	85
Dinant	91
Namen	92
Philippeville	93

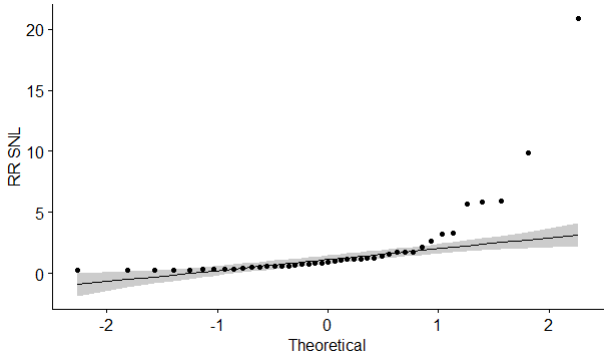
Appendix C: Q-Q plots to check for Normality assumption of the distributions of the relative risks from the three disease mapping models



(a) Normality plot of the distribution of x



(b) Normality plot of the distribution of y



(c) Normality plot of the distribution of z