A new centered spatio-temporal autologisticregression model with an application to local spread of plant diseases

Anne Gégout-Petit, Lucia Guérin-Dubrana, Shuxian Li

A new centered spatio-temporal autologistic regression model with an application to local spread of plant diseases

Anne Gégout-Petit[a], Lucia Guérin-Dubrana[b,c], Shuxian Li[b]

[a] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France,*
anne.gegout-petit@univ-lorraine.fr
[b] *Université de Bordeaux, ISVV, UMR-1065 INRA, France*
[c] *Bordeaux Sciences Agro, Gradignan, FRANCE*

Abstract

We propose a new spatio-temporal autologistic centered model for binary data on a lattice. Centering allows the self-regression coefficients to be interpreted by separating the large-scale structure from the small-scale structure. One of the coefficients determines the overall level (or average) of the process, the second determines the spatial autocorrelation. We discuss the existence of the joint distribution of the process and carry out numerical studies to highlight the interest of this type of centering. We suggest using the estimator that maximises the Pseudo Likelihood (denoted Maximum Pseudo Likelihood Estimator (MPLE) in the following) and we give a method for choosing the neighbourhood structure. We run simulations studies that show that the estimation method and model selection method work well. The method is applied to model and fit epidemiological data on Esca disease in a vineyard in the Bordeaux region.

Keywords: spatial-temporal modeling, large-scale model structure, binary response, Maximum Pseudo-Likelihood Estimation, Autologistic Model

1. Introduction

Since spatial and spatio-temporal data are commonly present in nature, the modelling of such data has increased the interest of many scientists from various fields such as ecology, epidemiology and image analysis. Binary data is of particular interest for modelling the occurrence of an event such as disease or death. Spatio-temporal and spatial binary data models are quite adequate to study the evolution of a known plant disease on a grid if propagation and interaction between neighbours is suspected.

40 years ago, Besag (1974) firstly presented an autologistic model for spatial binary data, assuming a simple dependence on surrounding neighbours. This model has been shown to be useful and then was extended by Gumpertz et al. (1997); Huffer and Wu (1998) to take into account the effects of some covariates.

More recently, Zhu et al. (2008) and Zheng and Zhu (2008) generalised autologistic regression models to simultaneously account for covariates, spatial covariates, and time dependence for binary data that are measured repeatedly over time on a grid.

However it is difficult to interpret the parameters in a non-centered autologistic regression model. This problem has been first pointed out by Caragea and Kaiser (2009) who presented a centered parameterisation for an autologistic spatial regression model to overcome interpretation problems. Following this work, Hughes et al. (2011) discussed in more detail the estimation of parameters and random field simulations from an centered autologistic model on a lattice. Then, Wang and Zheng (2013) presented a centered spatio-temporal autologistic regression model. They used and compared several

methods for estimating model parameters and coefficients. The drawback of this spatio-temporal model is that the temporal dependency is not causal and the state of a point at time t is linked with its state at time $t-1$ and $t+1$. This model has good mathematical properties but is not useful for interpretation by the practitioner. A review of the literature on autologistic models applied to binary data was recently given by Zhu and Zheng (2016).

In this paper, we present a new centered spatio-temporal autologistic model which depends only on the past. We will show, on simulated data, the advantage of this kind of centering over other autologistic models; indeed, we will see that, unlike the other models, it gives the expected average of the process. Since the joint distribution of such models is very complex to write, we recommend an estimation method based on the Maximisation of the Pseudo-Likelihood (MPL). We present a method for choosing the neighbourhood structure by using the value of the Pseudo-Likelihood. We apply the model for a better understanding of the spread of esca grapevine disease in a vineyard. We used leaf symptom data recorded in a vineyard in the Bordeaux region from 2004 to 2017.

The paper is organised as follows: in this section we first give the formalism to define random field, neighbours structure and autologistic models and we review the literature on spatial and spatio-temporal autologistic models. In Section 2 we present our new centered autologistic model and discuss the existence of the joint distribution of the spatio-temporal process. Then, in Section 3, we present several simulations results to compare the spatio-temporal autologistic model depending on the past under different centered parametrisations. In section 4, we present an inference algorithm to calcu-

late the Maximum Pseudo-Likelihood Estimator of our model. We evaluate it by calculating parameter estimates on simulated data. We also present a method for choosing between possible neighbourhood structures. We study its performance by comparing the selected neighbourhood structures on simulated data with the actual structure. Section 5 presents an analysis of real data on the spread of a disease in a vineyard. In the last section, we discuss the interest of the methodology and give some perspectives on this study.

1.1. Spatial autologistic models

Let $[Z]$ denote the distribution of random variable Z. Let $\mathbf{Z} = \{Z_i : i = 1, ..., n\}$ be the random field where $Z_i \in \{0, 1\}$ represents the state at the ith point s_i of a lattice $S = \{s_1, \ldots, s_n\}$. The distribution $[Z]$ is given by the conditional laws

$$[Z_i | Z_j, j \neq i] \sim Binary(p_i),$$

where for $1 \leq i \leq n$, $p_i = \mathbb{P}(Z_i = 1 | Z_j, j \neq i)$.

In addition, we have a neighbourhood structure: we assume that each position s_i is associated with a set N_i containing the neighbours of s_i. We suppose that this relation is symmetric that is $s_j \in N_i \Leftrightarrow s_i \in N_j$. This neighbourhood structure defines a non-oriented graph whose nodes are the locations s_i's and there is an edge between s_i and s_j if $s_j \in N_i$. From now on, we will use an notation abuse and confuse i and s_i in the formulae by using "$j \in N_i$" instead of "j such that $s_j \in N_i$".

If moreover we assume that the conditional distributions of the random

field $\mathbf{Z}$ satisfy the following Markov property:

$$[Z_i|Z_j, j \neq i] = [Z_i|Z_j, j \in N_i] \quad \text{for all} \;\; 1 \leq i \leq n, \tag{1}$$

then $\mathbf{Z}$ is said to be a Markov random field associated with the neighbour structure given by the N_i's. The conditional binary probability can be expressed in exponential family form:

$$\mathbb{P}(Z_i|Z_j, j \in N_i) = \frac{\exp\left(Z_i A_i(Z_j, j \in N_i)\right)}{1 + \exp\left(A_i(Z_j, j \in N_i)\right)} \tag{2}$$

where A_i is called a natural parameter function. When the Z_i's take values in $\{0, 1\}$ and that will be the case in this paper, the use of the logit function instead of the A_i to model the p_i is very common; it is defined by the following equation:

$$A_i(Z_j, j \in N_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \log\left(\frac{\mathbb{P}(Z_i = 1|Z_j, j \in N_i)}{\mathbb{P}(Z_i = 0|Z_j, j \in N_i)}\right).$$

Besag (1974) showed that the natural parameter functions must be of form:

$$A_i(Z_j, j \in N_i) = \text{logit}(p_i) = \alpha_i + \sum_{j \in N_i} \rho_{ij} Z_j, \tag{3}$$

with α_i a leading constant. Note that the regression coefficients ρ_{ij} may depend on site s_i and its spatial neighbour s_j. The variables Z_i's are independent from each other if for each $1 \leq i \leq n$, $\mathbb{P}(Z_i = 1|Z_j, j \neq i)$ does not depend on the Z_j. From Equations 2 and 3, we can easily see that it is equivalent to $\rho_{ij} = 0 \quad \forall (i, j) \in S \times S$ with $S = \{1, \ldots, n\}$. It means that the parameters ρ_{ij}'s reflect the dependences within the lattice. An obvious question is the existence of the joint distribution of the spatial process

$\{Z_i : i = 1, ..., n\}$ that is only defined through the conditional probabilities $[Z_i|Z_j, j \in N_i]$. Coefficients ρ_{ij} must satisfy certain restrictions for a joint distribution of the Z_i's to exist (Gaetan and Guyon, 2008); in particular, Besag (1974) showed that the symmetry condition $\rho_{ij} = \rho_{ji}$ is necessary for the joint distribution to exist.

The modelling can be generalized to include covariates $\mathbf{X} = \{\mathbf{X}_i, i = 1, ..., n\}$. In this case, the natural parameter function is given in Caragea and Kaiser (2009) by:

$$\text{logit}(p_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j \in N_i} \rho_{ij} Z_j, \tag{4}$$

with $\boldsymbol{\beta}$ is a k-vector of regression parameters. Caragea and Kaiser (2009) discussed the difficulties in interpreting model parameters for a non-centered autologistic model given by Equation 4. Indeed, let $p_i = \mathbb{P}(Z_i = 1|Z_j, j \in N_i, \mathbf{X}_i)$ so that the odds that $Z_i = 1$ in model (4) is $p_i/(1 - p_i)$. Let $c_i = \frac{exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1+exp(\mathbf{X}_i^T \boldsymbol{\beta})}$ being the probability of occurrence under the spatial independence model (when all the ρ_{ij}'s equal 0), the odds that $Z_i = 1$ is $c_i/(1 - c_i)$. Then the log odds ratio for model (4) relative to the independence model is:

$$log\left(\frac{p_i/(1-p_i)}{c_i/(1-c_i)}\right) = \sum_{j \in N_i} \rho_{ij} Z_j$$

In this case, the odds of $Z_i = 1$ in model (4) relative to the independence model increases for any nonzero neighbours, and can never decrease. This is not reasonable if most of neighbours are zeros and could bias the realisations towards 1.

To overcome these difficulties of interpretation, a centered spatial autologistic

model was introduced by Caragea and Kaiser (2009). In this model, the value of the Z_j's in the regression is centered by their expected "large-scale" value and the p_i's are given by:

$$\text{logit}(p_i) = \mathbf{X}_i^T\boldsymbol{\beta} + \sum_{j \in N_i} \rho_{ij} \left(Z_j - \frac{\exp(\mathbf{X}_j^T\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_j^T\boldsymbol{\beta})} \right). \tag{5}$$

The authors of Caragea and Kaiser (2009) pointed out that model (5) is similar to the parametrisation customarily used for auto-Gaussian models. They show that the alternative (centered) parameterization overcomes the difficulty of interpretation of the non centered model.

In the model given by Equation 5, the term $\mathbf{X}_i^T\boldsymbol{\beta}$ (called "large-scale model component by Caragea and Kaiser (2009)), determines the expected value of p_i. On average over the entire field, $\frac{\exp(\mathbf{X}_j^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_j^T\boldsymbol{\beta})}$ corresponds to the proportion of Z_i that are equal to 1. Variances, covariances and other high-order portions of the data structure are determined by the second term of Equation 5, Caragea and Kaiser (2009) called it the small-scale model component. About the modelling of Markov random field and the interpretation of the coefficients of $\text{logit}(p_i)$ in term of small or large scale-structure, see Kaiser and Cressie (2000) or Cressie (1993) p.114. In Caragea and Kaiser (2009), the parameters were estimated by MPL.

Hughes et al. (2011) focused on the methods for estimating the coefficients and parameters in the centered autologistic model. They used MPLE and also parametric bootstrap, Monte Carlo Maximum Likelihood (MCML) and MCMC Bayesian approaches to infer the parameters. They also discussed ways to optimise the effectiveness of their algorithms. They also compared the performance of the three approaches in an in-depth simulation study.

They found that "inference for regression parameters in the centered model is reliable only for reasonably large lattices ($n > 900$) and no more than moderate spatial dependence". They recommended the MPLE for its easier implementation and much faster execution. A package for the free software R is avalable estimating the centered spatial models parameters (Hughes, 2014). More recently, Wolters (2017) discussed coding and centering in spatial autologistic models and he recalled the good properties of MPLE for estimating non-centered models parameters. It is not always true in the centered case because PL can exhibit multiple local optima.

1.2. Spatio-temporal autologistic models

Now let $\mathbf{Z}_t$ to denote a random field indexed by discrete time t. Z_{it} for $i = 1, ..., n$ and $t \in \mathbb{Z}$ is a random binary variable indexed by position s_i and time t. And the covariates $\mathbf{X}$ are k-vectors indexed by i and t.

Zhu et al. (2005) generalised the autologistic regression models to account for covariates, spatial dependence, and temporal dependence simultaneously. The model specifies the joint distribution of $\{\mathbf{Z}_t : t \in \mathbb{Z}\}$ by a family of conditional distributions:

$$\begin{aligned} &\mathbb{P}(\mathbf{Z}_{t_1}, ..., \mathbf{Z}_{t_2} | \mathbf{Z}_t; t \neq t_1, ..., t_2) \\ &\propto \quad exp\left\{ \sum_{t'=t_1}^{t_2} \left[\sum_{i=1}^{n} \mathbf{X}_{i,t'}^T \boldsymbol{\beta} Z_{i,t'} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j \in N_i} \rho_1 Z_{i,t'} Z_{j,t'} \right] + \sum_{t'=t_1}^{t_2+1} \sum_{i=1}^{n} \rho_2 Z_{i,t'} Z_{i,t'-1} \right\}. \end{aligned} \tag{6}$$

for all $t_1, t_2 \in \mathbb{Z}^2$ such that $t_1 < t_2$, where $\mathbf{X}_{i,t}$ is the k-vector of the covariates at site s_i and time t. Note that the specification is consistent for all $t_1 < t_2$, and the joint distribution of $\{\mathbf{Z}_t : t \in \mathbb{Z}\}$ can be shown to exist by Theorem 2.1.1 of Guyon (1995). If $N_{i,t} = \{(j,t) : j \in N_i\} \cup \{(i, t-1), (i, t+1)\}$

denotes the neighbourhood set for the position s_i and the tth time point, the full conditional distribution of the model is

$$[Z_{i,t}|Z_{i',t'} : (i',t') \neq (i,t)] = [Z_{i,t}|Z_{i',t'} : (i',t') \in N_{i,t}] \quad \text{and}$$

$$\begin{aligned} &\text{logit}(\mathbb{P}(Z_{i,t} = 1|Z_{i',t'} : (i',t') \in N_{i,t}; \mathbf{X})) \\ &= \mathbf{X}_{i,t}^T \boldsymbol{\beta} + \sum_{j \in N_i} \rho_1 Z_{j,t'} + \rho_2 (Z_{i,t-1} + Z_{i,t+1}). \end{aligned} \tag{7}$$

Note that in model (6), the coefficients corresponding to the ρ_{ij} of Equations (3), (4), (5), are all the same and they equal ρ_1 meaning that all the spatial neighbour's relations have the same intensity. The coefficient ρ_2 corresponds to a "temporal" autoregression, which is the same regardless of spatial position and time. On the other hand, Zheng and Zhu (2008) pointed out that one drawback of the parameter estimation for the spatio-temporal autologistic regression model presented by Zhu et al. (2005), was based on MPLE whose statistical efficiency is not well established in the centered case. Zheng and Zhu (2008) suggest a fully Bayesian approach and compared it to estimation via MPL or MCMC Maximum Likelihood. Another drawback is the difficulty of using model (7) for a real application. Indeed it seems unrealistic to model the probability of an event occurring in terms of the future.

It is probably why Zhu et al. (2008) developed a spatio-temporal autologistic regression model which depends only on the past. They suggested to infer the model parameters by maximum Likelihood estimation. On one hand, they assume that the temporal dependencies satisfy the following prop-

erty $[\mathbf{Z_t}|\mathbf{Z}_{t'}, t' = t-1, t-2, ...] = [\mathbf{Z_t}|\mathbf{Z}_{t'}, t' = t-1, ..., t-\tau]$ i.e. the model is a Markov model of order τ (the distribution of Z_t depends to the past only through the τ last times). On the other hand, for a given time t, the spatial field is a spatial Markov random field and the conditional probabilities are given by

$$\begin{aligned} &\text{logit}\left[\mathbb{P}(Z_{it} = 1|Z_{jt}, j \in N_i; \mathbf{Z}_{t'}, t' = t-1, ..., t-\tau; \mathbf{X})\right] \quad (8) \\ &= \mathbf{X}_{i,t'}^T\boldsymbol{\beta} + \sum_{j\in N_i} \rho_1 Z_{j,t} + \sum_{s=1}^{\tau} \rho_{1+s} Z_{i,t-s}. \end{aligned}$$

Here the model allows the temporal autoregression to be of order greater than 1. And the coefficients ρ_{1+s} depend on the order s of the autoregression. For instance for the spread of an illness, we can expect that all these coefficients are positive which means that the probability increases with the age of the symptoms. However, nothing is said about the existence of the joint distribution of such a process in the paper.

To overcome the difficulties of interpretation identified by Caragea and Kaiser (2009) in the spatial case, Wang and Zheng (2013) were the first to develop a centered parametrisation version of Equation 8 in a spatio-temporal framework. They propose the following modelling:

$$\begin{aligned} &\text{logit}(\mathbb{P}(Z_{i,t} = 1|Z_{i',t'} : (i',t') \in N_{i,t}; \mathbf{X})) \quad (9) \\ &= \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \sum_{j\in N_i} \rho_1 Z^*_{j,t} + \rho_2(Z^*_{i,t-1} + Z^*_{i,t+1}), \end{aligned}$$

$$\text{with} \quad Z^*_{i,t} = Z_{i,t} - \frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta})}{1 + exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta})}. \quad (10)$$

They proposed Expectation-Maximisation (EM) algorithm to maximise the Pseudo-Likelihood and Monte Carlo Expectation-Maximisation Likelihood, as well as consider Bayesian inference to obtain the estimates of model parameters. They found that Monte Carlo Expectation-Maximisation Likelihood algorithm is optimal taking into account the criteria of calculation time and accuracy of the estimate. Further, they compared the statistical efficiency of these approaches.

In the next section we propose a new centered spatio-temporal model and we believe that such model is better adapted for parameter interpretation in a spatio-temporal context.

2. A new centered spatio-temporal autologistic model

2.1. Model specification

In this section, we propose a new spatio-temporal autologistic model, specified by Markov field Markov chain (Gaetan and Guyon, 2008). It can include spatio-temporal covariates. In order to avoid bias and interpretation problems caused by spatial self-regression, we propose to center the corresponding covariates term. We define the model as follows. First, we assume that, conditionally to the covariates, $\{\mathbf{Z}_t, t = 1, 2, ...\}$ is a Markov chain:

$$[\mathbf{Z}_t|\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ..., \mathbf{X}] = [\mathbf{Z}_t|\mathbf{Z}_{t-1}, \mathbf{X}_t]$$

where $\mathbf{X}$ is the spatio-temporal process of vector of the covariates $(X_{i,t})_{1\leq i\leq n, 1\leq t\leq T}$. Moreover, we assume that $\mathbf{Z}_t$ is a Markov random field conditional on $\mathbf{Z}_{t-1}$ with spatial neighbour structure N_i, that means

$$[Z_{i,t}|Z_{j,t}, j \neq i; \mathbf{Z}_{t-1}, \mathbf{X}_t] = [Z_{i,t}|Z_{j,t}, j \in N_i, \mathbf{Z}_{t-1}, \mathbf{X}_t].$$

More precisely, we define the conditional distribution of $Z_{i,t}$ by:

$$\text{logit}(\mathbb{P}(Z_{i,t} = 1|Z_{j,t}; j \neq i, \mathbf{Z}_{t-1}, \mathbf{X}_t)) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_1 \sum_{j \in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1}, \quad (11)$$

$$\text{where} \quad Z_{i,t}^{**} = Z_{i,t} - \frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1})}{1 + exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1})}. \quad (12)$$

We will discuss the interest of this centering in the next section but we can note that this new centered specification looks like a hierarchical model with a latent auto-regressive model of order 1 given by:

$$\begin{aligned}
\text{logit}(\mathbb{P}(\mathbf{Z}_{i,t} = 1|\xi_{i,t}, \mathbf{X}_{i,t})) &= \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \xi_{i,t}, \\
\xi_{i,t} &= \rho_2\xi_{i,t-1} + \omega_{i,t}, \qquad \text{with} \\
Z_{i,t-1} &\approx \xi_{i,t-1} \quad \text{and} \\
\sum_{j \in N_i} Z_{j,t}^{**} &= \sum_{j \in N_i} (Z_{j,t} - \frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta} + \rho_2 Z_{j,t-1})}{1 + exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta} + \rho_2 Z_{j,t-1})}) \approx \omega_{i,t}.
\end{aligned}$$

2.2. Model Interpretation

There are two main differences between one-step centered model of Wang and Zheng (2013) given by Equation 10 and our new centered model given by Equation 12. The first one is that we do not center de temporal term $Z_{i,t-1}$. Indeed, there is no difficulty for the interpretation of ρ_2 because unlike the $\sum_{j \in N_i} Z_{j,t}$, $Z_{i,t-1}$ is known at time t and can be treated like the other covariates. The second one is that the $Z_{j,t}$'s in the term of spatial autoregression,

are centered differently: the former is centered with $\frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta})}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta})}$, and the latter is centered with $\frac{exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1})}{1+exp(\mathbf{X}_{j,t}^T\boldsymbol{\beta}+\rho_2 Z_{j,t-1})}$. It has to be noted that due to the link function, the expectation $\mathbb{E}(Z_{it}|\mathbf{Z}_{t-1}, \mathbf{X}_t)$ equals $\frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1})}{1+exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta}+\rho_2 Z_{i,t-1})}$. The construction of this new centered autologistic model is explained as follows; the aim of the modelling is to separate the large- and small-scale structures. The parameters of spatio-temporal dependence ρ_1, ρ_2 can be interpreted and they have a practical interpretation if the event modelled is an illness:

- Instantaneous spatial dependence ρ_1. It quantifies the spatial autocorrelation between neighbours for the occurence of the event at each time. To model illness, we expect some kind of common sensibility quantified by $\rho_1 \geq 0$. Strong spatial dependence indicates a highly aggregated spatial structure, which makes it possible to identify and monitor the aggregated zones with high infection possibility.

- Temporal dependence ρ_2. It quantifies the temporal dependence on the previous year's status. Again, we expect that the illness likely remains at a place such that $0 \leq \rho_2$. ($\rho_2 < 0$ would indicate a temporal evolution with high frequency at 2-year cycle, this is not adapted for most of the biological processes); ρ_2 is a term of autoregressive regression. Strong temporal dependence can be interpreted as a smooth temporal evolution. If the external effects (covariates) are constant, the individuals have a tendency to keep their status. For instance, if we consider annual data (meaning that t refers to the year), this may indicate no need to monitor two consecutive years if the exterior factors are the same between two years.

2.3. Existence of the joint distribution

Many authors have discussed the existence of a joint multivariate distribution defined by a set of univariate conditional distributions for auto-models that include the autologistic models. Hammersley & Clifford were the first to work on this subject in 1971 (unpublished manuscript!) and the results were written and generalised for instance in Grimmett (1973); Kaiser and Cressie (2000); Gaetan and Guyon (2008). It is probably not possible to show the existence of the distribution of the whole spatio-temporal processes defined by Equation 11 for all years t and sites of the lattice s_i. In our case, the functionals that appear in the right side of Equation 11 are not invariant by permutating the temporal indices. Note that the model of Zhu et al. (2005) given by Equation 6 satisfies this invariance necessary for the joint distribution to exist, but it raises the problem of using the future $(Z_{i,t+1})$ to model the present $(Z_{i,t})$. To prove the existence of our process, we consider the framework of Markov chain of Markov fields presented in Guyon and Hardouin (2002). Using the Hammersley-Clifford theorem given for instance in Gaetan and Guyon (2008), we can show the existence of the joint law of spatial process $\mathbf{Z}_t$ for a fixed t given the past of $\mathbf{Z}_t$ and the current information about the covariates and derive an expression of this conditional joint spatial distribution. Thus, the existence of the distribution of the whole spatio-temporal process, is trivial by recursivity. In addition, we can also obtain the formula of the conditional transition probabilities of the spatial Markov chain. We have the following result.

Theorem 2.1. *Let $(\mathbf{Z}_t)_{(0\leq t\leq T)}$ be the spatio-temporal process defined by (11), let us denote $\mathcal{F}_t^X = \sigma\{\mathbf{X}_{i,s}, 1 \leq i \leq n, s \leq t\}$ and $\mathcal{F}_t^Z = \sigma\{Z_{i,s}, 1 \leq i \leq$*

$n, s \leq t\}$ *the* σ*-algebra generated by the covariates and the process of interest respectively.*

Given $\mathcal{F}_{t-1}^Z \wedge \mathcal{F}_t^X$ *the conditional joint law of* $\mathbf{Z}_t$ *denoted by* $\pi_t(. \mid \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z)$ *is well defined. Moreover, for* $\mathbf{z} = (z_1, \ldots, z_n) \in \{0,1\}^n$*, the spatial conditional joint law is of the form*

$$\begin{aligned}
\pi_t(\mathbf{z} \mid \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) &= C(\mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) \exp(\sum_{i \in S} \Phi_i(z_i) + \sum_{\{i,j\}} \Phi_i(z_i, z_j)) \quad \textit{with} \\
\Phi_i(z_i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) &= z_i \left(\mathbf{X}_{i,t}^T \boldsymbol{\beta} - \rho_1 \sum_{j \in N_i} \frac{exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 Z_{j,t-1})}{1 + exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 Z_{j,t-1})} + \rho_2 Z_{i,t-1} \right) \\
\Phi_i(z_i, z_j) &= \rho_1 \mathbb{1}_{\{j \in N_i\}} z_i z_j
\end{aligned}$$

and $C(\mathcal{F}_t^X, \mathcal{F}_{t-1}^Z)$ *can be considered as constant if the past of process* $\mathbf{Z}_t$ *and the current values at time* t *of the covariates are known.*

The transition probabilities of the Markov chain are

$$\begin{aligned}
\mathbb{P}(\mathbf{y}, \mathbf{z} \mid \mathcal{F}_t^X) &= C(y, \mathcal{F}_t^X) \exp\left(\sum_{i \in S} (z_i \mathbf{X}_{i,t}^T \boldsymbol{\beta} + \rho_1 \sum_{j \in N_i} z_i z_j) \right) \\
&\times \exp\left(\sum_{i \in S} z_i (\rho_2 y_i - \rho_1 \sum_{j \in N_i} \frac{exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 y_j)}{1 + exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 y_j)} - \rho_2 y_i) \right)
\end{aligned}$$

Proof: We use the Hammersley-Clifford theorem given for instance in Gaetan and Guyon (2008) in a conditional form. With the above notation, let us define two assumptions given by (13) and (14) that says that if

$$\begin{aligned}
&\text{logit}(\mathbb{P}(Z_{i,t} = z_i | Z_t^i = z^i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z)) \qquad (13) \\
&= A_i(z^i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) B_i(z_i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) + C_i(z_i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) + D_i(z^i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z)
\end{aligned}$$

and if $\forall i \neq j$, it exists α_i and $\rho_{ij} = \rho_{ji}$ such that

$$A_i(z^i, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) = \alpha_i(\mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) + \sum_{j \neq i} \rho_{ij} B_j(z_j, \mathcal{F}_t^X, \mathcal{F}_{t-1}^Z) \tag{14}$$

The conditional laws satisfying (13) and (14) are consistent with a joint distribution that is a Markov random field. If we rewrite (11) in the following form, it is easy to see that our model satisfies the required conditions:

$$\begin{aligned} &\text{logit}(\mathbb{P}(Z_{i,t} = 1 | Z_{j,t}; j \neq i, \mathbf{Z}_{t-1}, \mathbf{X}_t)) \\ &= \underbrace{\mathbf{X}_{i,t}^T \boldsymbol{\beta} - \rho_1 \sum_{j \in N_i} \frac{exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 Z_{j,t-1})}{1 + exp(\mathbf{X}_{j,t}^T \boldsymbol{\beta} + \rho_2 Z_{j,t-1})}) + \rho_2 Z_{i,t-1}}_{=\alpha_i(\mathcal{F}_t^X, \mathcal{F}_{t-1}^Z))} + \sum_{j \neq i} \underbrace{\rho_1 \mathbb{1}_{j \in N_i}}_{\rho_{ij}} \underbrace{Z_{j,t}}_{B_j(z_j)}. \end{aligned}$$

The expression of π comes directly from Hammersley-Clifford theorem and the transitions probabilities from Guyon and Hardouin (2002). □

3. Comparative simulation study

3.1. Simulation study objective

The idea of proposing the new centered model is to make an agreement between large-scale model and data structure. In particular, if parameters are intended to reflect an overall mean or the effect of covariates, then they should have a constant interpretation across varying levels of statistical dependence. In this section, we compare three models: the model without centering that Caragea and Kaiser called "traditional" and one-step centered as well as the new centered model, defined again below. We want to verify if the marginal structure of data reflects the large-scale structure. The three studied models are given by the general following equation differing according to the kind of centering of $Z_{i,t}^{**}$.

$$\text{logit}(p_{i,t}) = \mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_1 \sum_{j \in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1},$$

$$Z_{i,t}^{**} = Z_{i,t} \qquad \text{traditional model} \tag{15}$$

$$Z_{i,t}^{**} = Z_{i,t} - \frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta})}{1 + exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta})} \qquad \text{one-step model} \tag{16}$$

$$Z_{i,t}^{**} = Z_{i,t} - \frac{exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1})}{1 + exp(\mathbf{X}_{i,t}^T\boldsymbol{\beta} + \rho_2 Z_{i,t-1})} \qquad \text{new model} \tag{17}$$

The agreement between spatial large-scale model structures and marginal data structures have been already examined by Caragea and Kaiser (2009) for both centered and traditional spatial autologistic regression models. They showed that the realised trajectories from the traditional autologistic regression model can not reflect the large-scale structure, and this difficulty can be alleviated by centered parametrisation.

In this paper, we focus on examining the time variation of the large-scale model structure and marginal data structure for the three spatio-temporal autologistic models mentioned above in the case when the covariates depend only on the time t but not on the location on the grid. Note that Caragea and Kaiser (2009) has already carried out a simulation study in the purpose to verify the agreement between the spatial model and spatial data structures according to the different values of the spatial covariates.

Therefore, we have chosen to simulate trajectories of a dynamic process of Markov random field that have a temporal large-scale structure with a deterministic tendency. For site s_i at year t, we define a large model structure with one temporal covariate: $\text{logit}(p_{i,t}) = \beta_0 + \beta_1 X_{it} + \rho_1 \sum_{j \in N_i} Z_{j,t}^{**} + \rho_2 Z_{i,t-1}$,

where $X_{it} = X_t$ is a temporal covariate, that is spatial constant at year t. Thus the average large-scale model at year t is:

$$L_t = \frac{1}{n}\sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 X_t)}{1 + exp(\beta_0 + \beta_1 X_t)} = \frac{exp(\beta_0 + \beta_1 X_t)}{1 + exp(\beta_0 + \beta_1 X_t)}. \tag{18}$$

Moreover, we can also compute an average scale conditional to the past defined by:

$$C_t = \frac{1}{n}\sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 X_t + \rho_2 Z_{i,t-1})}{1 + exp(\beta_0 + \beta_1 X_t + \rho_2 Z_{i,t-1}))}. \tag{19}$$

It can be noted that this process is not deterministic but specific to each realisation of the field $\mathbf{Z}_{t-1}$.

To represent marginal data structure, the marginal empirical data mean of the Markov random field at year t is computed from a simulated field at time t as:

$$D_t = \frac{1}{n}\sum_{i=1}^{n} Z_{it}^{sim}. \tag{20}$$

The objective of the simulation studies presented here is to compare the behaviour of L_t, C_t and that of D_t for the three different models differing by the kind of centering of the spatial autoregression.

3.2. Sampling Algorithms

Hughes et al. (2011) proposed to use perfect sampling to generate Markov Random Field (MRF) samples. The advantage of the perfect sampling compared to Gibbs sampling is that we don't need the burn-in step, nor do we need to decide the spacing numbers. It gives us the exact draws from a given

target distribution when the algorithm completes, details are given in Kendall (2005) Kendall (2005). However, its algorithm running time is random even is still finite. We do not know at which moment the lower chain and the upper chain coalesce. So the number of repetitions is random. In our case, we have to generate a Markov chain Markov random field, the computation time of such a algorithm is quite difficult to control.

Here we use Gibbs sampler but start at a "perfect simulated" sample, we call it PGS sampling. It is less time consuming than perfect sampling, and do not need to decide burn-in and spacing when compared with Gibbs sampling. The PGS sampling was often used to generate the simulated trajectories of autologistic model (Zhu et al., 2008; Zheng and Zhu, 2008; Wang and Zheng, 2013).

3.3. Simulated data

We focus on two types of large-scale model structures, the first one without covariate and the second one with increasing temporal tendency given by a covariate depending only on the time. Both models were for data on a 20 by 20 lattice for 50 time units; the details being given in the two following sections 3.3.1 and 3.3.2. We simulated data according to these structures but with different values of "auto-regression parameter" (ρ_1, ρ_2), in order to evaluate the joint effects of (ρ_1, ρ_2) to the agreement between large-scale structure models and data structures. One trajectory of each of the three models is drawn for each configuration of (ρ_1, ρ_2). To study the dispersion of the empirical large scale structure and confirm the possible tendencies exhibited by the trajectories, we have performed 100 independent realisations of process D_t for each of the three models (traditional, one-step and new

centered) and specified by different values of (ρ_1, ρ_2) and then computed and drawn the empirical fluctuation dynamical intervals. Simulations results are presented in the following sections.

3.3.1. Model 1

We first consider a model without covariates (that is with only an intercept). It is given by $\text{logit}(p_{it}) = \beta_0 + \rho_1 \sum_{j \in N_i} Z^{**}_{j,t} + \rho_2 Z_{i,t-1}$, there is no temporal covariate in this model apart the term of temporal auto-regression. The baseline level of infection is chosen via β_0 such that $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.2$ leading to a constant expected model structure constant $L_t = 0.2$. The initial field is generated by independent Bernoulli variables with parameter $p = 0.2$. To see the effect of the intensity of autocorrelation given by the two parameters (ρ_1, ρ_2) on the difference between the models, we generate simulated data and draw one trajectory from each model with different values of parameters $(\rho_1, \rho_2) \in \{0.3, 0.5, 0.7\}^2$. The empirical confidence intervals were computed with 100 realisations of independent trajectories for each models and for different values of parameters $(\rho_1, \rho_2) \in \{0.5, 0.7\}^2$.

3.3.2. Model 2

We consider here a model with a temporal trend. We consider large-scale structures with one temporal covariate: $\beta_0 + \beta_1 X_t$. A said above, X_t is constant spatially for each year but show monotonic increasing with time: $X_t = t$.

We choose β_0 such that $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.1$ and the initial field is generated by independent Bernoulli variables with parameter $p = 0.1$. With this model, we have $L_t = \frac{exp(\beta_0+\beta_1 t)}{1+exp(\beta_0+\beta_1 t)}$, a monotonic increasing function. With the chosen

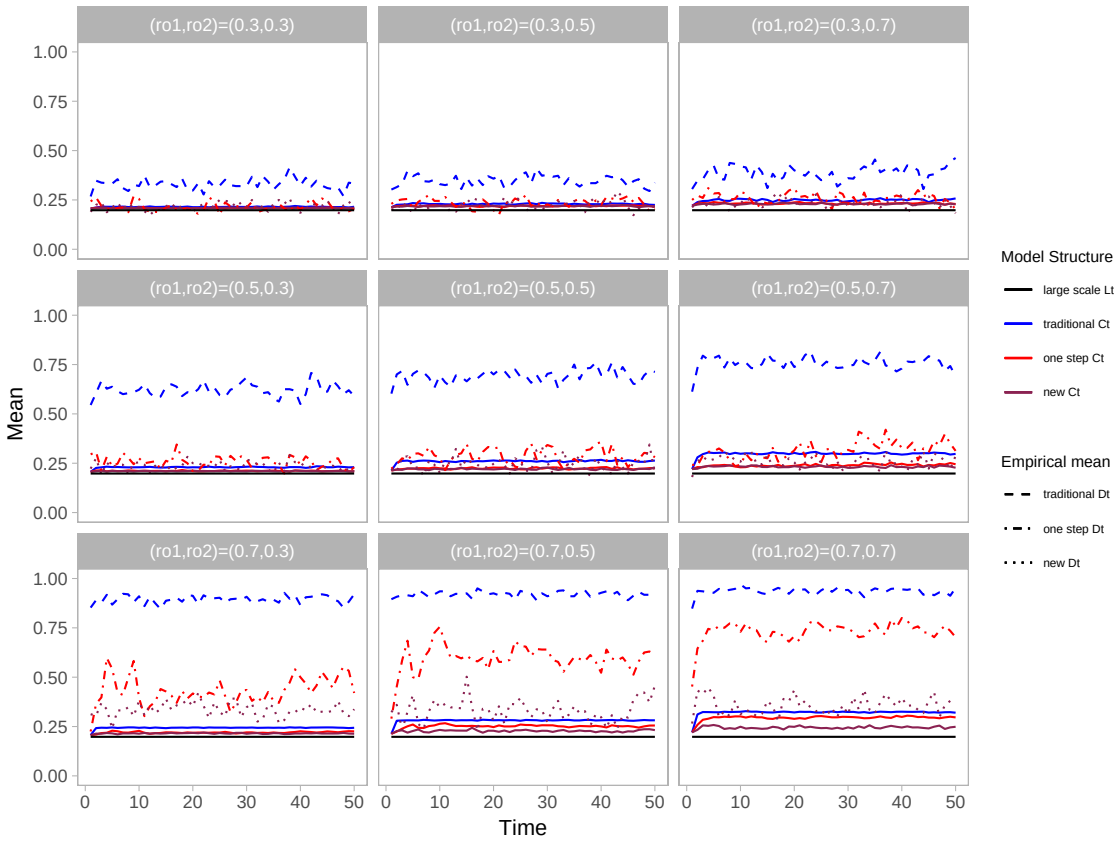

Figure 1: Comparison between large-scale model structure L_t (represented by black line), the expected means C_t according to the past (lines) and empirical mean of data structures D_t for the traditional (blue dashed), the one-step (red dot-dashed) and new (violet dots) centered models for different values of auto-regression parameters (ρ_1, ρ_2). The grid is 20×20, and $0 \leq t \leq 50$, baseline infection $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.2$.

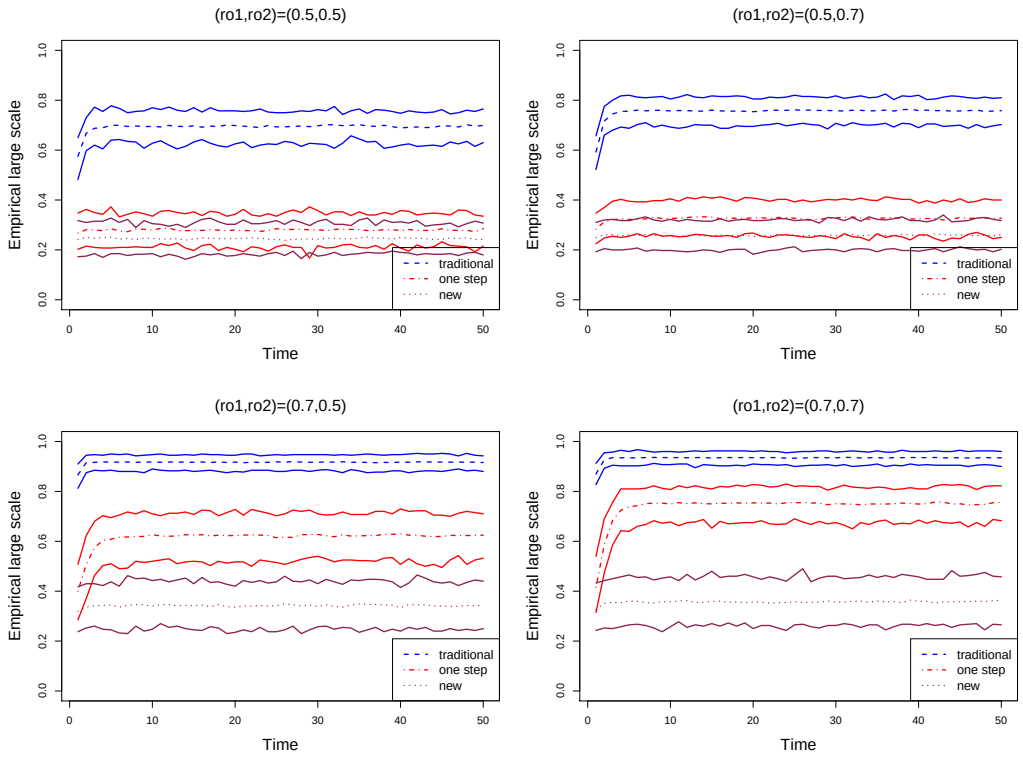

Figure 2: Empirical confidence curves for the large scale structures D_t for the model without covariate. The grid is 20×20, $t = 50$, baseline infection $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.2$. About the centering, traditional (resp. one-step and new centered) is drawn in blue dashed, (resp. red dot-dashed and violet dots).

coefficients, L_t increases from 0.1 at time one to 0.94 at time 50. Again, for a trajectory, we compute the conditional mean given the past. Unlike the model without covariates, we expect an empirical large-scale structure that increases with time.

3.3.3. Results

From Figure 2 and Figure 4, we see that the spread of the realisations of the empirical mean D_t is not very large and that for the given value of the parameter (ρ_1, ρ_2), the models may differ more or less. We see that scatter of the empirical average of the spatial field for each t is low enough to trust and interpret the difference between the individual curves shown in Figure 1.

In Figure 1 (resp. Figure 3) we draw one trajectory from each model without covariate (resp. with covariates) for different values of ρ_1, ρ_2 to study the difference between the models according to these parameters that reflect the dependence level between the $Z_{i,t}$. We see that the empirical mean D_t for the traditional model is always greater than the expected large scale structure and than the realisation of D_t for the one-step centered and new centered models. The value of D_t for the traditional model is not sensible to the value of the temporal autoregression parameter ρ_2 but it is very sensible to the spatial auto-regression parameter ρ_1 and it increases as ρ_1 increases.

Regarding the centered models, they show the same large scale behaviour when parameter ρ_1 equals 0.3 or 0.5. For $\rho_1 = 0.7$, the two centered models show different large scale behaviour but the D_t for new centered model agrees with the expected mean behaviour. The difference between them increases as ρ_2 increases.

In summary, the difference between data simulated according to the one-

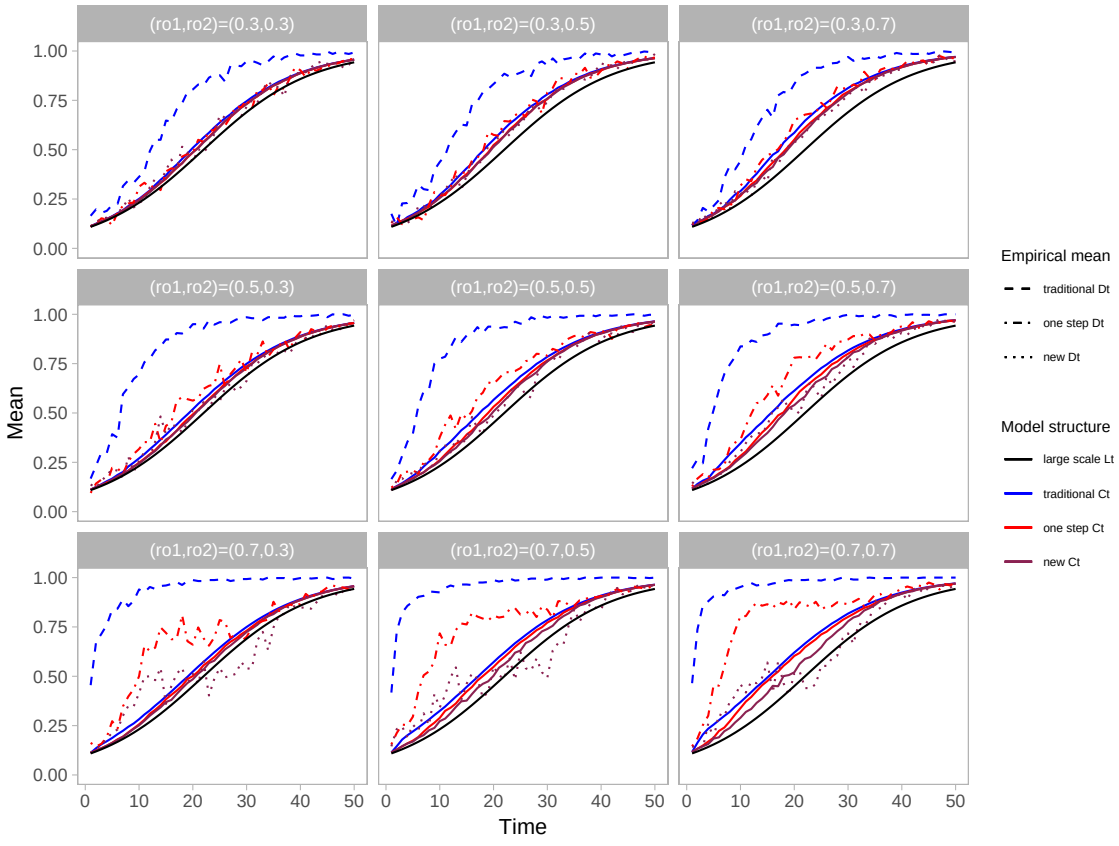

Figure 3: Comparison between large-scale model structure L_t (represented by black line), the expected means C_t according to the past (lines) and empirical mean of data structures D_t for the traditional (blue dashed), the one-step (red dot-dashed) and new (violet dots) centered models for different values of auto-regression parameters (ρ_1, ρ_2). The grid is 20×20, and $0 \leq t \leq 50$, baseline infection $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.1$, and covariate coefficient $\beta_1 = 0$, and covariate $X_t = t$.

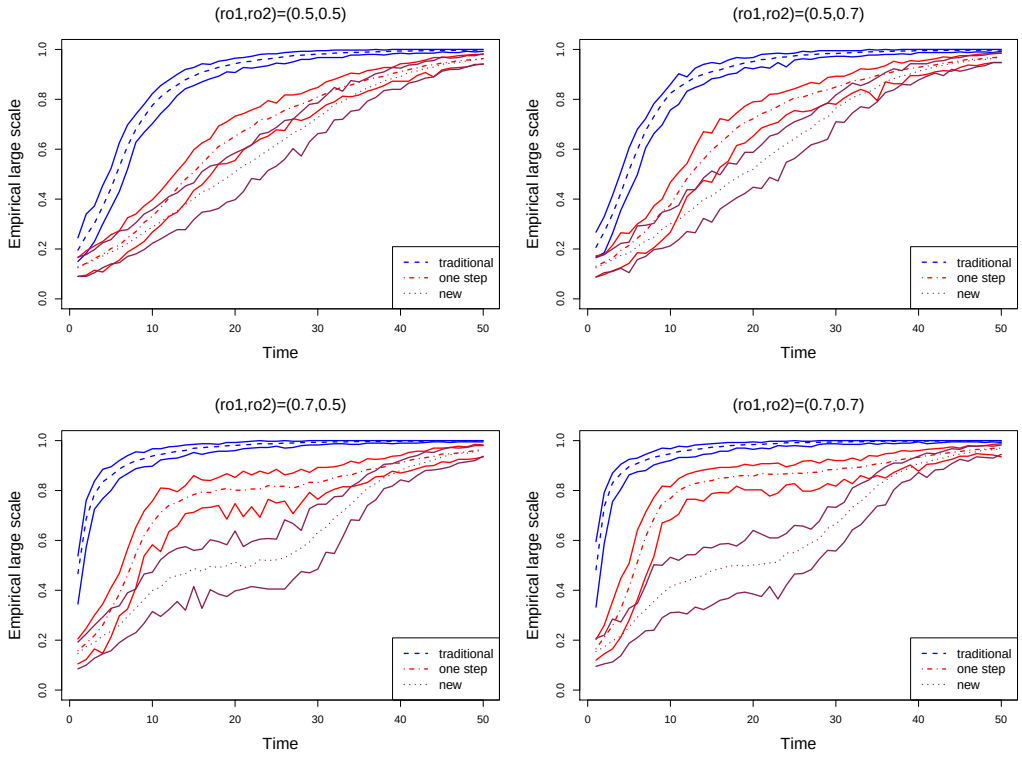

Figure 4: Empirical confidence curves for the large scale structures D_t for the model with temporal covariate $X_t = t$. The grid is 20×20, $t = 50$,baseline infection $\frac{exp(\beta_0)}{1+exp(\beta_0)} = 0.1$, and covariate coefficient $\beta_1 = 0$, and covariate $X_t = t$. About the centering, About the centering, traditional (resp. one-step and new centered) is drawn in blue dashed, (resp. red dot-dashed and violet dots).

step model and the new centered autologistic model are small except when both spatial and temporal dependence are relatively strong. It can be seen from the simulated data that the traditional or one-step centered model over-represents the large-scale structure and incorrectly differs from the expected structure of the model. As a result, interpretations are complex and can lead to erroneous conclusions.

4. Estimation

From now on, we only consider the new centered model. We propose here a method of estimation by Pseudo-Likelihood Maximisation (MPL) and a method for selecting the most likely neighbourhood structure of the data set. We show their performances by testing them on simulated data.

4.1. Estimation by Pseudo-Likelihood Maximisation

Since the structure of the new centered autologistic model is more complicated than the traditional or the one-step centered one, both Monte Carlo Maximum Likelihood Estimation (MCML) and Bayesian methods can be very heavy and sophisticated to implement. We propose to estimate the parameters of our model using the estimator that maximises Pseudo-Likelihood. It is very easy to implement. Some authors have studied the mathematical properties such as convergence of this estimator under the constraint of homogeneity and ergodicity of the Markov Random Field Markov Chain and other required assumptions can be found in Guyon and Hardouin (2002). The imbrication of the parameters in the definition of the centered variables $Z^{**}_{i,t}$'s (Equation 17) and the presence of covariates make these conditions hard to verify, that is why, in the following, we only look at the empirical

behaviour of this estimator. The Pseudo-Likelihood for our model is given by the following formula:

$$\begin{aligned} \mathcal{PL}(\boldsymbol{\beta}, \rho_1, \rho_2) &= \prod_{t=1}^{T} \left(\prod_{1 \leq i \leq n} p_{it} \right) \\ &= \prod_{t=1}^{T} \left(\prod_{1 \leq i \leq n} \frac{\exp\left(\mathbf{X}_{i,t}^T \boldsymbol{\beta} + \rho_1 \left(\sum_{j \in N_i} Z_{j,t}^{**} \right) + \rho_2 Z_{i,t-1} \right)}{1 + \exp\left(\mathbf{X}_{i,t}^T \boldsymbol{\beta} + \rho_1 \left(\sum_{j \in N_i} Z_{j,t}^{**} \right) + \rho_2 Z_{i,t-1} \right)} \right). \end{aligned} \tag{21}$$

MPL Estimator is the vector of parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \rho_1, \rho_2\}$ that maximises $\mathcal{PL}(\boldsymbol{\beta}, \rho_1, \rho_2)$. We see that the spatial auto-regression covariate $Z_{j,t}^{**}$ imbricated the couple of parameters (ρ_1, ρ_2) themselves so that it is not possible to consider $Z_{j,t}^{**}$ as a common "external covariate". For this reason, the maximisation has to be made by an Expectation-Maximisation algorithm. We give the details of this algorithm in the next section.

4.1.1. Algorithm

We applied the EMPL (Expectation Maximisation Pseudo-Likelihood) algorithm, the principle is the same as described in Zheng and Zhu (2008), but with two iteration steps to accelerate the numerical algorithm/calculation.

The steps are as follows:

- Initialisation: to obtain the estimation of $\boldsymbol{\theta_1} = (\boldsymbol{\beta}, \rho_2)$, denoted by $(\tilde{\boldsymbol{\beta}}, \tilde{\rho_2})$, from model $logit(p_{it}) = \mathbf{X}_{i,t}^T \boldsymbol{\beta} + \rho_2 Z_{i,t-1}$, we maximise the corresponding log Pseudo(partial)-Likelihood by Quasi-Newton.

- Step 2: to obtain the estimation of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho_1, \rho_2)$, denoted by $\breve{\boldsymbol{\theta}} = (\breve{\boldsymbol{\beta}}, \breve{\rho_1}, \breve{\rho_2})$, for the new centered autologistic model:

1. Initialization: Set initial values: $\boldsymbol{\theta}^0 = (\tilde{\boldsymbol{\beta}}, 1, \tilde{\rho}_2)$
2. Expectation: Given $\boldsymbol{\theta}^{l-1}$, compute the $Z_{j,t}^{**}$'s by removing the corresponded trend.
3. Maximisation: Obtain $\boldsymbol{\theta}^l$ by maximising the log Pseudo-Likelihood by Quasi-Newton.
4. Go to 2

- Obtain estimates $(\breve{\boldsymbol{\beta}}, \breve{\rho}_1, \breve{\rho}_2)$.

4.1.2. Variance of the MPLE

Even if the convergence properties of the Maximum Pseudo-Likelihood Estimator is well known under specified conditions discussed above, the variance of the estimator has to be carefully estimated. For the variance-covariance matrix of the coefficients, we propose to compute the matrix $\mathbf{U}'\mathbf{W}\mathbf{U}$ as if we were in the case of a classic variance in the case of Maximum Likelihood in the logistic case. The matrix $\mathbf{U}$ is a $nT \times p$ matrix defined by theses rows $\mathbf{U}_{it,\cdot} = (1, X_i^T, \sum_{j \in \mathcal{N}_i} Y_{jt}, Y_{i,t-1})$ for each (i,t), $1 \leq t \leq T$ and $1 \leq i \leq n$. $\mathbf{W}$ is the diagonal $nT \times nT$ matrix with coefficients being equal to $\breve{p}_{it}(1 - \breve{p}_{it})$ that depends on the estimation parameters $(\breve{\boldsymbol{\beta}}, \breve{\rho}_1, \breve{\rho}_2)$.

We compared the variances of the estimators with ones computed by "bootstrap" on simulated data.

4.2. Model choice

Although the estimation method is efficient for a given neighbourhood structure, it is necessary to select the best one. Indeed, in the context of modelling the occurrence of a disease, learning the structure of the neighbourhood is a way to understand the mechanisms of disease spread.

Firstly, we had the idea of adapting the ABC method (for Approximate Bayesian Computation) proposed by Grelaud et al. (2009) in order to choose the best model for the neighbourhood. ABC is a Likelihood-free inference method in the bayesian framework that is very convenient when the Likelihood is not available in a closed form. First introduced by Pritchard et al. (1999) and expanded in Beaumont et al. (2002) and Marjoram et al. (2003), ABC method was adapted by Grelaud et al. (2009) for model choice in Gibbs Random Fields (GRF). We first thought about using this method because our model is not so far from a GRF. But because of the centering parametrisation, it is not possible to produce a sufficient statistics in order to compute a simple distance between the simulated field and the observed one. We can see in equations of the spatial joint laws of Theorem 2.1, that the parameters $(\boldsymbol{\beta}, \rho_1)$ are nested in the potentials depending on the spatio-temporal field $\mathbf{Z}$. We tried this method with different statistics without success. However, this sophisticated method is not essential here because we observed on simulated data that the Log-Pseudo-Likelihood value is a very simple and performant indicator to choose the model of a given data set.

The approach we propose is the following: we use"experts opinions" to determine a set of possible neighbours structures to consider. For each structure, we estimate the parameters and simply choose the structure that optimises Log-Pseudo-Likelihood.

4.3. Simulation study

4.3.1. Model 1

The first configuration was again $20 * 20$ grid for 15 years without covariate. The initial field is generated by Bernoulli distribution with parameter

0.1 and model parameters are $\beta_0 = -1.4, \beta_1 = 0, \rho_1 = 0.5, \rho_2 = 0.5$.

We standardise the neighbourhood structure here: we suppose the points s_i's are on a grid, and we assume their spatial distribution is like a matrix – each point is located on the intersection of a row and a column. All models considered in this section to generate simulated data have the same structure of neighbourhood: each point has four neighbours on the same row (the four nearest that is in our case the two nearest on each side) and also two neighbours in the same column. Points of the first row (resp. second row) have only two (resp. three) neighbours on the same row and points on the first and second columns have only one neighbour on the same column (and so on for the last row or column). This configuration is denoted $v_r = 2$ (for 2 neighbours on each side in the row) and $v_c = 1$ (resp. 1 in the column). Figure 7 show such kind of neighbourhood for $v_r = 4$ and $v_c = 2$.

For our simulation study to infer the neighbour's structure, we assume that all the neighbourhood structures considered are regular that is, the definition of the neighbourhood is the same for all the points of the grid. And we only consider the neighbourhood structures defined by the number of neighbours on each side in the same row v_r (or in the same column v_c). We do not consider the possibility of having neighbours on the diagonal.

4.3.2. Model 2

Again a $20 * 20$ grid for 15 years with one temporal covariate with large variation first increasing from 1 to 8 the 8 first years and next decreasing by 1 until the year 15. $x(t) = t$ for $1 \leq 8$ and $x(t) = 14 - t$ for $1 \leq 7$, $\beta_0 = -2.8, \beta_1 = 0.1, \rho_1 = 0.5, \rho_2 = 0.5$. The structure of neighbourhood is given by $v_r = 2$ and $v_c = 1$.

4.3.3. Estimation results

We used again PGS sampling methods described in section 3.2 to simulate trajectories of the process on a $20 * 20$ grid for 15 years in different configurations. The purpose is to study the performance of the estimations and of the model selection procedure. We compute the estimations via EMPL algorithm detailed in section 4.1.

For the variance, we compared the values estimated by the method explained above on a sample, with the experimental value of the variance when we estimate the parameters of $B = 500$ independent repetitions of the same model. Results of the inference for Model 1 (resp. Model 2) is given in table 1 (resp. 2).

Figures 5 and 6 show the dispersion of $B = 500$ estimations of B independent simulations of each model.

	β_0	β_1	ρ_1	ρ_2
mean	-1.47(-1.4)	0.003(0)	0.519(0.5)	0.560(0.5)
est st.deviation	0.066	0.014	0.028	0.071
boot. st.deviation	0.083	0.018	0.034	0.068

Table 1: Maximum Pseudo-Likelihood estimation for Model 1 without covariate, true values are in brackets. Variance estimators is computed by our proposed method (est st.estimation) and by repetitions method (boot st.estimation)

All these results show the good performance of the Maximum Pseudo-Likelihood Estimator and its standard deviation. We have to note that the method is very easy to implement and the results are available almost instantly while it would be not the case with MCMC or Bayesian methods. It

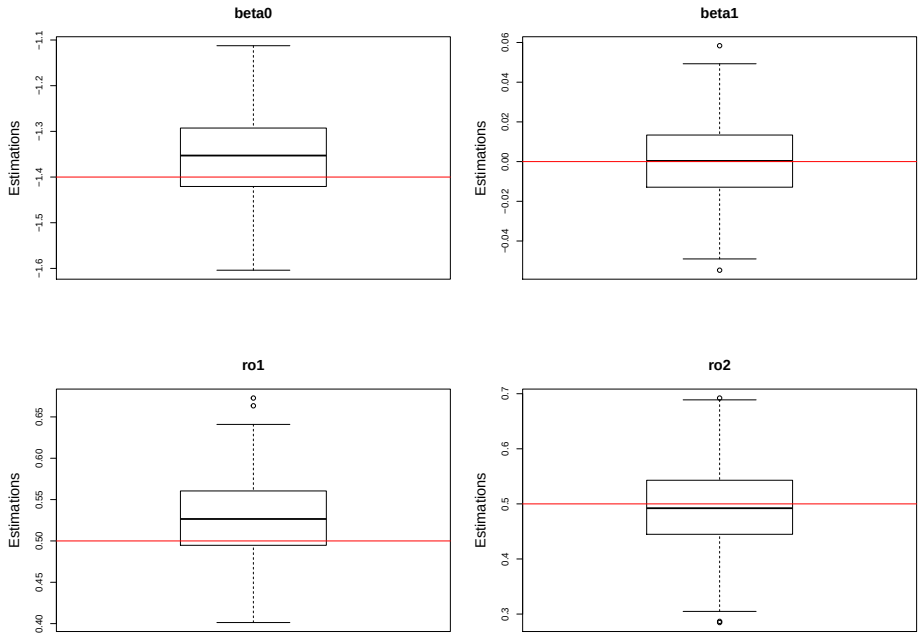

Figure 5: Box-plot of $B = 500$ estimations of the four parameters in Model 1. Red lines show the true values of parameters.

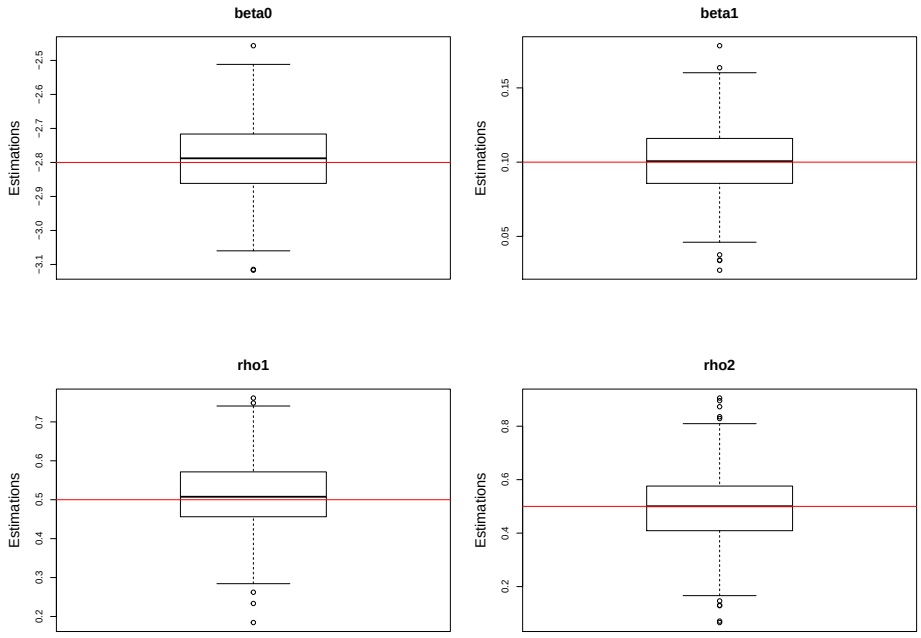

Figure 6: Box-plot of $B = 500$ estimations of the four parameters in Model 2. Red lines show the true values of parameters

	β_0	β_1	ρ_1	ρ_2
mean	-2.757(-2.8)	0.094(0.1)	0.488(0.5)	0.486(0.5)
st.deviation	0.097	0.021	0.042	0.130
boot. st.deviation	0.108	0.022	0.073	0.130

Table 2: Maximum Pseudo-Likelihood estimation for Model 2 with temporal covariate, true values are in brackets. Variance estimators is computed by our proposed method (est st.estimation) and by repetitions method (boot st.estimation).

should also be noted that we have made estimates on simulated data generated with different parameter values and that the method's performance has remained as good as above.

4.3.4. Model choice results

We first show the effects of different choices of neighbourhood graphs on the estimation of the spatial auto-regression parameter ρ_1 of Equation 11. For a given simulated dataset, we perform estimations of the parameters for different graphs of neighbourhood. Results are shown in Table 3 that confirms the intuitive result that the estimation of the spatial autoregressive parameter ρ_1 decreases with the number of neighbours. We have to notice that this decrease is not proportional to the number of neighbours of each points of the grid. Moreover estimation with a wrong neighbourhood structure does not affect the estimation of the other parameters of the models (regression on the past parametred by ρ_2 and on the covariate by $\boldsymbol{\beta}$).

To see the good performance of the model choice rule, we simulated 500 independent realisations of different kinds of models under three different neighbourhood structures, without (resp. with a temporal covariate) and for

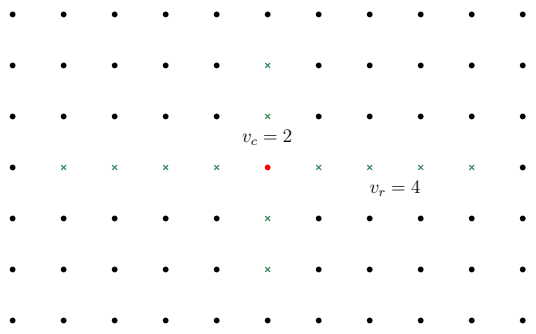

Figure 7: Structure of neighbourhood used for the simulations. Green crosses are the neighbours of the red point.

three different values of ρ_1. We estimated the parameters under six different neighbours structures and choosed the model with the biggest Pseudo-Likelihood. Results are shown in Table 4 (resp. Table 5).

We see that the ability of the rule to detect the true neighbours structure is globally good. However, it has better performance if the model does not include covariates and almost perfect while $\rho_1 \geq 0.3$. Note that if the value of ρ_1 is low, it means that the spatial autocorrelation with the neighbours is low and thus it is not so important to infer the neighbourhood structure properly. The performance of the rule is degraded by the presence of a covariate and again when the relative weight of the spatial auto-correlation in the model is lower.

5. Application to local spread of plant diseases

5.1. Context and modelling

The spectrum of applications of spatio-temporal autologistic models is large. Thus for instance it was applied for the inference of networks in order

Model	v_r	v_c	β_0	β_1	ρ_1	ρ_2
1	1	1	-1.434	0.005	0.604	0.539
2	*2*	*1*	*-1.470(-1.4)*	*0.003(0)*	*0.519(0.5)*	*0.560(0.5)*
3	2	2	-1.467	0.003	0.420	0.560
4	3	1	-1.468	0.004	0.427	0.563
5	3	2	-1.475	0.004	0.368	0.563
6	3	3	-1.464	0.004	0.317	0.567

Table 3: Estimation by Pseudo-Likelihood in different models for neighbourhood. v_r (resp. v_c) is the number of neighbours on each side of a point on the raw (resp. on the column). Estimations for the true model are in red and the true values of the parameters are in brackets.

to study competition in financial markets in Betancourt et al. (2018). In our paper, the model was built with application to plant epidemiology. Indeed, the purpose here is to analyse the spread of the esca grapevine trunk disease over a 14-year period in a vineyard of Bordeaux in France, including 1 980 vines in a block of 30 rows and 66 columns. Esca is a grapevine trunk disease that remains poorly understood but causing extensive damage in vineyard worldwide and resulting in major economic losses (Bertsch et al., 2013; Mugnai et al., 1999). Grapevine esca is a complex dieback disease associated with pathogenic fungi that degrade the woody part of the vine. It exhibits discoloured foliar symptoms (Mugnai et al., 1999; Surico et al., 2008; Lecomte et al., 2012). Leaf symptoms are erratic in the extent that they appear during late spring or summer, however they can appear on year and not the following one. The disease leads to a decrease in wine quality, and worse, to vine de-

	Selected Model																	
True	$\rho_1 = 0.3$						$\rho_1 = 0.4$						$\rho_1 = 0.5$					
Model	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	474	15	5	4	1	1	495	3	2	0	0	0	500	0	0	0	0	0
2	13	451	21	13	1	1	0	486	5	9	0	0	0	499	1	0	0	0
3	0	13	470	0	15	2	0	0	498	0	2	0	0	0	500	0	0	0

Table 4: Model choice by maximising the Pseudo-Likelihood for a model on a $20 * 20$ grid for 15 years without covariate. $\beta_0 = -1.4$, $\rho_2 = 0.5$ and three different values of ρ_1.

	Selected Model																	
True	$\rho_1 = 0.3$						$\rho_1 = 0.4$						$\rho_1 = 0.5$					
Model	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	357	60	27	24	15	17	401	42	32	14	2	9	452	27	7	9	2	3
2	60	287	48	73	15	17	40	344	45	52	9	10	18	424	24	31	1	2
3	18	51	314	12	66	39	15	38	390	2	34	21	4	31	438	1	25	1

Table 5: Model choice by maximising the Pseudo-Likelihood for a model on a $20 * 20$ grid for 15 years with a covariate. $\beta_0 = -2.8$, $\beta_1 = 0.1$, $\rho_2 = 0.5$ and three different values of ρ_1. $X_t = t$ for $t \leq 8$ and $16 - t$ for $8 \leq t \leq 15$.

cline and death at long term. In order to better understand the factors that drive the esca spread, several studies based on spatio-temporal mathematical modelling have been used. Some of them have focused on the contagiousness of symptomatic vines in order to improve the prophylactic control of esca. In Stefanini et al. (2000), a non- centered auto-logistic multinomial statistical model with autoregression on the past and on the neighbourhood was used to study the spatio-temporal dynamics of the esca at the scale of a vineyard. However, the modelling and the inference method were not discussed. More recently, a non-centered autologistic model with Bayesian inference was also used to analyse a 17-year dataseries resulting from the monitoring of one

vineyard since planting (Zanzotto et al., 2013). In Li et al. (2016), we used join count procedures to analyse aggregation and spread of the esca over an eight-year period.

The lattice data came from the esca disease monitoring for 14 years between 2004 and 2017 in a commercial vineyard of the Bordeaux region (vineyard 13 in Li et al. (2016)), which was planted in 1989 with the cultivar Cabernet Sauvignon (Vitis vinifera), and trained in accordance using the Guyot trellising system. Within the surveyed plot, the distances between rows and between vines were respectively 1.4 m and 1.2 m respectively. The foliar esca expression of contiguous vines was recorded each year at the end of August according to methods described in Li et al. (2016). The mean annual esca prevalence was 9.5%, varying annually between 1.8 and 16.8%.

We propose to analyse our data with the objectives to understand: (i) the effect of the status of a vine (symptomatic or not) the year preceding the observation; (ii) the effect of the frequencies of infected plants among the vine's neighbours the year preceding the observation on the occurrence of the symptom for the given vine. Moreover we want to capture the instantaneous spatial correlation between vines in the same year. We want to choose the best neighbourhood structure for the previous year's effect and the instantaneous effect. We are clearly in a context of selecting the best suited models in terms of past and instantaneous neighbourhood structures.

According to physiopathologists, the neighbourhood structures are ellipses shaped (see Figure 8). Indeed, vines are row planted leading to possible anisotropy. The neighbourhood of a vine i located in s_i is given by all vines included in a ellipse defined by its semi-major axis and semi-major axis.

These quantities are denoted v_r and v_c for the instantaneous neighbourhood $\mathcal{N}_i$ (resp. p_r and p_c for the past one $\mathcal{N}_i^p$).The model is the following. For a vine i, located in s_i at time t, the probability to present the symptom according to the history of the vineyards and the neighbourhood is given by:

$$logit(p_{it}) = \beta_0 + \beta_1 \left(\sum_{j \in N_i^p} Z_{j,t-1} \right) + \rho_1 \left(\sum_{j \in N_i} Z_{j,t}^{**} \right) + \rho_2 Z_{i,t-1} \qquad (22)$$

Twenty five neighbourhood structures were tested for the two auto-regressions (instantaneous and on the past) corresponding to 625 different models. A neighbourhood was defined by an ellipse around the vine with a number of neighbours v_r (resp. v_c) in the direction of the row (resp. column). v_r and v_c vary independently from 1 to 5 neighbours leading to 25 possible structures. We also defined p_l and p_c the corresponding parameters for the neighbourhood concerning the past.

5.2. Results

The most important result is that the value of the estimation of coefficient ρ_2 is very robust whatever the chosen neighbourhood structure, ($\rho_2 = 2.28$). It means that the risk for a vine that has already expressed the symptoms the previous year to express them again is multiplied by $\exp(2.28) = 9.7$ comparing to the same risk for a vine without expression the previous year. This result corroborates that of Guérin-Dubrana et al. (2013). These authors have shown that a declining vine has expressed esca symptoms on average two to three times the years before. The re-expression of leaf symptoms is frequent and reveals the advanced state of internal infection (Maher et al., 2012).

With respect to other terms of the regression, another important result is that the choice of the neighbourhood for the instantaneous correlation is more important than the one for the past regression. The 50 best models according to the PL are those with the instantaneous neighbourhood defined by an ellipse (that could also be a circle) with radius of 5 in a row direction and 4 in the other one. This effect is more important than the effect of the past that covers all the possibilities for the neighbourhood. The interpretation of this instantaneous autocorrelation is likely to be found in local environmental effects such as soil properties. This effect is not due to a spread of the illness. Note that it captures a little anisotropy showing more effect along the row, that is a standard result in vineyards because the vines are nursed along the row. The estimated coefficients and associated standard deviation for the best model are given in Table Table 6.

v_r	v_c	p_r	p_c	β_0	β_1	ρ_1	ρ_2
5	4	1	1	-3.04 (0.035)	0.178 (0.034)	0.135 (0.006)	2.28 (0.05)

Table 6: Instantaneous and past neighbour structures and estimated coefficients for the best model on the real data. Standard deviation of the parameters are in bracket.

We have already commented the estimation of autoregression ρ_2. The value of the one for β_0 indicates a spontaneous level of infection equal to $\exp(-3,04) = 0.05$. β_1 is the coefficient of regression that quantifies the spread of the illness, like in the logistic model, when the level of the illness is still low, the presence of a one more vine with symptoms in the neighbourhood at a time multiplies by $\exp(0.178) = 1.19$ the risk to present symptoms the following year. This result shows the small role of the neighbouring

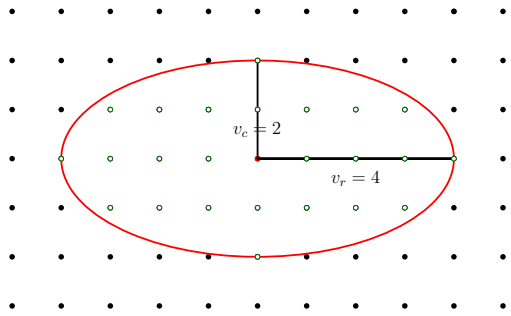

Figure 8: Structure of neighbourhood used for esca data. Green circles are the neighbours of the red point.

symptomatic vines on the disease occurrences. It confirms those of Li et al. (2016) which suggested a limited potential of secondary local spread from neighbouring symptomatic vines.

6. Discussion and conclusion

In this paper, we have proposed a new spatio-temporal model for the study of binary data evolving with time on a lattice. At each time t, the spatial covariates are centered at its expected value which depends on the value of the covariates and also on the values of the field in the past. Simulations studies in section 3, show the interest of this new centering to demonstrate that what is naturally taken to represent large-scale model structure in the traditional parameterization of an autologistic model does not necessarily reflect marginal structure in the data it generates. This new model allows the practitioner to make a good interpretation of the spatial regression parameter that was not possible in previous models. We have shown the ability of the Maximum Pseudo-Likelihood Estimator to infer quickly the value of the pa-

rameters. Maximising Pseudo-Likelihood also allows us to choose efficiently between models with different with different neighbourhood structures. Even if the whole spatio-temporal joint distribution of the process is not proved to exist, we still discuss the existence of the spatial joint law of the process at any time given the covariates and the past of the process. This approach seems coherent with the recursive construction of the process along the time.

The model and the method presented in this paper are very suitable and efficient for modelling the evolution of an illness on a lattice taking into account covariates and spatial auto-correlation. They allow to measure and quantify effects of the neighbourhood in the past on the occurrence of the illness at a given time. It should be noted that although we have thought this model by thinking of a repartition of plants on a lattice, it could be applied to other spatial repartitions provided that the structure of neighbours is well defined. Moreover, for reasons of sparsity we have chosen auto-regression coefficients ρ_1 and ρ_2 which do not depend on the site s_i or the neighbour s_j but the model can be easily complicated depending on the purpose of the modelling (for instance we can distinguish several kind of neighbours). The purpose of the application was to show that our methodology is easy to implement, the data at our disposal were very simple without covariates other than data from the past and the state of the neighbours. But the next step is the acquisition of spatial or spatio-temporal covariates (soil properties, vigour of the plant, water stress..) to better understand their effect on leaf symptoms. Temporal covariates such as weather information would also be interesting to incorporate. We also plan to develop a free software package for the R software that would be available for the analysis of spatio-temporal

binary data.

Acknowledgements

We wish to thank Avner Bar Hen and Cécile Hardouin for precious discussions at the beginning of this work. We wish to acknowledge the vine-grower who participated in this study and also Sylvie Bastien and David Morais for their excellent technical assistance. This research was supported by Bordeaux Sciences Agro, the Regional Council of Aquitaine, the JEAN POUPELAIN Foundation, the French Ministry of Agriculture and the Food-processing industry and Forest (CASDAR V1303).

References

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

C. Bertsch, M. Ramírez-Suero, M. Magnin-Robert, P. Larignon, J. Chong, E. Abou-Mansour, A. Spagnolo, C. Clément, and F. Fontaine. Grapevine trunk diseases: complex and still poorly understood. *Plant pathology*, 62 (2):243–265, 2013.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

B. Betancourt, A. Rodríguez, and N. Boyd. Investigating competition in financial markets: a sparse autologistic model for dynamic network data. *Journal of Applied Statistics*, 45(7):1157–1172, 2018.

P. C. Caragea and M. S. Kaiser. Autologistic models with interpretable parameters. *Journal of agricultural, biological, and environmental statistics*, 14(3):281–300, 2009.

N. Cressie. *Statistics for Spatial Data: Wiley Series in Probability and Statistics*. Wiley-Interscience New York, 1993.

C. Gaetan and X. Guyon. *Modélisation et statistique spatiales*, volume 63. Springer, 2008.

A. Grelaud, C. P. Robert, J.-M. Marin, F. Rodolphe, J.-F. Taly, et al. Abc likelihood-free methods for model choice in gibbs random fields. *Bayesian Analysis*, 4(2):317–335, 2009.

G. R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84, 1973.

L. Guérin-Dubrana, A. Labenne, J. C. Labrousse, S. Bastien, R. Patrice, and A. Gégout-Petit. Statistical analysis of grapevine mortality associated with esca or eutypa dieback foliar expression. *Phytopathologia Mediterranea*, 52 (2):276–288, 2013.

M. L. Gumpertz, J. M. Graham, and J. B. Ristaino. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 131–156, 1997.

X. Guyon. *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media, 1995.

X. Guyon and C. Hardouin. Markov chain markov field dynamics: models and statistics. *Statistics: A Journal of Theoretical and Applied Statistics*, 36(4):339–363, 2002.

F. W. Huffer and H. Wu. Markov chain monte carlo for autologistic regression models with application to the distribution of plant species. *Biometrics*, pages 509–524, 1998.

J. Hughes. ngspatial: A package for fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data. *R Journal*, 6(2), 2014.

J. Hughes, M. Haran, and P. C. Caragea. Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871, 2011.

M. S. Kaiser and N. Cressie. The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220, 2000.

W. S. Kendall. Notes on perfect simulation. *Markov Chain Monte Carlo: Innovations and Applications*, 7, 2005.

P. Lecomte, G. Darrieutort, J.-M. Liminana, G. Comont, A. Muruamendiaraz, F.-J. Legorburu, E. Choueiri, F. Jreijiri, R. El Amil, and M. Fermaud. New insights into esca of grapevine: the development of foliar symptoms and their association with xylem discoloration. *Plant Disease*, 96(7):924–934, 2012.

S. Li, F. Bonneu, J. Chadoeuf, D. Picart, A. Gégout-Petit, and L. Guerin-Dubrana. Spatial and temporal pattern analyses of esca grapevine disease in vineyards in france. *Phytopathology*, 107(1):59–69, 2016.

N. Maher, J. Piot, S. Bastien, J. Vallance, P. Rey, and L. Guérin-Dubrana. Wood necrosis in esca-affected vines: types, relationships and possible links with foliar symptom expression. *OENO One*, 46(1):15–27, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

L. Mugnai, A. Graniti, G. Surico, et al. Esca (black measles) and brown wood-streaking: two old and elusive diseases of grapevines. *Plant disease*, 83(5):404–418, 1999.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

F. M. Stefanini, G. Surico, G. Marchi, et al. Longitudinal analysis of symptom expression in grapevines affected by esca. *Phytopathologia Mediterranea*, 39(1):225–231, 2000.

G. Surico, L. Mugnai, and G. Marchi. The esca disease complex. In *Integrated management of diseases caused by fungi, phytoplasma and bacteria*, pages 119–136. Springer, 2008.

Z. Wang and Y. Zheng. Analysis of binary data via a centered spatial-temporal autologistic regression model. *Environmental and Ecological Statistics*, 20(1):37–57, 2013.

M. Wolters. Better autologistic regression. *Front. Appl. Math. Stat.*, pages 3–24, 2017.

A. Zanzotto, M. Gardiman, S. Serra, D. Bellotto, F. Bruno, F. Greco, and C. Trivisano. The spatiotemporal spread of esca disease in a cabernet sauvignon vineyard: a statistical analysis of field data. *Plant Pathology*, 62(6):1205–1213, 2013.

Y. Zheng and J. Zhu. Markov chain monte carlo for a spatial-temporal autologistic regression model. *Journal of Computational and Graphical Statistics*, 17(1), 2008.

J. Zhu and Y. Zheng. Autologistic regression models for spatio-temporal binary data. *Handbook of Discrete-Valued Time Series*, pages 367–387, 2016.

J. Zhu, H.-C. Huang, and J. Wu. Modeling spatial-temporal binary data using markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2):212–225, 2005.

J. Zhu, Y. Zheng, A. L. Carroll, and B. H. Aukema. Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood. *Journal of agricultural, biological, and environmental statistics*, 13 (1):84–98, 2008.