applied sciences

MDPI

Article

Prediction and Surveillance Sampling Assessment in Plant Nurseries and Fields

Nora C. Monsalve [1,*] and Antonio López-Quílez [2]

[1] Department of Operations Research and Statistics, University Centroccidental Lisandro Alvarado, Barquisimeto 3001, Venezuela
[2] Department of Statistics and Operations Research, University of Valencia, 46100 Burjassot, Spain
* Correspondence: nmonsalve@ucla.edu.ve; Tel.: +58-412-5177009 (ext. 3001)

Abstract: In this paper, we propose a structured additive regression (STAR) model for modeling the occurrence of a disease in fields or nurseries. The methodological approach involves a Gaussian field (GF) affected by a spatial process represented by an approximation to a Gaussian Markov random field (GMRF). This modeling allows the building of maps with prediction probabilities regarding the presence of a disease in plants using Bayesian kriging. The advantage of this modeling is its computational benefit when compared with known spatial hierarchical models and with the Bayesian inference based on Markov chain Monte Carlo (MCMC) methods. Inference through the use of the integrated nested Laplace approximation (INLA) with the stochastic partial differential equation (SPDE) approach facilitates the handling of large datasets in excellent computation times. Our approach allows the evaluation of different sampling strategies, from which we obtain inferences and prediction maps with similar behaviour to those obtained when we consider all subjects in the study population. The analysis of the different sampling strategies allows us to recognize the relevance of spatial components in the studied phenomenon. We demonstrate how Bayesian kriging can incorporate sources of uncertainty associated with the prediction parameters, which leads to more realistic and accurate estimation of the uncertainty. We illustrate the methodology with samplings of Citrus macrophylla affected by the tristeza virus (CTV) grown in a nursery.

Keywords: Bayesian kriging; Bayesian hierarchical models; Gaussian Markov random field (GMRF); integrated nested Laplace approximation (INLA); stochastic partial differential equation (SPDE)

Citation: Monsalve, N.C.; López-Quílez, A. Prediction and Surveillance Sampling Assessment in Plant Nurseries and Fields. *Appl. Sci.* **2022**, *12*, 9005. https://doi.org/10.3390/app12189005

Academic Editor: Roberto Romaniello

Received: 28 July 2022
Accepted: 5 September 2022
Published: 8 September 2022

1. Background

Geostatistical models are used to deal with situations in which we have observations made in a continuous region occurring in a defined spatial domain. In these models, the estimation of the response in unsampled locations can be seen as a prediction problem, and whether the response is normal or not is known as kriging prediction.

The simplest case of kriging prediction is the estimation of model parameters and the replacement of such estimates in the predictor equation as if they were the true values. This tends to be optimistic in the sense that it leads to an underestimation of the uncertainty, ignoring variability between parameter estimates and unknown true values. In a more realistic analysis, we rarely part from a Gaussian model, and the variogram is unknown since we do not know the real uncertainty of their parameters. However, from a Bayesian perspective, this estimation is not a problem.

Using Bayesian hierarchical models [1] and assuming non-Gaussian responses, we can take into account the uncertainty of the parameters. The key idea is to realize that these models can be considered structured additive regression (STAR) models [2]. This additive structure can be seen as a generalized linear mixed model depending on whether we add a random component. This type of model has also been applied to spatial point pattern data with random effects (or a hierarchical component) in other contexts [3].

Generalized mixed linear models have enjoyed a growing popularity due to their ability to model correlated observations. Their application range can go beyond the popular generalized linear models, but this involves more complex and difficult calculations. Various inference procedures have been proposed, among them the Bayesian approach through Markov chain Monte Carlo methods.

Bayesian hierarchical geostatistical models can be seen as a particular case of STAR models, where we assign Gaussian priors to all components of the additive predictor. In this case, we have latent Gaussian models [4], where variables are part of a latent Gaussian field. Gaussian fields play a dominant role in spatial statistics and in the geostatistical field [5–7] and are an important component in current spatial hierarchical models [1], since they provide one appropriate multivariate model with an explicit normalization constant and with good analytical properties.

The Gaussian processes and random fields have a long history, covering multiple approaches to representing spatial dependence and spatio-temporal dependence, such as covariance functions, spectral representations, reproducing kernel Hilbert spaces, and graph-based models [8]. The increasing popularity of Bayesian hierarchical models has made this situation very important because of the need to perform repeated simulations for model fitting, which may be impractical. Various methodologies have been proposed to address these limitations. Specifically, in this paper, we use the approach proposed by [9], in which a Gaussian field is approximated to a Gaussian Markov random field [4,10,11]. The main advantage of this representation is the remarkable improvement in calculation times and reduction in numerical difficulties associated with the analysis of generalized mixed linear models since the Gaussian Markov random fields are defined on sparse matrices and not on dense arrays.

This work describes how to use the intregrated nested Laplace approximation (INLA) with the stochastic partial differential equation (SPDE) approach via Bayesian hierarchical models. In addition, we illustrate the proposed methodology in an agricultural context and discuss the advantages of this methodology when choosing samples from a large population. This methodology can be used in other contexts and in other crops in order to improve knowledge of the movement patterns of agents causing a disease. This article presents a general strategy for estimation and prediction based on continuous spatial processes when we have observations made on a lattice of fixed locations via a Bayesian paradigm.

This article is organized as follows. After the background, in Section 2, we present the general form of the Bayesian hierarchical model proposed, with which we perform both the inference as the prediction of the presence of a disease in plants in unsampled locations and the descripton of a study case. In Section 3, we show the usefulness of our proposal in an agricultural context, and we present various sampling strategies in which the proposed methodology was applied. Finally, in Section 4, we present the discussion, concluding observations and future lines of research.

2. Method

In this section, we propose a structured additive regression (STAR) model for modeling the occurrence of a disease in fields or nurseries. The Bayesian hierarchical model with a spatial structure that we developed is able to predict the presence of a disease in plants in unobserved locations (Bayesian kriging), an event that in principle can originate in a lattice of fixed locations. The methodological approach involves a Gaussian field (GF) affected by a spatial process represented by an approximation to a Gaussian Markov random field (GMRF). The spatial effect is implemented through the stochastic partial differential equation (SPDE), and inference is done using the integrated nested Laplace approximation (INLA).

The Bayesian approach is appropriate for the spatial hierarchical model analysis because it allows both the observed data and model parameters to be considered as random variables [1], resulting in a more realistic and accurate estimation of uncertainty. With this

approach, it is easy to incorporate the prior information, which can be very helpful for distinguishing between spatial effects and ordinary linear non-spatial effects [12,13].

2.1. Bayesian Kriging for a Binary Response

In situations where we want to know the occurrence of an event of interest and the spatial process can be seen as a continuous process, we can follow [14] and consider a hierarchical model for geostatistical data. In particular, we describe Bayesian kriging and its application to the presence or absence of a disease in plants.

Specifically, let Y_i be a random Bernoulli variable that represents the presence (1) or absence (0) at location i ($i = 1, \ldots, n$), and let π_i be the probability of the presence of disease. We assume for Y_i

$$\begin{array}{c} Y_i \sim \text{Ber}(\pi_i) \\ \text{logit}(\pi_i) = \beta_0 + W_i \end{array} \tag{1}$$

where β_0 represents the intercept of the linear predictor, and W_i represents the random effect with a spatial structure, while the relationship between π_i and the covariates of interest and the random effect is modeled by the usual logit link. This proposal does not include the effect of covariates; therefore, the probability of the presence of disease is determined only by the intercept and by the spatial random effect.

W_i is assumed to be Gaussian with a covariance matrix $\sigma_W^2 H(\phi)$depending on the distance between locations and with hyperparameters σ_W^2 and ϕ representing the variance (partial sill) and the range of the spatial effect, respectively. We assume the following distribution for W_i:

$$W \sim N(0, \sigma_W^2 H(\phi)) \tag{2}$$

The structure of $H(\phi)$ is determined by the Matern function

$$C(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^{\nu} K_{\nu}(\kappa h)$$

K_{ν} is the modified Bessel function of the second-type and order $\nu > 0$. The parameter μ is usually fixed and measures the degree of smoothing of the whole process; its integer value determines the differentiability mean quadratic of the process. κ is a scale parameter related to the range. The spatial correlation function $C(H)$ depends on the locations s_i and s_j only through the Euclidean distance $h = \| s_i - s_j \| \in \Re$.

The modeling approach can be augmented by incorporating a pure error term known as the nugget effect in classic kriging. This effect describes the "noise" associated with the replication of measurement at each location, as is usual when using the Bayesian approach to assign a Gaussian distribution.

Under the Bayesian paradigm, we have to specify the prior distributions for each parameter involved in the model $(\beta_0, \sigma_W^2, \phi)$. In this regard, the usual choice is to deal with independent priors for the parameters [1], i.e.,

$$p(\beta_0, \sigma_W^2, \phi) = p(\beta_0) p(\sigma_W^2) p(\phi) \tag{3}$$

When we want to express some initial vague but useful knowledge about the parameters, a non-informative Gaussian prior distribution is usually chosen as candidate distributions for β_0 and an inverse gamma distribution for σ_W^2. The specification of ϕ will depend on the choice of the correlation function, which determines the covariance matrix H [1]. The final choice for the priors will depend on the type of modeling and parameterization chosen to be defined.

Expressions of (1)–(3) contain all our knowledge about the posterior distribution but do not produce closed expressions for the posterior distributions of the parameters. The general form of the posterior distribution for the variables $y = \{y_1, \ldots, y_n\}$ denoted by $\pi(y|x, \theta)$ with $\theta = (\theta_1^T, \theta_2^T)$ with dim $(\theta) = 2$ is the following:

$$\begin{aligned}\pi(x,\theta|y) &\propto \pi(\theta)\pi(x|\theta)\Pi_i\pi(y_i|x_i,\theta)\\ &\propto \pi(\theta)|Q(\theta)|^{n/2}\exp\left(-\tfrac{1}{2}x^TQ(\theta)x+\textstyle\sum_i \log\pi(y_i|x_i,\theta)\right)\end{aligned} \tag{4}$$

where $Q(\theta)$ is non-singular. The approximation to the marginal posterior of $\pi(x_i|y)$, $\pi(\theta|y)$ and $\pi(\theta_j|y)$ is discussed below.

2.2. Implementation of Bayesian Kriging with INLA

The basic idea of the modeling proposal is to realize that the hierarchical models (1) can be seen as structured additive regression (STAR) models. In other words, they are models in which the mean of the response variable Y_i is linked to a structured predictor which represents the effects of different covariates in an additive way. Specifically, these spatial hierarchical Bayesian models are the so-called latent Gaussian models [4], where Gaussian priors are assigned to all the components of the additive predictor. From this perspective, all latent Gaussian variables can be part of a vector and form the latent Gaussian field.

The approach based on integrated nested Laplace approximation (INLA) is a methodology introduced by [4,15] for statistical inference in latent Gaussian models. INLA provides a quick and efficient method of performing approximations to the marginal posterior density of the hyperparameters $\tilde{\pi}(\theta|y)$ and to the full conditionals of the posterior marginals of the latent variables $\tilde{\pi}(x_i|\theta,y)$, $i=1,\ldots,n$.

The key to this new approach of inference is in approximations to the marginal posterior of xi by nested approximations of

$$\tilde{\pi}(x_i|y)=\int \tilde{\pi}(x_i|\theta,y)\tilde{\pi}(\theta|y)d\theta \approx \sum_{k=1}^{K}\tilde{\pi}(x_i|\theta_k,y)\tilde{\pi}(\theta_k|y)\Delta_k \tag{5}$$

where $\tilde{\pi}(\cdot|\cdot)$ is an approximate conditional density. The approximations of (5) are calculated by approximations $\pi(\theta|y)$ y $\pi(x_i|\theta,y)$ using numerical integration (finite sum) on θ. The marginal posterior for the hyperparameters $\tilde{\pi}(\theta_j|y)$, $j=1,\ldots,m$ are determined similarly.

The inference is based on the approximation $\tilde{\pi}(\theta|y)$ of the marginal posterior of θ:

$$\tilde{\pi}(\theta|y)\propto \frac{\pi(x,\theta,y)}{\tilde{\pi}_G(x|\theta,y)}\Big|_{x=x^*(\theta)} \tag{6}$$

where $\tilde{\pi}_G(x|\theta,y)$ is the Gaussian approximation of the full conditionals x, and $x^*(\theta)$ is the mode of the full conditional of x for a given θ. The sign of proportionality is due to the fact that the normalization constant for $\pi(x,\theta|Y)$ is unknown. This expression is equivalent to a Laplace approximation [16] and this suggests that the approximation error is relative and of order $O(n^{-3/2})$ after renormalization.

Note that $\tilde{\pi}(\theta|y)$ tends to stray too far from Gaussianity; therefore, this approach determines approximations of $\tilde{\pi}(\theta|y)$ and $\tilde{\pi}(x_i|y)$ in a non-parametric way. The main tool for inference is in the application of the Laplace approximation to $\pi(x_i|\theta,y)$.

The Matern covariance function appears naturally in various scientific fields [17]. However, Ref. [9] establish an approximation between the Gaussian field and the Matern covariance function using a weak stochastic approximation to a stochastic partial differential equation (SPDE) as follows:

$$\begin{gathered}(\kappa^2-\Delta)^{\alpha/2}x(u)=W(u),\ u\in\Re^d\\ \alpha=\nu+d/2,\ \kappa>0,\ \nu>0\end{gathered} \tag{7}$$

where $(\kappa^2-\Delta)^{\alpha/2}$ is a pseudo-differential operator. W has a Gaussian distribution of white noise with unitary variance; Δ is the Laplacian

$$\Delta=\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2} \tag{8}$$

and marginal variance

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}} \tag{9}$$

Hereinafter, any solution of the form (7) is called a Matern field. The solutions limit the SPDE approach when $\kappa \to 0$ or $\nu \to 0$ do not have Matern covariance functions. However, there is a solution when $\kappa = 0$ or $\nu = 0$ whether the random measures are well defined. When $\alpha \geq 2$, the null space of the differential operator is not trivial and contains, for example, the functions $\exp(\kappa e^T u)$ for all $\|e\| = 1$. Matern fields are the only stationary solutions to a stochastic partial differential equation (SPDE).

The goal of the SPDE approach is to find a Gaussian Markov random field (GMRF) with a neighborhood structure and a precision sparse matrix Q that best represents the Matern field. Given this representation, inferences can be made using the GMRF found and its good properties of calculation.

Basically, the SPDE approach uses a finite representation to define the Matern field as a linear combination of base functions, which are defined by a triangulation into domain D. This triangulation divides D into a set of triangles that are not intercepted and united by at least one common edge or corner. First, the initial vertices of the triangles are placed at locations $s_1, \ldots, s_n$, and later, additional vertices are added in order to obtain a useful triangulation desired for spatial prediction.

Considering the triangulation, the representation of the base function of a Matern field $X(s)$ is given by

$$X(s) = \sum_{i=1}^{n} \psi_l(s) w_l \tag{10}$$

where n is the total number of vertices, $\{\psi_l(s)\}$ are the basis functions and w_l are weights with a Gaussian distribution. The functions $\{\psi_l(s)\}$ are selected as linear segments in each triangle, i.e., $\psi_l(s)$ is 1 on vertex l and 0 on the other vertices. The height of each triangle (the space field value at each vertex of the triangle) is given by the weight w_l, and the values inside the triangle are determined by linear interpolation.

The key point of the SPDE approach is the finite representation (10) that establishes the link between Gaussian fields (GF) $X(s)$ and the GMRF defined by Gaussian weights w_l. These weights may be assigned a Markovian structure as shown in [9].

In particular, the precision matrix Q of GMRF is defined by the equation $w_l \sim N(0, Q_S^{-1})$ as a function of κ^2 for $\alpha = 1, 2, \ldots, \nu = 0, 1, 2, \ldots$ and $\alpha = \nu + 1$.

Under this perspective, for each vertex $i = 1, \cdots, n$, the full hierarchical model structure can be stated as follows:

$$\begin{gathered} Y_i \sim \text{Ber}(ß_i) \\ \text{logit}(\pi_i) = \beta_0 + W_i \\ \pi(\beta_0) \propto 1 \\ W \sim N(0, \sigma_W^2 H(\phi)) \end{gathered} \tag{11}$$

In contrast to WinBUGS software [18], regarding assignation of the priors, the correlation function is not modeled directly. In this case, the numerical solution to the Gaussian field is made using an approximate weak solution to a stochastic partial differential equation (SPDE) as a Gaussian Markov random field [4,15]. This solution requires defining two new parameters, κ and τ, which determine the range of the spatial effect and the total variance. More precisely, the range is approximated by the expression $\phi \approx \sqrt{8}/\kappa$, while the variance is defined as $\sigma_W^2 = \frac{1}{4\pi\kappa^2\tau^2}$. κ and τ have the following prior distributions:

$$\begin{gathered} 2\log\kappa \sim N(m_\kappa, \sigma_\kappa^2) \\ \log\tau \sim N(m_\tau, \sigma_\tau^2) \end{gathered} \tag{12}$$

Thus, the variance of the spatial component $\sigma_W^2 H(\phi)$ is given by the matrix $Q(\theta)$. We assume the reparameterization $\theta_1 = \log\tau$ and $\theta_2 = 2\log\kappa$. The mean m_κ is chosen

reasonably according to the size of the region, while the mean m_τ is chosen so that the variation of the field is 1.

2.3. Research Object

The utility of the proposed methodology is illustrated through a dataset obtained from a nursery, formed of 10,920 *Citrus macrophylla* saplings. The 10,920 saplings are distributed in 40 rows of 273 saplings each. The saplings are placed on 20 mounds. Each mound is compound of two rows of saplings. The distance between any two saplings in the same row is between 15 and 18 cm; however, the final distance considered between every two saplings is the midpoint between 15 and 18 cm, i.e., 16.5 cm. Furthermore, the distance between two rows within a mound is 40 cm, and the distance between two adjacent rows of a different mound is 70 cm.

The analysis was performed on 10,920 saplings in search of the *Citrus Tristeza virus* (CTV; Family: *Closteroviridae*; Genus: *Closterovirus*). Figure 1 shows the distribution of the disease throughout the nursery. There is a total of 443 diseased saplings (red dots), representing an infection rate of 4.05%. The nursery is seen as a continuous region in which any sapling may become diseased with the virus at any point given the proximity of the saplings. This consideration can be due to the large number of saplings planted and to the low proportion of saplings infected with the citrus tristeza virus.

Figure 1 only represents the triangulation underlying the SPDE approach that we have applied and which is based Bayesian kriging. Each vertex of the mesh is an observed point or a point prediction, with the red dots indicating infected saplings and black dots representing uninfected saplings.

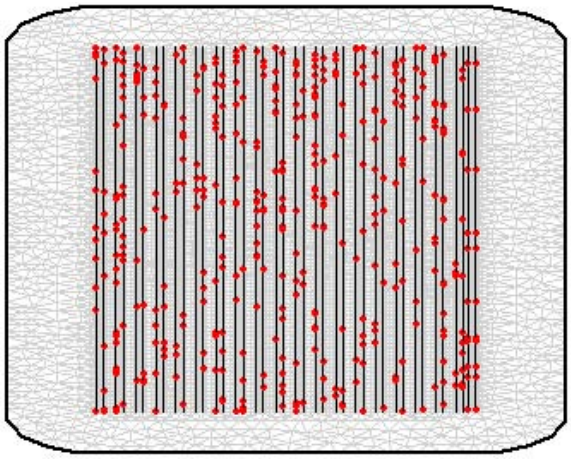


Figure 1. Sampling locations for the presence and absence of CTV in the nursery—mesh.

By applying the model defined in Section 2.2 over the observed dataset, we obtained the estimation of the posterior parameters of interest shown in Table 1.

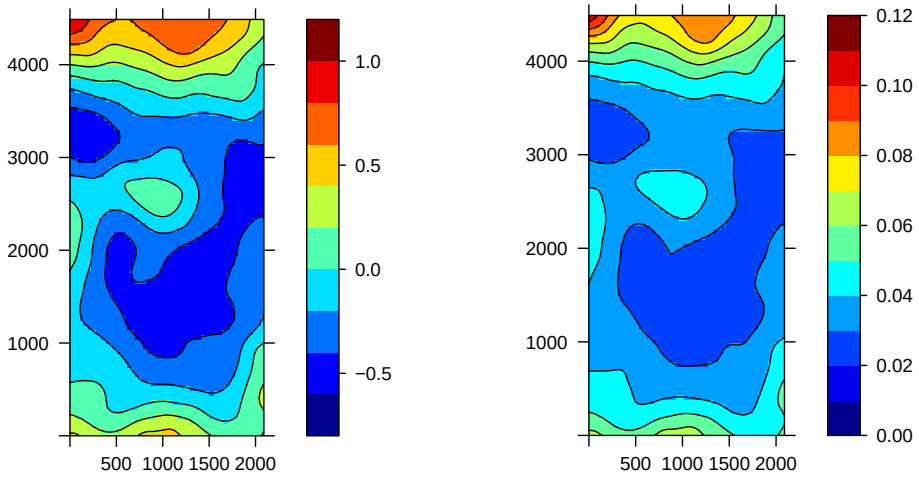
Table 1. Posterior distribution of the parameters using the complete dataset.

Parameter	Mean	S.D.	$Q_{0.025}$	$Q_{0.50}$	$Q_{0.975}$
β_0 (Intercept)	−3.13	0.31	−3.77	−3.14	−2.45
κ	0.00217	0.0063	0.00074	0.0019	0.0049
τ (cm)	3.55	0.0153	1.198	3.12	8.38

Q = Quantile; S.D. = standard desviation.

According to the definition of ϕ, we find that the range is equal to 1302.869 or approximately about 13 cm. Since this is the distance where the correlation is close to 0.10, it can be inferred that data are characterized by a strong correlation for distances less than or equal to 13 cm. Therefore, we can conclude that correlation decreases after this distance. Clearly the presence of the disease is determined by a spatial effect. In particular, infection occurs between plants located in the same row over distances of less than or equal to 13 cm in any of the mounds.

Figure 2a shows the posterior mean of spatial effect W_i. We observe how the spatial component reaches positive values to the north and south of the nursery, as well as negative values, and achieves values close to zero in the center. In this map, we may recognize the areas with higher risks. This can be explained by the action of the wind that introduces the aphids infected with the virus to the nursery and which carry the disease. The variance in random spatial effect σ^2_W is equal to 0.13. As the variability takes a small value, we can conclude that the spatial component reflects the pattern of contagion between close saplings located in the same row.

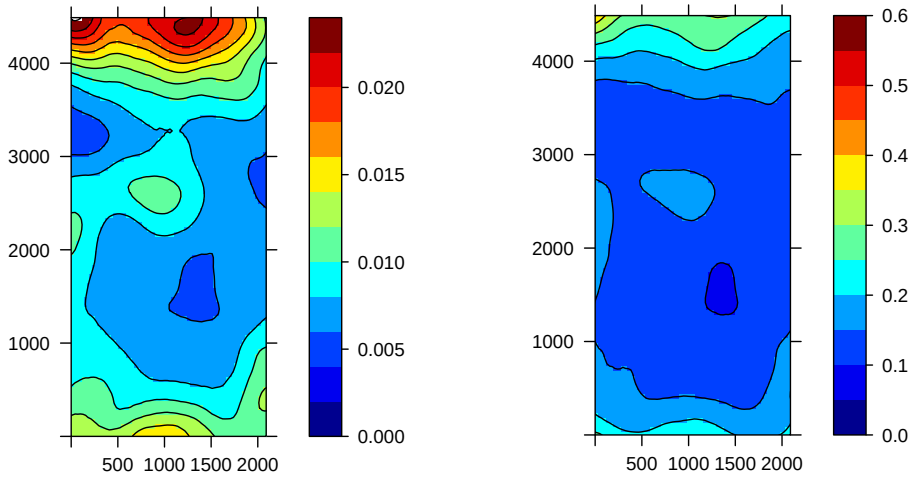


(**a**) (**b**)

Figure 2. Posterior mean of the spatial effect (**a**) and of $\pi_i|Y$ (**b**) using the complete dataset.

In order to understand the behavior of disease in the nursery, maps were generated with an estimate of the probabilities ($\pi_i|Y$) both at observed sites as well as unobserved sites. Figure 2b shows the posterior mean probability $\pi_i|Y$, while Figure 3a,b shows the quartiles of $\pi_i|Y$. Thus, we obtain not only a point estimate of the probability of the disease of a subject but also an assessment of the uncertainty in this estimate. These figures confirm that the probability of finding the tristeza virus is higher towards the edges of the nursery where the influence of the wind is present.

This type of modeling was used in the context of fisheries, with very good results [19]. The estimates and results presented in this paper were obtained using the INLA program through the R [20] package of the same name [21].



(**a**) (**b**)

Figure 3. First (**a**) and third (**b**) quartile for $\pi_i|Y$ for the complete dataset.

3. Results

3.1. Sampling Assessment in the Study of Agricultural Diseases

In some areas we may need to know information about a population to determine the prevalence and infection rate of a virus because of interest in the presence/absence of a disease or because we just want to know the causes or possible risk factors using epidemiological studies. In either case, analyzing the entire population is very expensive, and in some cases impossible.

Instead of analyzing all individuals in the population under study, it is preferable to measure the variables in a part of it, obtaining a sample. Working with a sample has the advantages of being faster and cheaper. Additionally, if the sample is chosen correctly, the information obtained leads to reasonable and reliable estimates.

When the goal is to understand the dynamics of a disease, or when we want to conduct an epidemiological study whose results can be extrapolated to a general population, a prerequisite is that the sample is representative. The best option to obtaining a representative sample is to randomly choose the individuals through a random sampling method.

Random sampling can be performed in different ways, the most common methods being single sampling, systematic, stratified, and cluster sampling. In random sampling, all individuals have the same probability of being chosen. The elements that form part of the sample are randomly selected using random numbers.

The 10,920 saplings are for experimental purposes and are analyzed in order to study the behavior of the tristeza virus in a controlled environment. The key for achieving this purpose is to analyze only one part of the subjects, i.e., to get a sample of all the nursery. Therefore, in situations such as this, it is very beneficial to determine sampling strategies that help understand of the dynamics of the disease with the lowest investment of resources, i.e., without analyzing all the saplings. In this sense, we propose a calibration procedure for the sample and a comparison between various sampling methods.

3.2. Calibration Sampling

In situations similar to that analyzed, where there are a large number of saplings to analyze, it is advisable to perform a calibration process before the sampling method. Through this process, we can know which areas of the nursery present with an increased risk of infection. This process can help us to propose sampling methods that can improve sample selection.

The calibration process begins when dividing the nursery into 9 horizontal bands, each consisting of 500 points depending on the coordinate values of x for the sapling i with $i = 1, \ldots, n = 10{,}920$. Figure 4 illustrates the configuration of the bands schematically.

To find the probabilities in all locations of the nursery (point observed, prediction point), we fit the proposal modeling under the INLA-SPDE approach, and we use the mesh or triangulation shown in Figure 1. Thanks to the projection that we built on triangulation, it is possible to study the spatial process as a continuous process. In Figure 4, we can observe that the greatest risks are located in bands 1 and 9 of the nursery.

Once we conclude the calibration process, we can determine which sampling methods provide the greatest benefits in terms of sample size and prediction errors. For this reason, we calculated the measures of discrepancy between the probabilities obtained throughout the nursery (projection based on triangulation defined in Figure 1) and those obtained under different sampling methods. The projection points in each sample are defined on the triangulation defined for all the nursery (Figure 1). This way, we can compare probabilities at the same points of projection.

The discrepancy measures or the prediction errors are obtained for each random sampling from successive simulations. The measures defined are: mean square error (MSE), absolute error (ABSE) and coeficient of variation (CV).

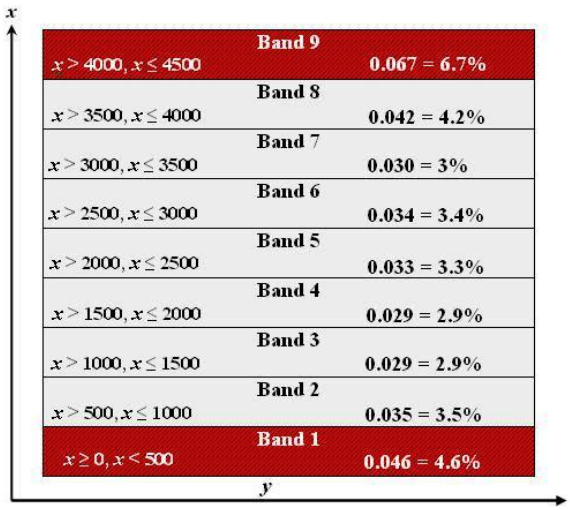


Figure 4. Configuration of nursery for the calibration process; values for x (**left**) and percentage of diseased saplings (**right**).

3.3. Application of the Sampling Methods

Table 2 shows the summary of the measurements obtained using random sampling. This method is conceptually simple and involves taking the individuals at random from a list. In our case, the list is made up of 10,920 saplings.

In order to propose the best sampling design, we tested various percentages and sample sizes under different sampling strategies. Table 2 presents both the estimated parameters and the different measures of error for random samples of 10%, 15%, 20% and 25%.

Under this method, smaller errors occur in those random samples of 25%. With 25% of representation, smaller discrepancies are obtained under the mean square error and the coefficient of variation. However, no significant improvements in the absolute error are found.

Table 3 presents the summary of the measurements obtained using a systematic sampling. In this case, the first subject is chosen randomly, and the rest are conditioned by this choice. This method is simple to apply in practice and has the advantage of not needing a framework survey. It can be applied in most situations, and the only precaution that should be considered is to check that the characteristic studied does not have a periodicity coincident with the sample.

Table 2. Posterior distribution of the parameters and prediction errors for simple random samples considered.

Measure	Sample 10%	Sample 15%	Sample 20%	Sample 25%
κ	0.0058	0.0084	0.0044	0.0029
ϕ (cm)	6.26	3.83	8.39	12.44
τ (cm)	19.44	11.81	6.76	4.82
MSE	3.1419	2.8638	2.8448	2.4847
ABSE	0.0134	0.0116	0.0118	0.0106
CV	0.2025	0.2183	0.2075	0.1689

In the samples that represent 50%, we defined a systematic jump equal to two, i.e., we consider one in every 2 plants. In samples that represent 25%, we consider one in every four plants, while for the remaining percentages, one in every five and one in every eleven, respectively, was chosen.

Table 3. Posterior distribution of the parameters and prediction errors for the systematic samples considered.

Measure	Sample 50%	Sample 25%	Sample 20%	Sample 9%
κ	0.0032	0.0046	0.0036	0.0026
ϕ (cm)	8.58	6.03	7.70	10.81
τ (cm)	7.98	4.33	4.20	4.80
MSE	2.4409	2.7979	2.9392	3.7203
ABSE	0.0109	0.0109	0.0110	0.0130
CV	0.1775	0.2017	0.2391	0.3313

Clearly, with a systematic 50% sample, we have smaller errors due to the large size of the sample. However, when analysing the rest of the systematic samples, we observe that the 25% samples have smaller errors. The discrepancy measure of these samples compared with the 20% samples is similar. However, the mean square error (MSE) for the latter is greater. The measure that recognizes greater discrepancy remains the mean square error.

Finally, according to the results in Table 3 and due to the need to study both smaller samples as the most representative, we may conclude that under this scheme, it is advisable to select a 25% sample of all saplings in the nursery.

The third method used is a mixed sampling method, where stratified sampling is combined with the random sampling method. In this case, we build blocks or layers considering the value of x coordinate of each tree i with $i = 1, \ldots, 10{,}920$. We form three blocks or subsets of data, the first block being made up of those subjects located in band 1, the second by the saplings located in bands 2, 3, 4, 5, 6, 7 and 8, and the third block by saplings located in band 9. Once the blocks are formed, we take random samples of different sizes in each of them (see Table 4).

Table 5 presents a summary of the measures obtained using a mixed sampling and shows that as n increases, it progressively decreases the error measure considered. However, the differences between the errors of the last samples is not significant. Therefore, in such situations, a mixed sampling method can be used with sample sizes of 30% in block 1, 30% in block 2 and 20% in block 3, or samples in each block of 35% , 35% and 25%, respectively. With both sampling schemes, we have approximately 23% and 31% of all the nursery. As in the other sampling methods, the mean square error (MSE) is the statistic which recognizes major differences.

Table 4. Percentages used in the stratified random sampling method.

	Block 1	**Block 2**	**Block 3**	**Total**
100%	20%	20%	10%	12.23%
N	1240	1200	8480	10,920
n	248	240	848	1336
100%	25%	25%	15%	16.67%
N	1240	1200	8480	10,920
n	310	300	1211	1821
100%	30%	30%	20%	22.97%
N	1240	1200	8480	10,920
n	413	400	1696	2509
100%	35%	35%	25%	30.58%
N	1240	1200	8480	10,920
n	620	600	2120	3340

Table 5. Posterior distribution of the parameters and prediction errors in the stratified random samples considered.

Measure	**20%, 20%, 10%**	**25%, 25%, 15%**	**30%, 30%, 20%**	**35%, 35%, 25%**
κ	0.0042	0.0031	0.0024	0.0091
ϕ (cm)	6.98	10.36	14.68	9.35
τ (cm)	9.82	3.80	3.917	2.91
MSE	8.2617	4.0836	3.4858	3.0376
ABSE	0.0195	0.0140	0.0130	0.0124
CV	0.3985	0.3264	0.2969	0.2505

3.4. Estimation and Prediction from Sampling Assessment

Having evaluated the effect of choosing random samples under different sampling methods, we observed that the most appropriate sampling method when we have data of this nature is a random samples method with equal percentages of 25%. With these samples, we obtained lower prediction errors. This result seems appropriate since it is an intermediate sampling scheme between a sample percentage recommended by [22] and the 10% reported by most literature on sampling.

The proposed modeling approach is able to capture the behavior of the entire dataset when we consider smaller sample sizes. This was not only observed with random samples of 25% but in the maps obtained from the random samples of 10%, 15% and 20%. Similar behavior was obtained in the maps generated from the other sampling methods considered.

The modeling is robust in the presence of little data and in the absence of explanatory variables. Once again, we recognized the importance of a random spatial effect and its determining effect on the dynamics of the disease (Figure 5a). The greatest risks are still presented towards the edges of the nursery and the northern and southern borders, where there is greater wind action (Figure 5b).

We only show the maps of the spatial effect and the estimated uncertainties of the posterior probabilities for random samples of 25%. By comparing the map obtained with random samples of 25% with the map of the entire nursery, we observed similar behavior between them. That is, the pattern observed in the maps using the entire data is preserved and can be seen in the maps of random samples of 25% (Figure 6a,b).

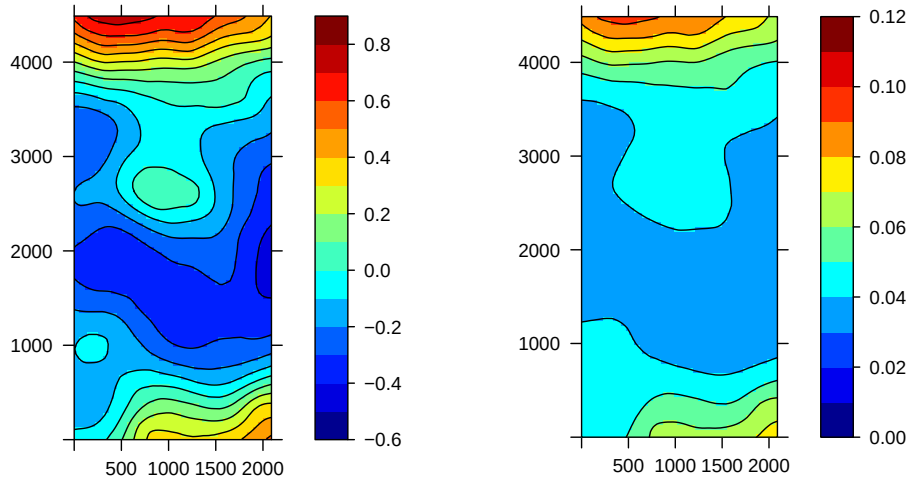


Figure 5. Posterior mean of the spatial effect (**a**) and of $\pi_i|Y$ (**b**) for random samples of 25%.

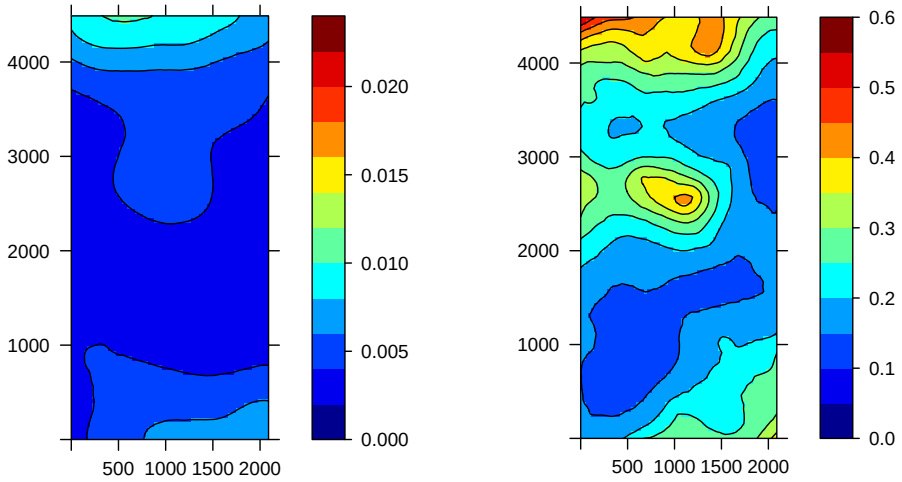


Figure 6. First (**a**) and third (**b**) quartiles of $\pi_i|Y$ for random samples of 25%.

Following the definitions for the range ϕ and the total variance of the spatial effect σ_W^2, we have, respectively, that the spatial correlation maximum is reached in approximately 975.32, i.e., in 9.75 cm, while the variability equals 0.041. Comparing these estimates with those obtained using the entire dataset for the nursery, we have that both the range (ϕ) and the spatial variance (σ_W^2) are lower in random samples of 25%. This result seems logical and is a product of the sampling method. However, despite these differences, we can say that the range continues to describe the existence of a pattern of contagion among nearby saplings, i.e., between saplings located in the same row. This behaviour is observed especially in those rows located to the north and south of the nursery.

The validity of the modeling is verified through the effective number of parameters p_D derived from calculation of the DIC and predictive measures generated from the main calculations of INLA. In particular, the predictive measure evaluated was the probability integral transform (PIT). In all the adjusted models, we obtained values of pD less than the number of data considered. On the other hand, histograms by PIT showed behavior close to a uniform distribution in all cases.

Our modeling does not consider augmenting the additive structure with the influence of a pure error term (called the nugget effect in kriging terminology) since we find that this random effect has a very large variability. This shows that in situations like the one analyzed, this effect does not distinguish sources of variability different than the variability caused by the spatial random effect. This result agrees with the findings of [11], who

demonstrated that when using binomial models, there is a marked sensitivity resulting from the choice of priors assumed on the random effect precision parameters.

4. Discussion

First, we demonstrate that our proposed method of modeling may capture behavior patterns and draw risk maps which are similar to those obtained from the population from which the sample has been extracted. From risk maps and from estimates of the parameters, surveillance strategies and policies can be established to control the distribution of diseases in agricultural contexts.

Secondly, we demonstrate that this methodology allows prediction maps of uncertainty to be built relatively simply and quickly when combining the SPDE approach with the INLA methodology. This approach facilitates the handling of large datasets with excellent computation times. The biggest advantage of our modeling compared with classical geostatistical methods is its computational benefit in performing adjustment and prediction.

Another advantage of this approach is its generality, as it allows us to perform a Bayesian analysis directly and calculate criteria and various predictive measures that facilitate the comparison of complex models.

Thirdly, the INLA methodology can be used in any type of spatial data, and it can even deal with continuous nonstationary and anisotropic phenomena.

We also demonstrate how we obtain consistent inferences from samples in epidemiological studies that handle a significant amount of data and demonstrate the utility of the methodology in the presence of few data and in the absence of explanatory variables.

It is clear that, by using Bayesian kriging, it is possible to incorporate sources of uncertainty into the model associated with the prediction parameters and thus find more realistic estimates, contrary to what happens in classical geostatistics [7]. The INLA methodology combined with the SPDE approach offers an excellent theoretical framework for phenomena that need prediction where fixed effects, structured and unstructured Gaussian random effects are combined linearly in a linear predictor, and the elements of the linear predictor are observed through one or more likelihoods [21]. The illustration of the methodology with real data allows us to recognize its usefulness in epidemiological studies, not only in the agricultural context.

The Bayesian INLA-SPDE approach is a complementary method to the traditional spatial data mining approach to obtain the prediction of critical points of vulnerable species and accurately inform management decisions [23]. The proposed methodology recognizes small-scale spatial dependence, but it may be interesting to add a component or surface that can capture and improve large-scale variability and improve risk smoothing as a future line of research.

Another future line of research would be tp extend a class of additive spatio-temporal SPDE models and investigate the distinction between them, (1) extending their temporal effect, allowing a random walk process in time, (2) varying the spatial correlation function, and (3) running a simulation study to evaluate the effect of erroneously specifying spatial and temporal models [24,25].

Finally, we conclude that in situations similar to those analyzed where there are a large number of saplings to study, it is advisable to perform different sampling methods. The results we have found from the proposed modeling show it is possible to evaluate different sampling strategies, from which we can get the best results, and with which it is possible to obtain more realistic estimates and prediction maps with behaviour similar to that observed in the study population.

Author Contributions: Conceptualisation, A.L.-Q.; Methodology, N.C.M. and A.L.-Q.; Software, N.C.M.; Writing—original draft preparation, N.C.M.; Writing—review and editing, N.C.M. and A.L.-Q.; Supervision, A.L.-Q.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grant PID2019-106341GB-I00 from the Spanish Ministerio de Ciencia e Innovación—Agencia Estatal de Investigación, jointly financed by the European Regional Development Funds (ERDF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors want to thank the members of the Instituto Valenciano de Investigaciones Agrarias for providing the dataset used in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSE	Absolute Square Error
CTV	Citrus Tristeza Virus
CV	Coefficient of Variation
DIC	Deviance Information Criterion
GF	Gaussian Field
GMRF	Gaussian Markov Random Field
INLA	Intregrated Nested Laplace Approximation
MCMC	Markov Chain Monte Carlo
MSE	Mean Square Error
PIT	Probability Integral Transform
SPDE	Stochastic Partial Differential Equation
STAR	Structured Additive Regression

References

1. Banerjee, S.; Carlin, B.P.; Gelf, A.E. *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed.; Chapman & Hall/CRC Monographs on Statistics & Applied Probability; Chapman & Hall/CRC: Boca Raton, FL, USA, 2004.
2. Chien, L.C.; Bangdiwala, S.I.; The implementation of Bayesian structural additive regression models in multi-city time series air pollution and human health studies. *Stoch. Environ. Res. Risk Assess.* **2012**, *26*, 1041–1051. [CrossRef]
3. King, R.; Illian, J.B.; King, S.E.; Nightingale, G.F.; Hendrichsen, D.K.; A Bayesian Approach to Fitting Gibbs Processes with Temporal Random Effects. *J. Agric. Biol. Environ. Stat.* **2012**, *17*, 601–622. [CrossRef]
4. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 319–392. [CrossRef]
5. Cressie, N. *Statistics for Spatial Data*; Wiley: New York, NY, USA, 1993.
6. Stein, M.L. *Interpolation of Spatial Data: Some Theory for Kriging*; Springer: New York, NY, USA, 1999.
7. Diggle, P.J.; Ribeiro, P.J. *Model-based Geostatistics*; Springer: Berlin/Heidelberg, Germany, 2007.
8. Lindgren, F.; Bolin, D.; Rue, H. The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spat. Stat.* **2022**, *50*, 100599. [CrossRef]
9. Lindgren, F.; Rue, H.; Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *J. R. Stat. Soc. Ser. B* **2011**, *73*, 423–498. [CrossRef]
10. Moraga, P.; Dean, C.; Inoue, J.; Morawiecki, P.; Noureen, S.R.; Wang, F. Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique. *Spat Spatio-Temporal Epidemiol.* **2021**, *39*, 100440. https:/doi.org/10.1016/j.sste.2021.100440. [CrossRef] [PubMed]
11. Roos, M.; Held, L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Int. Soc. Bayesian Anal.* **2011**, *6*, 259–278. [CrossRef]
12. Finley, A.O.; Banerjee, S.; Ek, A.R.; McRoberts, R.E. Bayesian multivariate process modeling for prediction of forest attributes. *J. Agric. Biol. Environ. Stat.* **2008**, *13*, 60–83. [CrossRef]
13. Gaudard, M.; Ramsey, P.; Stephens, M. Interactive Data Mining and Design of Experiments: The JMP® Partition and Custom Design Plataforms, Group. 26. 2006. Available online: http://northhavengroup.com/pdfs/PartitionandDOEFinalCopy.pdf (accessed on 1 August 2022).
14. Diggle, P.J.; Tawn, J.A.; Moyeed, R.A. Model-based geostatistics (with discussion). *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1998**, *47*, 299–350. [CrossRef]
15. Rue, H.; Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *J. Stat. Plan. Inference* **2007**, *137*, 3177–3192. [CrossRef]

16. Tierney, L.; Kadane, J.B. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86. [CrossRef]
17. Guttorp, P.; Gneiting, T. Studies in the history of probability and statistics XLIX on the Matern correlation family. *Biometrika* **2006**, *93*, 989–995. [CrossRef]
18. Spiegelhalter, D.J.; Thomas, A.; Best, N.; Lunn, D. *WinBUGS User Manual*; Version 1.4; MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine: London, UK, 2003. Available online: https://www.mrc-bsu.cam.ac.uk/software/bugs/ (accessed on 1 February 2019).
19. Munoz, F.M.; Pennino, M.G.; Conesa, D.; López-Quílez, A.; Bellido, J.M. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian model. *Stoch. Environ. Res. Risk Assess.* **2013**, *27*, 1171–1180. [CrossRef]
20. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: http://www.R-project.org (accessed on 12 August 2022).
21. Rue, H.; Riebler, A.; Sørbye, S.H.; Illian, J.B.; Simpson, D.P.; Lindgren, F.K. Bayesian computing with INLA: A review. *Annu. Rev. Stat. Its Appl.* **2017**, *4*, 395–421. [CrossRef]
22. Gottwald, T.R.; DaGraça, J.V.; Bassanezi, R.B. Citrus huanglongbing: The pathogen and its impact. *Plant Health Prog.* **2007**, *8*, 31. [CrossRef]
23. Lezama-Ochoa, N.; Pennino, M.G.; Hall, M.; López, J.; Murua, H. Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular mobular). *Sci. Rep.* **2020**, *10*, 18822. [CrossRef] [PubMed]
24. Kifle, Y.W.; Hens, N.; Faes, C. Using additive and coupled spatiotemporal SPDE models: A flexible illustration for predicting occurrence of Culicoides species. *Spat. Spatio-Temporal Epidemiol.* **2017**, *23*, 11–34. [CrossRef] [PubMed]
25. Fioravanti, G.; Martino, S.; Carmeletti, M.; Cattani, G. Spatio-temporal modelling of PM_10 daily concentrations in Italy using the SPDE approach. *Atmos. Environ.* **2021**, *248*, 118192. [CrossRef]