OXFORD

Journal of Geophysics and Engineering

Journal of Geophysics and Engineering (2023) **20**, 523–537 https://doi.org/10.1093/jge/gxad023

Hydrogeological spatial modelling: A comparison between frequentist and Bayesian statistics

Jason M. Romero[1,2,*], **Daniel C. Salazar**[3] **and Carlos E. Melo**[1,2]

[1] Faculty of Engineering, Universidad Distrital Francisco José de Caldas, 110231, Bogotá D.C., Colombia
[2] Master in Information and Communications Science, Universidad Distrital Francisco José de Caldas, 110231, Bogotá D.C., Colombia
[3] Cadastral Engineering and Geodesy, Universidad Distrital Francisco José de Caldas, 110231, Bogotá D.C., Colombia

*Corresponding author: Jason M. Romero. E-mail: jamromeror@correo.udistrital.edu.co

Received 21 June 2022, revised 3 January 2023

Abstract

Traditional and modern spatial prediction techniques are applied in the analysis of water quality, evidencing a new approach that allows the modeling of a hydrogeological system in the central area of Boyacá, Colombia. The objective is to determine the quality status of groundwater for human consumption. In the process, spatial predictions were made based on frequentist methods (kriging, cokriging) and Bayesian methods (R-INLA stochastic partial differential equation) as an alternative to Markov chain Monte Carlo methods that require a large computational cost. From the application of these methods, a comparison is made by statistical tests that determine the goodness of adjustment of the predictions. Our interest here is in the implementation of future more robust, economic and scientific solutions, particularly for hydrogeological data, and the proper management of water resources. Finally, the vulnerability of aquifers is analyzed with the DRASTIC method, which takes into account the surrounding media of the aquifers through the variables depth (D), net recharge (R), type of aquifer (A), soil (S), topography (T), impact of the vadose zone (I) and hydraulic conductivity (C).

Keywords: Kriging, INLA, Bayesian methods, water quality

1. Introduction

Water, as a basic resource for living beings and human settlements, is essential for the socioeconomic progress of the population. The importance of this resource lies in the environmental services of supply and potable use in productive activities such as agriculture and industry. However, the growth of cities and the economic problems associated with high demand for water increases the pressure on water sources (UNESCO 2014), having as determinants the demographic factors represented in increasing demand for these resources (Montaño 2019). In the central zone of Boyacá, the water resource is susceptible to shortages, being the third most affected state in the country as regards the capacity of its resource. Furthermore, Boyacá is one of the states with the highest percentage of pesticide use, mainly organophosphorus pesticides, that require an adequate water management strategy. Therefore, estimation of quality indices for the hydrogeological system is sought in the central area of Boyacá, which includes several physical and chemical parameters that do not immediately provide a proper diagnosis themselves, since their variation throughout the territory must be considered, and so a broader picture of the problem is necessary, to infer the state of the resource in those areas where it was not possible to obtain measurements, in order to allocate better use of the resource for its protection and conservation.

For continuous estimation, most of the methods used to date in the evaluation of spatial uncertainty for the prediction of water quality indices are within the frequentist framework or conventional geostatistical approach that is part of

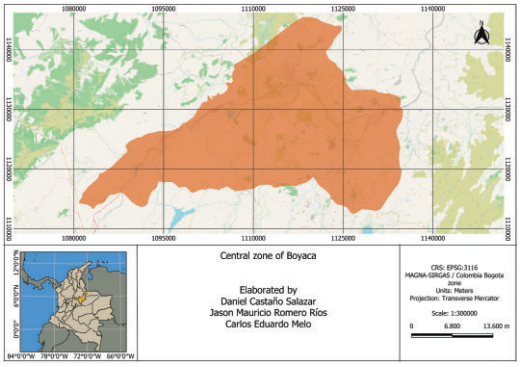


Figure 1. Central location of Boyacá. See https://rpubs.com/dcastsala413/boyaca.

Gaussian fields (GFs). Given the need to generate more accurate and cost-effective predictions, in the last few years many methods of spatial design with different approaches have been developed. Bayesian inference turns out to be a good option to analyze spatial hierarchical models, allowing the observed data and the model parameters to be random variables (Banerjee *et al.* 2014). The integrated nested Laplace approach (INLA) methodology (Rue *et al.* 2009) has major computational advantages in generating spatial models with a Bayesian approach (Lezama-Ochoa *et al.* 2020). Some of the most recent developments in this methodology were carried out by (Blangiardo & Cameletti 2015), (Lindgren 2018), (Moraga 2019), (Gómez-Rubio 2020) and (Vilela *et al.* 2021), generating interest in comparing the methodology to conventional geostatistics with Bayesian geostatistical methods, applied to hydrogeologic data measured in the central zone of Boyacá, Colombia.

Sections 2 and 3 describe the location and data of the case study; the theoretical approach of the methods used are found in sections 4 and 5. The calculation of indices of quality and vulnerability is presented in section 6, followed by the results, presented in sections 7 and 8. The document ends with a discussion and conclusions.

2. Generalities

2.1. Location

The study area, shown in figure 1, is located in Boyacá, Colombia, on the eastern mountain range in the Andean natural region, partially covering the municipalities of Paipa, Tuta, Sotaquirá, Cómbita, Duitama, Tibasosa, Firavitoba, Santa Rosa de Viterbo, Sogamoso and Nobsa with an area of 859 km^2, which covers the central part of Boyacá. In the region, the annual precipitation varies between 500 and 1000 mm yr^{-1}, and the lowest values are in the valley of the Chicamocha River with variations between 150 and 330 mm yr^{-1}. As for the temperature, the zone presents an average of 14°C, with daily variations ranging from 5°C to 24°C

Table 1. Rating of the water quality index. (Brown *et al.* 1970).

WQI value	Categories
<50	Excellent
50–100	Good
100–200	Poor
200–300	Very poor
>300	Unfit for human consumption

2.2. Materials and methods

The data were obtained from the Colombian Geologic Service (SGC) by the Groundwater Exploration Group, formulating the hydrogeologycal conceptual model of the state of Boyacá and representing point measurements of hydrogeochemical properties. The data correspond to 275 measurements at points of water extraction (wells, cisterns and springs) in the year 2016, making a collection of physical and chemical parameters: electrical conductivity, pH, temperature, calcium (Ca), magnesium (Mg), nitrates (NO_3), manganese (Mn), sulfate (SO_4), carbonates (HCO_3), alkalinity, hardness, turbidity, sodium (Na), total iron (FeT), potassium (K), chlorine (Cl), color and phosphate (PO_4), according to Murugesan *et al.* (2010), Krishna kumar *et al.* (2014) and Singh *et al.* (2016).

3. Quality indices and water vulnerability

The estimation of the water quality index (WQI) is obtained from multiplicative and weighted techniques, assigning specific weights in relation to their concentration (Brown *et al.* 1970). For the calculation of the quality index, the relative weights are determined as in equation (1); each parameter (w_i) is assigned a weight, represented by a number from 1 to 5, according to its importance. This index evaluates the quality of groundwater for human usage.

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i}. \tag{1}$$

After the calculation of the relative weights, a scale of valuation for the quality for each parameter is established as in equation (2) from the ratio of the concentration of each ion (C_i) and its standard value for drinking water (s_i):

$$q_i = \frac{C_i}{s_i} \times 100. \tag{2}$$

Finally, the calculation of the WQI is given by equation (3), and its classification is given by the categories in table 1.

Table 2. Rates of vulnerability, DRASTIC method.

Rate	Vulnerability
Very low	23–64
Low	65–105
Moderate	106–146
High	147–187
Very high	188–230

$$\mathrm{WQI} = \sum_{i=1}^{n} W_i q_i. \tag{3}$$

4. DRASTIC

This method consists of the evaluation of seven parameters that describe the geologic conditions of the aquifer environment with respect to their ability to prevent or facilitate the flow of pollutants into the aquifer. The parameters to be evaluated are depth (D), net recharge (R), aquifer (A), soil classification (S), topography (T), impact of the non-vadose zone (I) and hydraulic conductivity (C) (Aller *et al.* 1987). For each of the parameters a weight is assigned according to its impact on the vulnerability of the aquifers, as shown in equation (4). The rating of the vulnerability is given according to table 2.

$$\begin{aligned}\mathrm{IV} = {} & D_\mathrm{r}D_\mathrm{w} + R_\mathrm{r}R_\mathrm{w} + A_\mathrm{r}A_\mathrm{w} + S_\mathrm{r}S_\mathrm{w} \\ & + T_\mathrm{r}T_\mathrm{w} + I_\mathrm{r}I_\mathrm{w} + C_\mathrm{r}C_\mathrm{w}.\end{aligned} \tag{4}$$

5. Geostatistical spatial methods

5.1. *Frequentist spatial statistical method*

These statistical methods are based on the prediction of the autocorrelation degree of a phenomenon in a certain space (R^d) from observed samples. The spatial phenomenon is defined as a stochastic process z in a location s_i that takes values in a space, which can be n-dimensional, but in practice is most often in $n = 2$, $s_i = (x_i, y_i)$, and $n = 3$, $s_i = (x_i, y_i, z_i)$, dimensions.

5.2. *Stationarity*

When it comes to predicting a regionalized variable $z(s)$, we evidently assume that the process studied has some stability that can be detected, otherwise it will be impossible to predict (Puraivan 2017). Assuming that the spatial process $z(s)$ is stationary, in a weak sense, we define the following moments.

5.3. *Moments of a regionalized variable*

- Average of a regionalized variable. The average function exists and does not depend on location:

$$E[z(s_i)] = \mu. \tag{5}$$

- Variance of a regionalized variable. The variance function exists for all (s) belonging to the domain of the region under study:

$$V[z(s_i)] = \sigma^2. \tag{6}$$

- Covariance of a regionalized variable. The covariance exists and is a unique function of the vector $Z(s + h)$ of separation h:

$$C[z(s_i), z(s_j)] = C(h) = C_{ij}. \tag{7}$$

The C_{ij} function is called the covariogram or stationary covariance function (Cressie 1992).

This implies that the mathematical expectation is constant and the covariance function is invariant in space, which means that the process $z(s)$ is Gaussian, and for any location $(s_1, \ldots, s_n)$, $Z = [z(s_1), \ldots, z(s_n)]'$ has a normal multivariate distribution.

5.4. *Functions of spatial correlation*

The development of a geostatistical analysis begins with the determination of the spatial dependence of the measured data of a variable, known as structural analysis. The spatial dependence is studied with the semivariogram, covariogram and correlogram.

5.4.1. Variogram and semivariogram. The experimental semivariogram $\gamma(h)$ is a useful tool for making spatial estimations. The semivariogram allows us to detect the spatial correlation at different distances and directions. It is represented by a curve that measures the degree of continuity of the case study in order to define whether the spatial phenomenon fulfills the principle of seasonality; it is simplified as follows (Melo 2012):

$$2N(h)\gamma(h) = \sum_{(s_i,s_j)\in N(h)}^{N(h)} [z(s_i) - z(s_j)]^2, \tag{8}$$

where $\gamma(h)$ is known as the semivariogram and characterizes the properties of spatial dependence of the process, estimated by the method of moments, through the experimental semivariogram, which is calculated as in Wackernagel (2003). $z(s_i)$ is the spatial phenomenon measured in the space (s_i), $z(s_j)$ is another measure at a distance h from $z(s_i)$ and $N(h)$ is the number of pairs separated by h.

5.4.2. Covariogram and correlogram. The covariogram represents the joint sample spatial variability between pairs of observations at a distance h. Its mathematical representation is given by equation (7); μ represents the average value at all points of the region of study, and N is the number of pairs of points separated by a distance h.

The correlogram measures the correlation coefficient that exists between the measured variable $z[s_i]$ at site s_i and another sample value $z[s_j]$ measured at a distance h:

$$\rho(h) = \frac{C(h)}{C(0)}. \tag{9}$$

This spatial correlation reflects the fact that we are working with a regionalized variable and under the assumption of weak stationarity.

5.5. *Kriging*

5.5.1. Simple kriging. In this case of kriging, the average value of the spatial process μ is known and constant. Also, the variogram and the sill σ^2 are known. Therefore, the Gaussian spatial process is given by

$$z(s) = \mu(s) + \varepsilon(s), \tag{10}$$

where $E[\varepsilon(s)] = 0$ and $E[\varepsilon(s)^2] = \sigma^2$, with $\mu(s) = x'(s)\beta$. If there is no trend, then $x'(s) = 1$, $\beta = \beta_0$ and

$$Z = X\beta + \varepsilon, \tag{11}$$

where μ represents the average of the spatial process z and ϵ is the error. So, we must estimate z at a location s_0: (x_0, y_0), given by

$$\widehat{z}(s_0) = \mu + \widehat{\varepsilon}(s_0), \tag{12}$$

where

$$\widehat{\varepsilon}(s_0) = \sum_{i=1}^{n} \lambda_i \varepsilon(s_i). \tag{13}$$

In equation (12), we observe that the average value of the function is invariant according to each measure and therefore it is a fixed part in the process, while the errors are dependent on each measure and therefore are the random part of the process. Furthermore, for simple kriging, the weights λ_i have no type of restriction given that $E(\varepsilon) = 0$. On the other hand, the variance error of predictions is given by

$$\sigma^2_{\mathrm{SK}}(s_0) = \sigma^2 - \lambda' c. \tag{14}$$

5.5.2. Ordinary kriging. For this prediction method, it is assumed that the mean of the Gaussian process $\widehat{z}(s_0)$ is unknown and constant (Melo 2015); the predictor is defined in the following expression:

$$\widehat{z}(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i), \tag{15}$$

where λ_i are weights, which must add 1 to comply with the unbiasedness condition. The weights are chosen to minimize the estimation variance error $\sigma^2_{\mathrm{OK}}(s_0)$:

$$\sigma^2_{\mathrm{OK}}(s_0) = V[\widehat{z}(s_0) - z(s_0)] = \sigma^2 - (\lambda' c + \phi), \tag{16}$$

where σ^2 denotes the variance error of $z(s_0)$, and (ϕ, λ) corresponds to the Lagrangian function parameters, where ϕ is the Lagrange multiplier.

5.5.3. Cokriging. This is a method of spatial prediction derived from kriging, with the difference that it incorporates the use of covariables that provide additional information about the spatial phenomenon of the modeled object:

$$\gamma_{Z_k Z_l}(h) = \frac{1}{2N(h)} \sum_{(s_i, s_j) \in N(h)}^{N(h)} [z_k(s_i) - z_l(s_j)]^2. \tag{17}$$

For cokriging modeling, we have the cross-semivariogram equation (17), where the representation of semivariance is not only between points of a variable but also for one variable with respect to another. In particular, cokriging always gives a variance of estimation less than or equal to kriging (Emery 2013).

5.5.4. Simple cokriging. Analogous to simple kriging, the mean value μ that the process takes is known and constant. The approach consists of a linear combination of the available Z_i observations and linear combinations of the observations of the related variables (Ginzo Villamayor & Febrero Bande 2015). Starting with equation (12), the error component of a second variable is added:

$$\widehat{z}_k(s_0) = \mu_{z_k} + \widehat{\varepsilon}_{z_k}(s_0) + \widehat{\varepsilon}_{z_l}(s_0), \tag{18}$$

where

$$\widehat{\varepsilon}_{z_k}(s_0) = \sum_{i=1}^{n_{z_k}} \lambda_i \varepsilon_{z_k}(s_i), \tag{19}$$

$$\widehat{\varepsilon}_{z_l}(s_0) = \sum_{j=1}^{n_{z_l}} \theta_j \varepsilon_{z_l}(s_j). \tag{20}$$

Then, substituting equations (19) and (20) in (18), we have the following expression for the estimator $\widehat{Z}(s_0)$:

$$\widehat{z}_k(s_0) = \mu_{z_k} + \sum_{i=1}^{n_{z_k}} \lambda_i \varepsilon_{z_k}(s_i) + \sum_{j=1}^{n_{z_l}} \theta_j \varepsilon_{z_l}(s_j). \tag{21}$$

Finally, the variance of the estimator will be given by the following expression:

$$\sigma^2_{\mathrm{SCK}}(s_0) = \sigma^2 - (\lambda' c_{\varepsilon_{Z_k}} + \theta' c_{\varepsilon_{Z_l}}), \tag{22}$$

where $c_{\varepsilon_{Z_k}}$ is the covariance vector between each measure and the prediction points for the variable Z_k, and $c_{\varepsilon_{Z_l}}$ corresponds to the same covariance vector for the variable Z_l. Furthermore, for cokriging prediction methods, as many error or estimator components were added, as the case may be, as auxiliary variables are added to the modeling of the spatial phenomenon; moreover, a weight parameter corresponds to each component.

5.6. *Ordinary cokriging*

For estimation with ordinary cokriging we start from equation (15), adding the sum of the measurements of the auxiliary variable with their respective weights as seen in the following expression:

$$\widehat{z_k}(s_0) = \sum_{i=1}^{n_{Z_k}} \lambda_i z_k(s_i) + \sum_{j=1}^{n_{Z_l}} \theta_j z_l(s_j). \tag{23}$$

The variance analogous to equation (16) is given by the prediction variance error of $z_k(s_0)$ and the Lagrangian function parameters $(\lambda, \theta, \phi_1, \phi_2)$, where ϕ_1 and ϕ_2 are the Lagrange multipliers associated with the constraints $\lambda' \mathbf{1} = 1$ and $\theta' \mathbf{1} = 0$, respectively:

$$\sigma^2_{\mathrm{OCK}}(s_0) = \sigma^2 - (\lambda' c_{Z_k} + \theta' c_{Z_l} + \phi_1). \tag{24}$$

6. Bayesian spatial statistical method

Until recently, geostatistics was based on the theoretical basis for data models established by Cressie (1992), which allows the modeling of all the possibilities from a GF, which constitute an important component of current spatial hierarchical models (Banerjee *et al.* 2014), such as the case of the application of kriging that satisfies the conditions of spatial correlation functions, stationarity and isotropy. GFs are one of the few appropriate multivariate models with an explicit normalization factor and with good analytical properties, yet carry with them a problem of high computational cost when n is large due to the factorization of covariance matrices of dimension $n \times n$ (Monsalve 2013). This is because at the time of applying statistical inference, it is generally necessary to evaluate the probability function or the latent distribution of the Gaussian field, and therefore we need to make extensive calculations with matrices. To address the problem of large n, Lindgren *et al.* (2011) proposed applying an approximation, replacing a GF with a Gaussian Markov random field (GMRF).

6.1. *Link between GF and GMRF*

The proposed model involves a GF affected by a spatial process and represented as a discretely indexed GMRF, where its probability function depends on a neighborhood structure. This differs from the classical processes of kriging prediction, since what it does is perform a linear combination of the n random variables, leading to the problem of computational cost. This in turn allows them to make inferences in a much more conscious way (thanks to the Markov property), without the need to look for a valid family of covariance functions, including oscillatory, non-stationarity or on differential varieties of problems; moreover, working with GMRFs makes integration with the methodology of INLA possible in a natural way (Muñoz 2012).

The link between a GF and a GMRF is developed by Lindgren *et al.* (2011) using a stochastic partial differential equation (SPDE) as a Matérn covariance function of a GF process as an alternative to the traditional covariance, which guarantees weak stationarity for its smoothing parameter but simplifies the calculations by making use of the theory of the GMRF, defined as

$$k(s_i, s_j) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (k\|s_i - s_j\|)^{\nu} K_{\nu}(k\|s_i - s_j\|), \tag{25}$$

where K_{ν} is the modified Bessel function of the second type and order $\nu > 0$, κ is a parameter of the spatial dependency distance, $\Gamma(\nu)$ the Gamma function, $\|v - u\|$ the Euclidean distance and σ^2 the marginal variance. In order for the Bessel function to comply with the isotropy principle, a relationship is established between the Gaussian field and the Matérn covariance function as a solution of stochastic partial differential equations with the following expression:

$$(\kappa^{2-\Delta})^{\frac{\alpha}{2}}(\tau X(s)) = W(\mu), \quad \alpha = \nu + \frac{d}{2}, \quad \kappa > 0, \nu > 0, \tag{26}$$

where $(\kappa^{2-\Delta})^{\frac{\alpha}{2}}$ is a pseudo-differential operator, W is the process of Gaussian spatial innovation of white noise with unit variance, Δ is the Laplacian defined as $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ and τ refers to the marginal variance through the relationship

$$\tau^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + \frac{d}{2}) \times 4\pi^{\frac{d}{2}} \kappa^{2\nu} \sigma^2}. \tag{27}$$

The use of the SPDE approach allows the construction of a GMRF approximation to its solutions, which presents important computational advantages, by the property of the Markov field where the probability density function depends on the neighborhood structure. The mathematical notation representing the distribution of a GMRF, where $x \sim N(\mu, Q^{-1})$, with mean μ and symmetric positive definite precision matrix Q (inverse of the covariance matrix), is

summarized as (Monsalve 2013)

$$\pi(x) = (2\pi)^{-\frac{n}{2}}|Q|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'Q(x-\mu)\right). \tag{28}$$

The fact that the density function only depends on some components of x means that, at the time of applying Bayesian inference, the covariance matrices of order $n \times n$ are smaller than the global covariance matrices valid in the case of Gaussian fields. For example, matrix factorization, which normally requires $O(n^3)$ for a dense matrix, is $O(n)$ or $O(n^{3/2})$ and $O(n^2)$ for the sparse GMRF matrix (Blangiardo & Cameletti 2015). In addition, the introduction of random parameters in the coefficients of the SPDE allows us to consider it as part of a hierarchical model and fit them into a Bayesian inference scheme, which makes it feasible to implement the INLA methodology.

A GMRF is defined by a sparse precision matrix Q, which represents the conditional behavior of the field in relation to a neighborhood structure:

$$p(x_i \mid x_{i-1}) = p(x_i \mid x_{\vartheta_i}). \tag{29}$$

This is equivalent to saying that, given the neighborhood structure ϑ_i, x_i and x_{ϑ_i} are independent. This means that a strong relation exists between the conditional independence and the precision matrix Q. In fact, for any pair (i, j) with $i \neq j$, it will be, according to (Monsalve 2013),

$$x_{i \perp x_j} \mid x_{\{i,j\}} \Leftrightarrow Q_{i,j} = 0. \tag{30}$$

This property allows us to make fast factorizations of Q. The calculation of the matrix Q is made from a triangulation of the domain of the zone of interest, which seeks to approach the behavior of a Markov field that better represents the Matérn field, making it possible to make inference on the GMRF.

The SPDE approach defines the Matérn field in a single representation over a neighborhood with a triangulation structure. The representation of the base function of the Matérn field $X(s)$ is given by

$$X(s) = \sum_{i=1}^{n} \omega_i(s)\omega_i, \tag{31}$$

where n is the total number of vertices, $\omega_i(s)$ are the base functions and ω_i are weights with Gaussian distribution. The functions $\omega_i(s)$ are selected so that there are linear pieces in each triangle, i.e. $\omega_i(s)$ is 1 at the vertex and 0 at the other vertices. Finally, the weights ω_i represent the height of the triangles that are part of the net. The values of the triangles are estimated with linear predictions (Monsalve 2013).

6.2. *The INLA environment*

The algorithm is based on the definition of the probability distribution of the observed variable $Y = (y_1, y_2, \ldots, y_n)^t$, but specifically, a distribution given by φ_i is assigned to each y_i which will be described by a predictor η_i; however, this relationship will not be direct but is posed through a bond function in such a way that $f(\varphi_i) = \eta_i$ and the predictor will have the form

$$\eta_i = \beta_0 + \sum_{q=1}^{Q} \beta_q x_{iq} + \sum_{l=1}^{L} f_l(z_{il}), \tag{32}$$

where β_0 represents the intercept of the model, and the coefficients β represent the linear relationship between the predictor and a set of explanatory variables x_i, $i = 1, 2, \ldots, n$. Finally, the component $f_l(z_l)$ is one or more functions in terms of a subset of covariables z_i, $i = 1, 2, \ldots, n$, which, depending on their definition, can capture nonlinear relationships, temporal trends or spatial effects; then, all observable and non-observable effects will be captured in a vector $\Theta = \{\beta_0, \beta, f\}$. This is the reason why the INLA method can be applied in various disciplines (Blangiardo & Cameletti 2015).

In addition, the estimation requires the definition of a set of hyperparameters $\Psi = (\psi_1, \psi_2, \ldots, \psi_k)$ associated with the a priori distribution under which the algorithm is performed. So, if independence is assumed in the n observations, the probability distribution is given by

$$P(Y \mid \Theta, \Psi) = \prod_{i=1}^{n} P(y_i \mid \theta_i, \Psi). \tag{33}$$

From this perspective, the objective of the Bayesian estimation with the INLA method is to find the subsequent marginal distribution of the elements of Θ,

$$\int P(\theta_i \mid y) = P(\theta_i \mid \Psi, Y)P(\Psi \mid Y)d(\Psi), \tag{34}$$

for which the algorithm develops two tasks, the estimation of $P(\Psi|Y)$, where the hyperparameters of the a priori distribution are used to obtain $P(\theta_i|\Psi, Y)$, and finally to calculate $P(\theta_i|y)$.

The approximation to the subsequent marginal distribution consists in estimating the later ones of interest through Laplace's method of approximation of the hyperparameters, taking into account (Blangiardo & Cameletti 2015)

$$p(\psi \mid y) = \frac{p(\theta, \psi \mid y)}{p(\theta \mid \psi, y)} \propto \frac{p(\psi)p(\theta \mid \psi)p(y \mid \theta)}{p(\theta \mid \psi, y)} \approx \left.\frac{p(\psi)p(\theta \mid \psi)p(y \mid \theta)}{\tilde{p}(\theta \mid \psi, y)}\right|_{\theta=\theta^*(\psi)}. \tag{35}$$

The importance of INLA lies in the definition of $p(\theta|\psi, y)$ and $\theta = \theta^*(\psi)$, which refers to the point where the function corresponds to the maximum or most likely value of x for θ, given that it is obtained from Laplace's transformation by the expansion of the Taylor series evaluated at $x = x_0$ taking into account that (Blangiardo & Cameletti 2015)

$$\log f(x) \approx \log f(x_0) + (x - x_0)\frac{\partial \log f(x)}{\partial x}\bigg|_{x=x_0} + \frac{(x - x_0)^2}{2}\frac{\partial^2(\log f(x))}{\partial x^2}. \quad (36)$$

Linearizing and setting x_0 as $x^* = \text{argmax}_x \log f(x)$, then $\frac{\partial \log f(x)}{\partial x}\big|_{x=x_*} = 0$ and this approach becomes

$$\log f(x) \approx \log f(x^*) + \frac{(x - x_0)^2}{2}\frac{\partial^2 \log f(x))}{\partial x^2}\bigg|_{x=x^*}. \quad (37)$$

The interest integral is given by

$$\int f(x)\,dx \approx \int \exp\left\{ \log f(x^*) + \frac{(x - x^*)}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{x=x^*} \right\} dx, \quad (38)$$

$$\int f(x)dx \approx e^{\log f(x^*)} \int \exp\left\{ \frac{(x-x^*)}{2}\frac{\partial^2 \log f(x)}{\partial x^2}\bigg|_{x=x^*} \right\} dx. \quad (39)$$

This integrand is equivalent to the density of a normal distribution, establishing the variance

$$\sigma^{2*} = -\left(\frac{\partial^2 \log f(x)}{\partial x^2}\right)^{-1},$$

obtaining

$$\int f(x)\,dx \approx e^{\log f(x^*)} \int \exp\left\{ -\frac{(x - x^*)^2}{2\sigma^{2*}} \right\} dx. \quad (40)$$

This expression defines a distribution with average value x^*, and the variance σ^{2^*} is the integral evaluated in the interval (α, β) given by

$$\int_\alpha^\beta f(x)\,dx \simeq f(x^*)\sqrt{2\pi\sigma^{2*}}(\phi(\beta) - \phi(\alpha)), \quad (41)$$

where $\phi()$ is the normal density.

This allows different approaches than the Gaussian, with faster calculation, but errors can occur in the location of the later mean and/or errors due to the lack of symmetry (Lindgren & Rue 2015). Laplace's complete approach is more accurate, but requires more computational time, while the simplistic version is quick to calculate and generates the most accurate approximations (Rue *et al.* 2009). The key to this methodology is to approximate the subsequent marginals of x_i by nested approximations of

$$\tilde{p}(\theta_i \mid y) \approx \sum_{k=1}^{K} \tilde{p}(\theta_i \mid \psi_k, y)\tilde{p}(\psi_k \mid y)\Delta_k \quad (42)$$

from the discretization of the problem and the search for a square to produce a set of points relevant to the hyperparameters together with the weights Δ_k, to approximate this distribution using numerical integration (finite sum) on θ, which consists in the estimation of a curve from $\tilde{p}(\theta_i \mid y)$ from a set of points placed in some space that can be m-dimensional (Monsalve 2013). The objective of this modeling is not to calculate the joint posterior distribution, taking into account equation (28), but to approximate the marginal or subsequent distribution of the observable and non-observable elements. The distribution of variables $Y = \{y_1, \ldots, y_n\}$ is simplified with $\pi(y|\theta, x)$ by approximating to $\pi(\theta_i|y)$, $\pi(x|y)$, $\pi(x_i|y)$.

We assume that the result Y_i follows a Gaussian distribution as follows:

$$Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, 2, 3, \ldots, n, \quad (43)$$

$$\mu_i = \beta_0 + Z(s_i). \quad (44)$$

With the methodology used by INLA, following the notation of the equation (32), the linear predictor η_i is specified in a function of μ_i through the linkage function $g(\mu_i) = \eta_i$, equivalent to

$$E(y_i \mid \beta_0, \ldots, \beta_Q, X_{i1}, \ldots, X_{iQ}) = \beta_0 + \sum_{q=1}^{Q} \beta_q X_{iq}. \quad (45)$$

The prior distribution for β and σ^2 is calculated through a process, inferring the parameters of regression $\beta_0, \ldots, \beta_Q$ and variance σ^2.

7. Calculation of quality indices

7.1. WQI

Initially, the ionic balance (IB) is calculated, where it is assumed that the solution must be electrically neutral and therefore it must be verified that the sum of the cations is equal to the anions. Maximum values are accepted in the percentage of ionic balance error of 10%. From 275 samples, about 64 are removed for having an IB greater than 10%, resulting in 211 representative samples. The parameters under consideration in the calculation of WQI were pH, electrical conductivity, turbidity, alkalinity, sulfate, phosphate, nitrates, calcium, magnesium, manganese and sodium (Murugesan *et al.* 2010; Singh *et al.* 2016).

Taking into account the concentration of each parameter for the calculation of the index, following the methodology developed by Brown *et al.* (1970), the technique is a weighting of the parameters through the quality standards that were taken into consideration: those proposed by the World Health Organization (WHO 2011), and the Colombian standard, taking into account Colombian

Table 3. Quality parameters for calculation of the water quality index.

Parameter	Minimum	First quartile	Median	Mean	Third quartile	Maximum
pH	4.6	6.0	6.6	6.517	7.1	8.1
Electrical conductivity	8.00	50.65	110.00	248.96	303.00	**2691**
Turbidity	0.3	1.9	5.0	43.19	20.50	**2302**
Alkalinity	0.3	9.25	29.00	71.63	84.00	**945.00**
Phosphate	0.10	0.10	0.10	0.15	0.10	2.80
Nitrates	0.00	0.1	0.7	3.587	2.7	**105**
Sulfate	0.00	2.20	6.50	37.49	25.95	**559**
Calcium	0.20	4.10	10.60	26.07	35.5	**184**
Magnesium	0.10	1.00	2.20	5.90	5.80	**92.60**
Manganese	0.00	0.05	0.05	0.202	0.076	4.00
Sodium	0.1	1.75	4.6	19.5	14.30	**407**

Table 4. Standards and relative weights selected from the physicochemical parameters of water.

Chemical parameters	Standards (WHO 2011)	Standards (Resolución 2115)	Selected standards	Weight ω_i	Relative weight $W_i = \frac{\omega_i}{\sum_{i=1}^{n} \omega_i}$
pH	6.5–8.5	6.5–9.0	6.5–8.5	1	0.02631
Electrical conductivity (μ)	500	1000	500	5	0.13157
Turbidity (NTU)	2	2	2	4	0.10526
Alkalinity (mg l^{-1})	—	200	200	4	0.10526
Phosphate (mg l^{-1})	—	0.5	0.5	1	0.02631
Nitrates (mg l^{-1})	45	10	10	5	0.13157
Sulfate (mg l^{-1})	250	250	250	4	0.10526
Calcium (mg l^{-1})	75	60	60	4	0.10526
Magnesium (mg l^{-1})	50	36	36	3	0.07894
Manganese (mg l^{-1})	0.1	0.1	0.1	3	0.07894
Sodium (mg l^{-1})	200	—	200	4	0.10526
				$\Sigma\omega_i = 38$	$\Sigma W_i = 1$

resolution number 2115 of June 22 2007, where the guidelines are analyzed to be taken into account in the control and monitoring system for the quality of water for human consumption.

From the standards, it was decided to assign greater weight to electrical conductivity and nitrates, because, as set out in table 3, these are the parameters that are more distant from the permitted value. High electrical conductivity values mean there are salts and ionic amounts of all dissolved substances in the groundwater (Ca, Na and Mg); if these salts are consumed at high levels there can be negative effects on human health, causing conditions such as increased arterial pressure, abdominal ailments, renal diseases, or hepatic cirrhosis (Murugesan *et al.* 2010). The concentration of salts is related to their wide diffusion in igneous rocks, sedimentation and metamorphism. The concentration of nitrates may be related to intensive fertilizer practices for agricultural activities and the use of different fertilizers, accompanied by soil-specific conditions such as permeability and contamination, which allow the circulation of these ions towards aquifers, contaminating them; this is reinforced by the dumping of human and animal trash (Pacheco *et al.* 2004).

The parameters used for the calculation of the WQI index are reflected in table 4, in the "Selected standards" column.

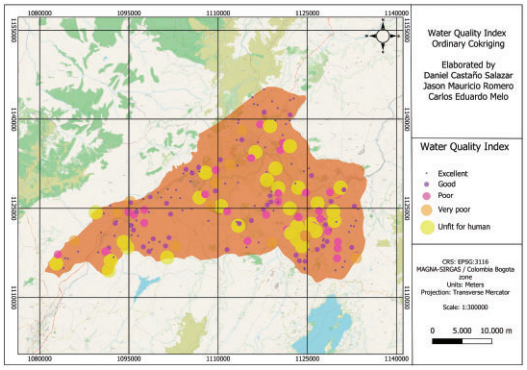


Figure 2. Water quality index map. See https://rpubs.com/jasonromero/IndexWQI.

In figure 2 and table 5 it can be identified that there are values for which the quality index of the groundwater for the area is suitable for human consumption, with a minimum value of 6.50, being a positive indicator for the present investigation; however, there is a maximum that exceeds the levels of quality of groundwater suitable for consumption, a value of around 12 143.53; this may be located in the area where the highest values for electrical conductivity and nitrates were identified, directly related to the industrialized cities of the area.

Table 5. Statistical summary for the calculation of the water quality index.

Parameter	Minimum	First quartile	Median	Mean	Third quartile	Maximum
Quality index (WQI)	6.50	30.70	65.58	266.79	156.32	**12143.54**

From the 211 samples, about 35.54% correspond to excellent water for human consumption, 18.00% refer to good water, and the other ranges appear with less frequency, below 18%. Thus, more than half of the points sampled showed optimal conditions for human consumption.

8. Exploratory spatial data analysis

8.1. WQI

The WQI variable has a positive asymmetric histogram, where the mean is 266.79 and is greater than the median (65.58), which indicates the possible non-normality of the variable. This being so, the solution is to transform the data.

Applying the Box–Cox transformation, as shown in figure 3b, the histogram indicates that now the data follow, in an approximate way, a normal distribution. This transformation will be taken into account when forecasting.

With the aim of determining if the variation of the value of our variable with space is equal in all directions, and detecting isotropic behavior, figure 4 shows the semivariogram in the directions 0°, 45° and 135°. They have the same behavior, so there is no variability from the separation of the location of the sampled points.

Two conclusions can be drawn from this figure. First, there is no clear anisotropic behavior in the WQI on the sampled scale, as the ranges and nuggets are approximately similar. In spite of this, in the directions NE–SW (45°) and E–W (90°) the sill reaches lower values than in the other two directions, its behavior also being different at greater distances (Gallardo

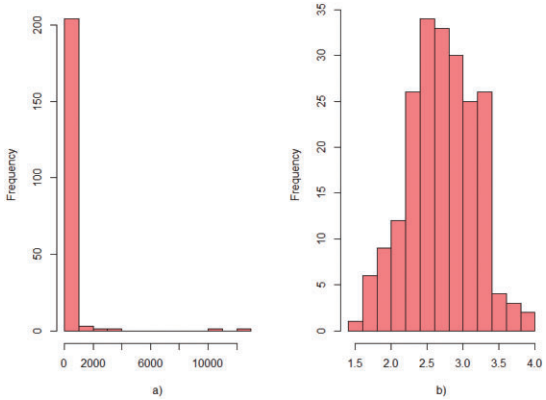


Figure 3. (a) No normal distribution of WQI. (b) Box–Cox transformation distribution of WQI. (R Core Team 2021).

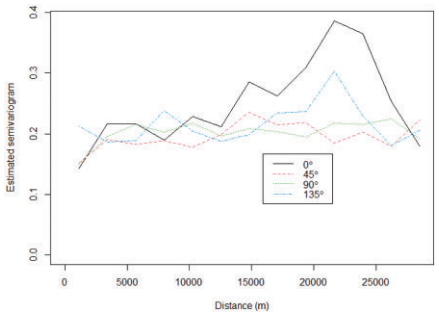


Figure 4. Classical experimental semivariograms of the water quality index in four directions of space. (R Core Team 2021).

2006). The regionalized variable is intricately stationary and isotropic to the stochastic process.

8.2. Kriging

From the experimental semivariograms, it can be observed that their behavior is similar and comparable; however, the estimates from the median and the mean cut to 10% present dispersion between the lags. Based on the above analysis, the robust experimental estimators and the classical have less dispersion, so we choose the robust, with better results in the presence of atypical data.

8.3. Adjustment of the semivariogram

The adjustment of the semivariogram was carried out with a Matérn model considering three methods of adjustment: maximum likelihood (ML), restricted maximum likelihood (RML) and weighted least squares (WLS). The adjustment parameters are shown in table 6.

WLS estimation is selected because with this setting, there is less distance from the theoretical curve of the Matérn semivariogram to the points of the theoretical semivariogram; furthermore, is the only one that allows the nugget to approach that of the experimental semivariogram (see figure 5).

8.4. Ordinary cokriging

According to Emery (2013), there are two reasons that justify the use of the cokriging method, including a model with auxiliary variables, instead of applying a kriging estimation separately from each of them. The first is that with the information provided by the auxiliary variables, cokriging always gives an estimation variance less than or equal to kriging. The second

Table 6. Semivariogram parameters adjusted for ML, RML and WLS.

Estimation	Nugget (c_0)	Sill (c_1)	Range (a)	κ
ML	0.10	0.28	7800	1.2
RML	0.1	0.25	9200	0.9
WLS	0.24	0.28	8800	0.33

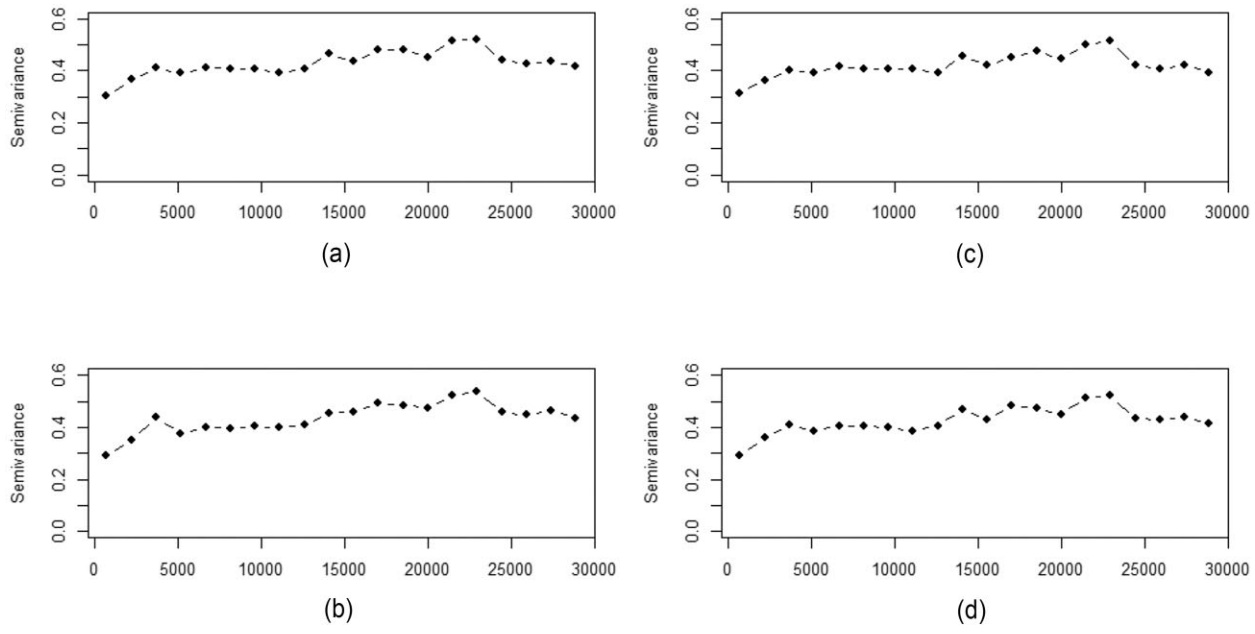


Figure 5. Estimation of experimental semivariograms for WQI. (a) Robust. (b) Classic. (c) Medium. (d) Cut average. (R Core Team 2021).

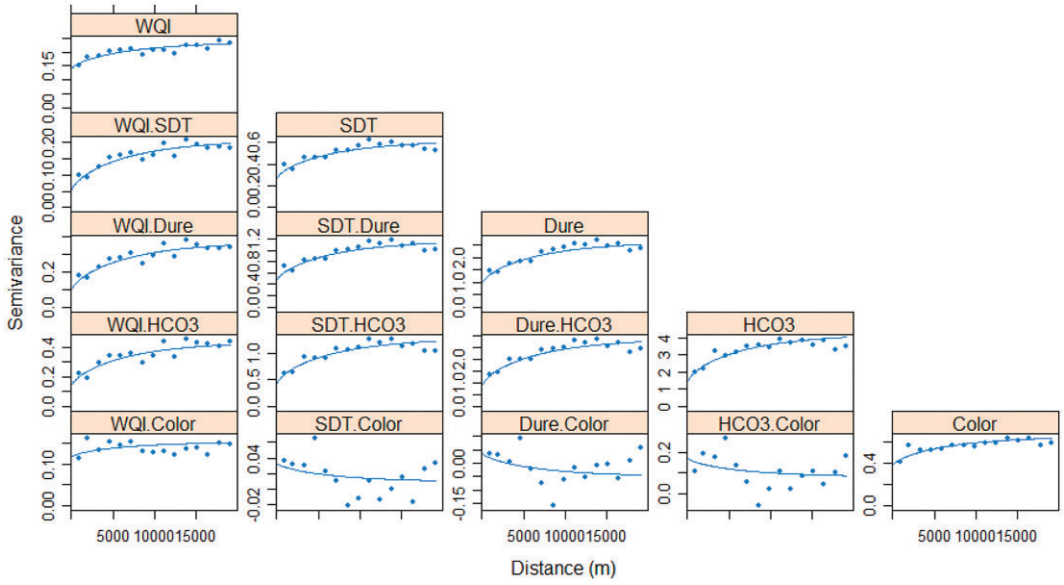


Figure 6. Adjustment of the linear model of coregionalization to the multivariable WQI semivariogram. (R Core Team 2021).

is that the consistency of the estimation results is improved, since the linear relationships between the variables are taken into account. For example, when variables represent proportions, the sum of their estimates is equal to 100%, a situation that is not necessarily fulfilled when performing kriging of each variable separately. That is why physicochemical properties not taken into account for the calculation of the quality index in the study area are chosen to be added in the model. That is, make a prediction of the WQI based on your information and on that of some auxiliary variables that spatially correlate with it and can model the behavior of the index. For this, a linear model of coregionalization is adjusted to a multivariable sample variogram based on the previously adjusted semivariogram.

The moment estimator of the cross-semivariance function is given by the cross semivariogram (Bogaert *et al.* 1995). The linear model of coregionalization assumes that all single and crossed semivariograms can be expressed as a linear combination of the same theoretical models.

The adjustment of the experimental cross-variogram describes the covariance between the auxiliary variables, where the variance represented is between points of one variable with respect to another; in figure 6, the correct adjustment of the cross-variograms can be observed.

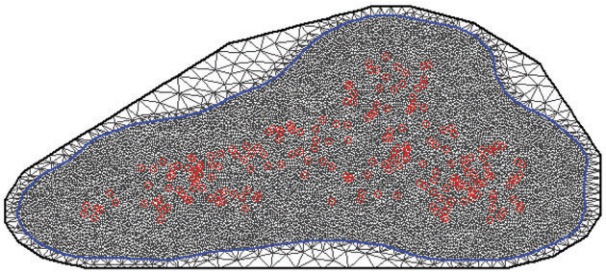


Figure 7. Triangulation selected, indicating the points sampled. (R Core Team 2021).

8.5. *SPDE approach*

To capture the quality index trend across the entire domain of the study area, and for computational convenience, a refined and restricted Delaunay triangulation spatial mesh is constructed (Vilela *et al.* 2021). The more triangles we have, the more precise will be our approach, but we have to pay for it in terms of computational time. The desired mesh will have small triangles where the data is dense, larger where the data is more scattered, and even larger where there are no observations.

Given the triangulation, the representation of the base function for the field $X(s)$ is given by equation (26) and shown in figure 7. For the formulation of the model, taking into account equations (43) and (44), it is assumed that the WQI follows a Gaussian distribution, expressed as the sum of an intersection β and a spatially structured random effect that follows a Gaussian process with the Matérn function of covariance (Moraga 2019):

$$\text{formula} = \text{WQI} \sim 0 + \beta_0 + \beta_x + f(\text{spatial.field}, \text{model} = \text{spde}). \quad (46)$$

Table 7. Estimated best model hyperparameters by R-INLA SPDE (WQI).

Parameters	Mean	SD	2.5%	50%	97.5%
σ_e^2	0.092	0.023	0.054	0.089	0.147
σ^2	0.049	0.034	0.011	0.040	0.141
r	3029.715	1901.317	764.2877	2569.764	7981.901

In the formula, the response variable is included on the left-hand side and the fixed β_x effects that refer to the $Z(s_i)$ of equation (44) and random effects are on the right-hand side; the intersection is eliminated (adding 0), and we add it as a covariable term (adding β_0; Moraga 2019). As fixed effects, the Secchi disk transparency (SDT), hardness, HCO_3 and color are selected as auxiliary variables. To evaluate the best models, the density of the parameters β_x and their significance is determined; if the density function is within the range 0.025–0.975 it does not contain zero as a vertical line.

From the hyperparameters obtained, listed in table 7, it is evident that the variation of the measurement error and the spatial term of the parameter of the Matérn covariance function (σ_e^2 and σ^2) are both close to zero, indicating good results, and with a spatial range (r) of 3029.715 m for the rank of correlation derived empirically from $\rho = \sqrt{8\nu}/\kappa$, $\kappa = 0.001$ and $\nu = 1.484$.

The WQI is determined by $e^{\beta_0} = 0.95$, which means that anywhere in the area the approximate average index is 34.137 with a probability of 95%, being a good indicator of water suitable for human consumption in the area. Looking at figure 8 and table 8, the SDT and color variables have a positive effect on the value of WQI. For color, the higher the value of this variable, the higher the WQI, being one of the organoleptic parameters indicating the quality of water for

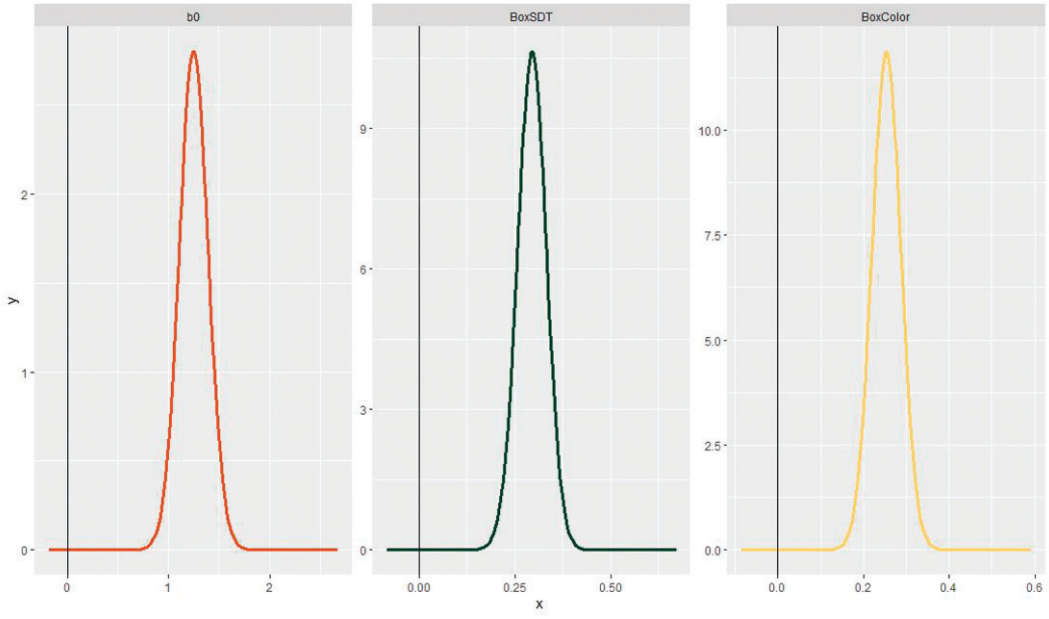


Figure 8. Density of the x-parameters and their significance according to the linear model. (R Core Team 2021).

Table 8. Actual values of the parameter β for the linear model.

Function	β	SDT	Color
e^{β}	34.137	3.827	3.632

human consumption and related to the dissolved substances and suspended parts contained in the body of water; with an increase of 3.632 mg l^{-1} in color, the index is increased by one unit. The SDT is a good measure of water quality, expressing the content of particles of salts, metals and minerals in the groundwater; with an increase of 3.827 mg l^{-1}, the WQI increases by one unit.

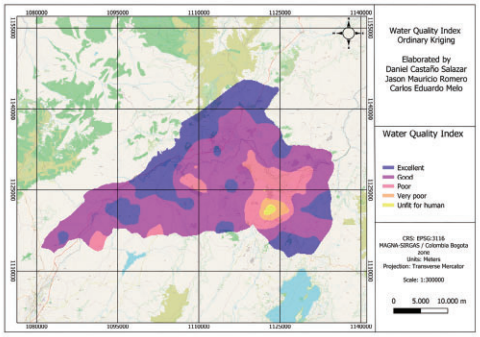
8.6. WQI predictions

The point estimates of concentration are very similar between kriging, cokriging and the INLA-adjusted model, providing similar estimations of the underlying spatial process.

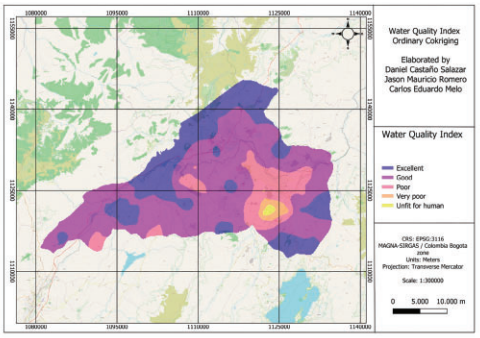
On the basis of table 9, it can be seen that in the central area of Boyacá, most bodies of underground water present ideal conditions for human consumption, with the highest percentage of zones with a WQI for good water for human consumption, followed by water in excellent condition, with a total shareholding together of 86.715% area, indicating that about 90% of water quality index values are between 25.94 and 100. Finally, by comparing the results obtained in figure 9a, b, c and d, it can be determined that the areas with groundwater with the best quality parameters correspond to the NE and SE zones, which are natural forest areas. The zones with high indices refer to the urban area of the municipalities of Sogamoso, Firavitoba, Paipa and Cómbita, which may be the result of discharges of domestic waste water resulting in superficial or direct contamination of groundwater from the pollution of some important water resources in the area such as the basin of the Sogamoso and Chicamocha rivers, mainly associated with the lack of wastewater treatment plants (Corpoboyacá 2015).

Table 9. Categories of water quality index for the central area of Boyacá from the SPDE forecast map.

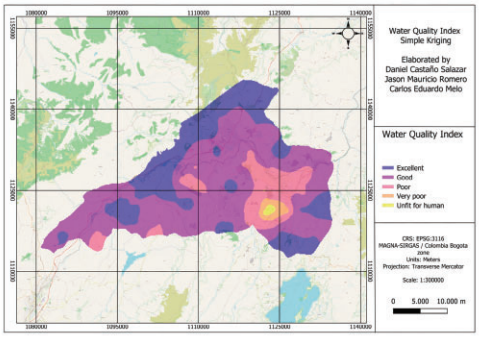
WQI	Categories	Area (km^2)
<50	Excellent	29.651
50–100	Good	**57.064**
100–200	Poor	11.944
200–300	Very poor	1.338
>300	Unfit for human consumption	0



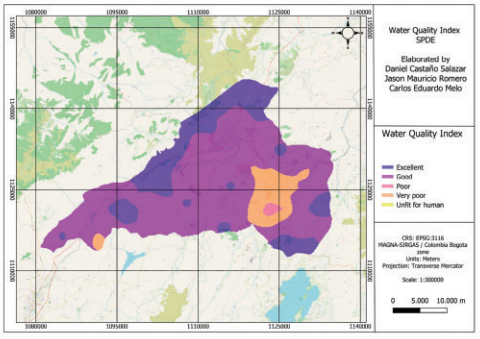
(a) Ordinary Kriging. See:
https://rpubs.com/jasonromero/KOWQI



(c) Ordinary CoKriging. See:
https://rpubs.com/jasonromero/COKOWQI



(b) Simple Kriging. See:
https://rpubs.com/jasonromero/KSWQI



(d) SPDE. See:
https://rpubs.com/jasonromero/SPDEWQI

Figure 9. WQI predictions using kriging and INLA-SPDE.

8.7. Cross validation

With the aim of achieving an approach to determine which of the prediction methods offers greater precision and evaluating the SPDE approach in terms of its robustness, ease of computation and automation versus the kriging approach, the mean absolute error (MAE) is calculated and the root mean square error (RMSE) is used to explain the ability of the entire distribution of the subsequent reconstruction.

For the calculation of cross validation, and to ensure that it takes into account the same number of neighbors, knowing that kriging takes into account a global covariance model and SPDE a discretely indexed covariance model, the cross validation was performed using the spatial range of table 7 as the radial distance, removing the observations of the training set. The methods of prediction have similarities in terms of their results: the lowest MAE and RMSE are from cokriging, taking into account the auxiliary variables SDT, hardness, HCO_3 and color, followed by the covariable SPDE, which takes into account only the effect of the auxiliary variables SDT and color.

To reinforce and verify the precision of the methods, it is not only the root mean square error and the mean absolute error that are taken into account, because we are trying to compare frequentist and Bayesian techniques. The statistic from the Diebold–Mariano test (Drachal 2018) determines whether the difference is significant (with predictive values) with 95% credibility. The null hypothesis of the statistic is that there are no relevant differences between the two models; the alternate hypothesis is that the method of prediction 2 is more accurate than method 1, rejecting the null hypothesis and accepting the alternate hypothesis when the *P*-value is greater than 5% alpha. Using this, we determine that the SPDE approach without and with auxiliary variables has better prognostic precision versus kriging and cokriging (see table 10 and 11).

Table 10. Cross validation results, WQI.

Method	MAE	RMSE
Ordinary kriging	0.356	0.445
Simple kriging	0.353	0.441
Ordinary cokriging	**0.296**	**0.369**
SPDE without covariables	0.374	0.464
SPDE with covariables	**0.292**	**0.360**

Table 11. Diebold–Mariano Test (Drachal 2018): comparison of the forecast precision for the water quality index.

Method 1	Method 2	*P*-value
Ordinary kriging	Simple kriging	0.028
Simple kriging	**Ordinary cokriging**	1
Ordinary kriging	**SPDE without covariables**	0.912
Simple kriging	**SPDE without covariables**	0.967
Ordinary cokriging	**SPDE without covariables**	1
Ordinary cokriging	**SPDE with covariables**	0.243

9. DRASTIC results

Figure 10 shows the result of the vulnerability index for the study area.

Based on the methodology proposed for the assessment of the intrinsic vulnerability of aquatic species to pollution (Quintero 2010), five DRASTIC index intervals are distinguished through the territory. Table 12 shows the area distribution and the related percentage over the total area of each rank of intrinsic vulnerability.

10. Discussion

The use of Bayesian spatial models using the INLA approach provides a robust and flexible approach especially useful for analyzing the behavior of hydrogeological systems, considering the effect of the structure of spatial dependence and providing the possibility of making inferences, providing statistical evidence to establish the existence of direct or inverse interactions between the quality indices and explanatory variables.

For this study, use of the INLA approach showed that quantitative parameters influenced the prediction. It would be of great interest to add qualitative parameters such as in Moghaddam *et al.* (2022); the combination of prediction results with qualitative parameters can contribute significantly to the evaluation of the water quality index, in addition to showing how the spatial statistics methodology kriging can be further improved by adding Bayesian networks. This leads to more dynamic results, with a better approach to the reality of the territory that allows studying the incidence and extension of hydrogeological parameters, informing management decisions with greater precision, even as multilevel spatial random effect structures that represent all the explicit spatial processes that can be integrated into the distribution pattern of quality indices (Vilela *et al.* 2021). In addition to this, Delaunay's triangulation allows us to collect more information in areas with more observations, which contributes to more accurate predictions (Vilela *et al.* 2021). However, it is important to highlight the results of studies such as Huang *et al.* (2017) and Poggio *et al.* (2016) where the use of triangulated meshes in INLA-SPDE meant the computation time beccame quite long when dealing with non-Gaussian likelihood families.

11. Conclusions

Being a continuous spatial phenomenon, geostatistics provides the right tools to address the problem, modeling the phenomenon as a realization of a spatial stochastic process,

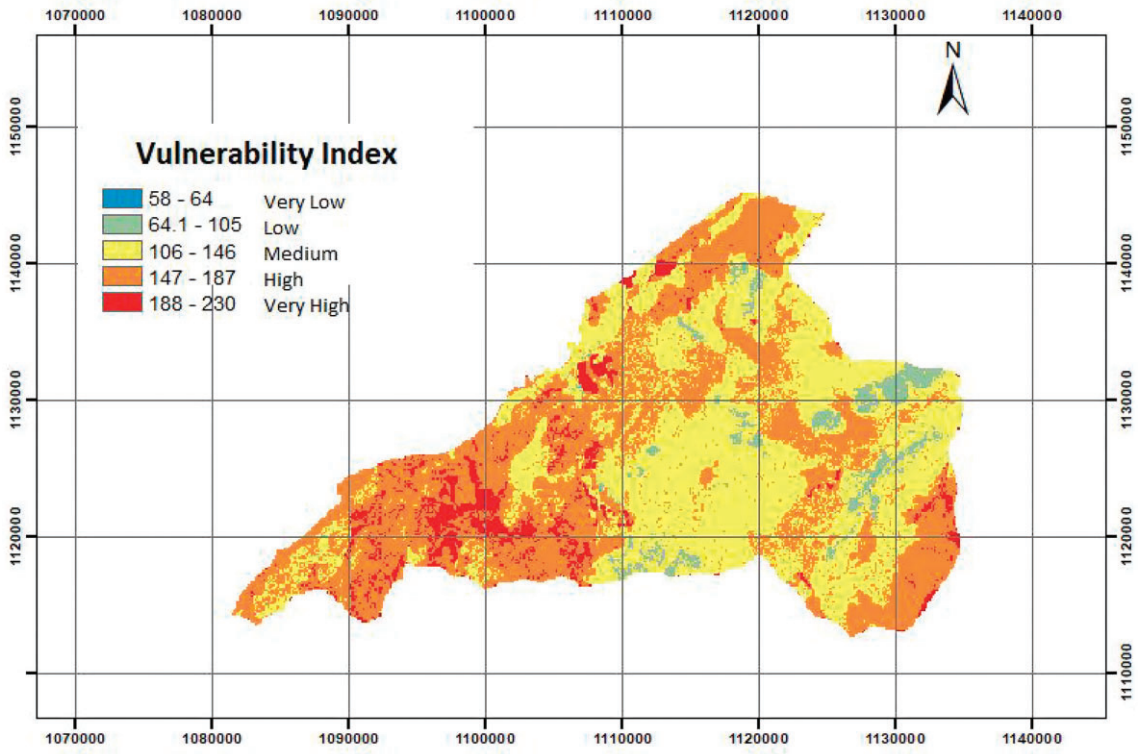


Figure 10. Intrinsic vulnerability of groundwater in the central zone of Boyacá, using the DRASTIC methodology.

Table 12. Area distribution of the ranks of intrinsic vulnerability.

Rank	Area (km^2)	Percentage
Very low	0.034	0.0039
Low	34.332	3.9961
Moderate	386.070	44.96
High	385.787	44.93
Very high	52.466	6.11

combining the SPDE approach with the INLA methodology, providing an important framework for calculating Bayesian inference in complex models with spatial structure and, at the same time, facilitating the handling of large sets, being suitable for any type of spatial data including stationary continuous phenomena and anisotropics with better results than kriging methods for this study in the regional geological data of Boyaća, Colombia, and demonstrating that the hierarchical spatial predictions made with the Bayesian model have a low computational cost. Nevertheless, the frequentist models with covariables had similar values of precision to INLA, as shown by the MAE and RMSE metrics. The resulting predictions reveal areas where water is fit for human consumption corresponding to areas with high and very high vulnerability, where water bodies are shallow, exposing them to pollutants that cause serious damage of the hydrological system for the study area in the medium-long term.

The DRASTIC method is of great importance for establishing critical areas with bodies of water in need of maintenance, for tracking potential sources of contamination, and to as a guideline to define development objectives in the central area of Boyacá. The results indicate that the largest proportion of land is concentrated in areas with a moderate and high degree of vulnerability, mainly due to the fact that they are not very deep bodies of groundwater and, by the lithology of the area, affected by erosion and leaching of feldspar and magnesium calcite found in the lithologic units of the zone, together with anthropogenic activities related to decreases in production levels in farming areas due to over-exploitation, pollution from the use of agrochemicals and solid waste dumping.

Acknowledgements

This work was partially funded and supported by Core Spatial Data Research (Faculty of Engineering, COL0013969, Universidad Distrital Francisco José de Caldas) and Applied Statistics in Experimental Research, Industry and Biotechnology (COL0004469, Universidad Nacional de Colombia).

Conflict of interest statement. None declared.

Group contributors. NIDE (spatial data research nucleus) working group of the Universidad Distrital Francisco José de Caldas.

References

Aller, L., Bennett, T., Lehr, J., Petty, R. & Hackett, G., 1987. *DRASTIC: A standardized system for evaluating groundwater pollution potential using hydrogeologic setting*. Technical report 600/2-87/035, Environmental Protection Agency, Washington, D.C.

Banerjee, S., Carlin, B.G. & Gelfand, A.E. 2014. *Hierarchical Modeling and Analysis for Spatial Data*, Routledge.

Blangiardo, M. & Cameletti, M., 2015. *Spatial and Spatio-temporal Bayesian Models with R-INLA*, John Wiley & Sons.

Bogaert, P., Mahau, P. & Beckers, F., 1995. *The spatial interpolation of agro-climatic data. Cokriging software and source code, User's manual*. Working paper, Food and Agriculture Organization.

Brown, R.M., McClelland, N., Deininger, R.A. & Tozer, R., 1970. A water quality index: Do we dare? *Water Sewage Works*, **117**, 339–343.

Corpoboyacá, 2015. *Diagnosis of water in the middle and upper basin of the Chicamocha river*. Technical report, Corpoboyacá.

Cressie, E., 1992. *Statistics for Spatial Data*, John Wiley & Sons.

Drachal, K., 2018. multMDM: Multivariate version of the Diebold–Mariano test. Available at: https://cran.r-project.org/package=multDM.

Emery, X., 2013. *Geostatistics*, Faculty of Physical and Mathematical Sciences, University of Chile.

Gallardo, A., 2006. Geostatistics, *Ecosystems*, 15.

Ginzo Villamayor, M. & Febrero Bande, M., 2015. *Análisis geostadístico de datos funcionales*. Poster presentation, Universidade de Santiago de Compostela.

Gómez-Rubio, V., 2020. *Bayesian Inference with INLA*, Chapman & Hall / CRC.

Huang, J., Malone, B., Minasny, B., McBratney, A. & Triantafilis, J., 2017. Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping, *Science of The Total Environment*, **609**, 621–632.

Krishna kumar, S., Lokeshkumaran, A., Magesh, N.S., Godson, Prince S. & Chandrasekar, N., 2015. Hydro-geochemistry and application of water quality index (WQI) for groundwater quality assessment, Anna Nagar, part of Chennai City, Tamil Nadu, India, *Applied Water Science*, **5**, 335–343.

Lezama-Ochoa, N., Pennino, M., Hall, M., Lopez, J. & Murua, H., 2020. Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the spinetail devil ray (*Mobula mobular*), *Scientific Reports*, **10**, 18822.

Lindgren, F., 2018. Spatially varying mesh quality. Available at: https://www.maths.ed.ac.uk/flindgre/posts/2018-07-22-spatially-varying-mesh-quality/.

Lindgren, F. & Rue, H., 2015. Bayesian spatial modelling with R-INLA, *Journal of Statistical Software*, **63**, 1–25.

Melo, C., 2012. *Análisis geoestadístico espacio tiempo basado en distancias y splines con aplicaciones*, PhD thesis, Universitat de Barcelona.

Melo, C., 2015. *Estadística espacial teoría y aplicaciones*, Universidad Distrital Francisco José de Caldas, Faculty of Engineering.

Moghaddam, H.K., Rajaei, A., Rahimzadeh kivi, Z. & Moghaddam, H.K., 2022. Prediction of qualitative parameters concentration in the groundwater resources using the Bayesian approach, *Groundwater for Sustainable Development*, **17**, 100758.

Monsalve, G., 2013. *Spatial Bayesian hierarchical models in agricultural epidemiology*, PhD thesis, Universitat Politècnica de València.

Montaño, B., 2019. *El crecimiento de la población y la escasez hídrica*. Congreso nacional del agua 2019: Innovación y sostenibilidad, pp. 509–519.

Moraga, P., 2019. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*, Chapman & Hall / CRC.

Muñoz, F., 2012. *Geostatistics in heterogeneous regions with distance based on cost*, PhD thesis, Universitat de Valencia.

Murugesan, V., Krishnaraj, S., Vijayaragavan, K., Ganthi, R., Sabarathinam, C., Anandhan, P., Manivannan, R. & Vasudevan, S., 2010. Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, Tamil Nadu, India, *Environmental Monitoring and Assessment*, **171**, 595–609.

Pacheco, Á., Cabrera, A. & Pérez, R., 2004. Diagnosis of the groundwater quality in municipal systems of supply in the state of Yucatan, Mexico, *Engineering*, **8**, 165–179.

Poggio, L., Gimona, A., Spezia, L. & Brewer, M., 2016. Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA, *Geoderma*, **277**, 69–82.

Puraivan, E., 2017. *Geostatistical analysis of air pollution data in Santiago, Chile, using SPDE with INLA estimation method*, Master's thesis, Universidad de Valparaíso.

Quintero, M., 2010. Propuesta metodológica para la evaluación de la vulnerabilidad intrínseca de los acuíferos a la contaminación. Technical report, Ministerio de Ambiente, Vivienda y Desarrollo Territorial, Colombia.

R Core Team, 2021. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rue, H. & Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Rue, H., Martino, S. & Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society Series B*, **71**, 319–392.

Singh, S., Hussain, A. & Dubey, S., 2016. Water quality index development for groundwater quality assessment of Greater Noida sub-basin, Uttar Pradesh, India, *Cogent Engineering*, **3**, 1177155.

UNESCO, 2014. *The United Nations world water development report 2014: Water and energy*, United Nations Educational, Scientific and Cultural Organization.

Vilela, R., *et al.*, 2021. Use of an INLA latent Gaussian modeling approach to assess bird population changes due to the development of offshore wind farms, *Frontiers in Marine Science*, **8**, 701332.

Wackernagel, H., 2003. *Multivariate Geostatistics: An Introduction with Applications*, 3rd edn, Springer.

WHO, 2011. *Guidelines for Drinking-Water Quality*, 4th edn, World Health Organization.