Introducing Bayesian priors to semi-variogram parameter estimation using fewer observations

Y. Zhang, Leo Pichon, J.A. Taylor, B. Oger, B. Tisseyre

Introducing Bayesian priors to semi-variogram parameter estimation using fewer observations

Yulin Zhang[1], Léo Pichon[1], James A. Taylor[1], Baptiste Oger[1], Bruno Tisseyre[1]
[1] ITAP, Univ Montpellier, INRAE, Institut Agro, 2 Pl. Pierre Viala, 34000, Montpellier, France
yulin.zhang@supagro.fr

Abstract

Correct estimation of variogram parameters relies on having a sufficiently large dataset. However, operational agri-datasets are often not large enough for variogram fitting. This article presents a new approach to estimating semi-variogram parameters from a small dataset by using a Bayesian approach. The three variogram parameters of the Spherical-Plus-Nugget model were fitted to the semi-variances of a vineyard water stress indicator. Two sources of prior information (i.e. using ancillary data, and using some simplistic assumptions), and six reduced datasets were tested. The results showed that using prior information introduced less variability in estimation results than with the classical approach. The priors extracted from the Sentinel-2 data significantly improved the estimation of the nugget effect, which allowed better preservation of the spatial pattern of kriging predictions.

Keywords: geostatistics, semi-variogram modeling, ancillary data, Bayesian modeling,

Introduction

Precision agriculture requires high spatial (and temporal) resolution datasets to make decisions at the within-field level. The kriging technique is frequently used to improve the spatial data coverage rate (Rajabi et al., 2018), but accurate kriging predictions require at least 100 data points for good semi-variogram computation (Oliver and Webster, 2014). However, many operational agri-datasets are smaller than this leading to unreliable variogram fitting. The role of ancillary data in reducing the impact of low data availability on variogram estimation is important, as the ancillary data may be spatially correlated with the targeted agronomic property and hence share similar spatial variability. Thus, the more abundant ancillary data can be used as a surrogate for geostatistical analysis. For example, it is possible to improve sampling schemes by considering ancillary data (Kerry and Oliver, 2004; 2008) directly used the nugget:sill ratio calculated from the ancillary data in variogram modeling of a variable of interest. However, in their approach, the relationship between the two types of data was uncertain and the information provided by the ancillary data was unable to be updated when actual observations were made. A possible solution to this limitation is the use of a Bayesian framework that provides an adequate way to combine prior information and observations, and to account for uncertainty (McElreath, 2016). Therefore, the objective of this study was to carry out semi-variogram modeling using few data points under a probabilistic framework, with prior information extracted from relevant ancillary data. The proposed approach considered variogram model parameters as random variables, which are characterized by certain probabilistic distributions and are updated by using the actual semi-variances derived from available data from small datasets.

Materials and methods

General approach

This work proposed a novel method to estimate three variogram parameters, nugget (c0), partial sill (c1), and range (r), using fewer data points. It described how the semi-variances were standardized using the sill variance, and how the nugget:sill ratio and range were estimated using a Bayesian approach with the standardized semi-variances. The best estimate of the standardized nugget:sill ratio was then back-transformed to obtain the n c0, and c1. A case study was presented with specific prior information to estimate a viticulture variable with varying prior information sources and varying observational dataset sizes. Lastly, the resulting estimations were evaluated both quantitatively and qualitatively.

Proposed methodology

Estimation and standardization of the variogram model

The spherical-plus-nugget (SPN) variogram model was used in this study. The model was fitted from a set of observed semi-variances, ***s***, using the method presented by Oliver and Webster (2014). A function was developed to estimate the SPN parameters stored in vector ***v*** [c0, c1, r] while accounting for a given ***s***. A black-box optimizer *optim()* was used to optimize ***v*** by minimizing the difference between predicted and observed semi-variances. After obtaining ***v***, all values in ***s*** were divided by the actual sill value (c0 + c1). The resulting set of semi-variances, ***s'***, were used for parameter estimation.

Bayesian update using grid approximation

To estimate the nugget:sill ratio and range, the prior probability density functions (p.d.f.) of the two parameters were defined. These latter were probabilistic descriptions of possible parameter values before actual observations were obtained. Each prior p.d.f. was represented approximately by a step-by-step calculation of the density of possible parameter values between two numerical bounds. These values were separated by a constant distance, forming a regular discretization grid containing N nodes. N^2 combinations of possible values of nugget:sill ratio and range were generated and stored in vector ***u***. The prior joint log-probability of observing each possible combination in ***u*** was computed using density functions provided by R (R Core Team, 2022) and denoted ***w****. The joint posterior p.d.f. of the two parameters, describing probabilities of possible values after considering actual observations, were obtained by the following steps.

1) The log-likelihoods of observing a specific ***s'*** given each combination in ***u*** was calculated and stored in vector ***I***.
2) According to the Bayes' law, each prior joint log-probability in ***w**** was summed by the corresponding log-likehood in ***I*** so that the former was reweighted by actual observations in ***s'***, which generated an approximation of the joint posterior p.d.f., denoted ***w***.

The first 10 combinations in ***u*** with the largest values in ***w*** were selected. The mean nugget:sill ratio and range were computed using these 10 values. The ***v*** parameters were back-transformed using the estimated nugget:sill ratio, range values, and actual sill variance obtained previously.

Case study

The study field was a 1.3 ha non-irrigated vineyard (*Vitis vinifera* cv Grenache) located near Corbières in southern France (Fig. 1b) (WGS 84: 43.1692°N, 2.5629°E). It was planted in 1989 with a density of 4000 vines ha^{-1}. Training and management practices were typical for this region. Observations of shoot growth characterization were collected at 97 within-field sites (Fig. 1a) using the iG-Apex index as proposed by Pichon et al. (2021). The iG-Apex index varies from 1 (full shoot growth) to 0 (total cessation of shoot growth), and is a surrogate for vine water restriction. Observations were collected weekly in 2020 from week 25 to week 34, generating 10 temporal points, that resulted in 97 time-series. The date at which the iG-Apex reached the value 0.5, noted dG0.5, was chosen as the agronomic variable of interest in this paper. The dG0.5 was spatially structured at the within-field level (Fig. 1c).

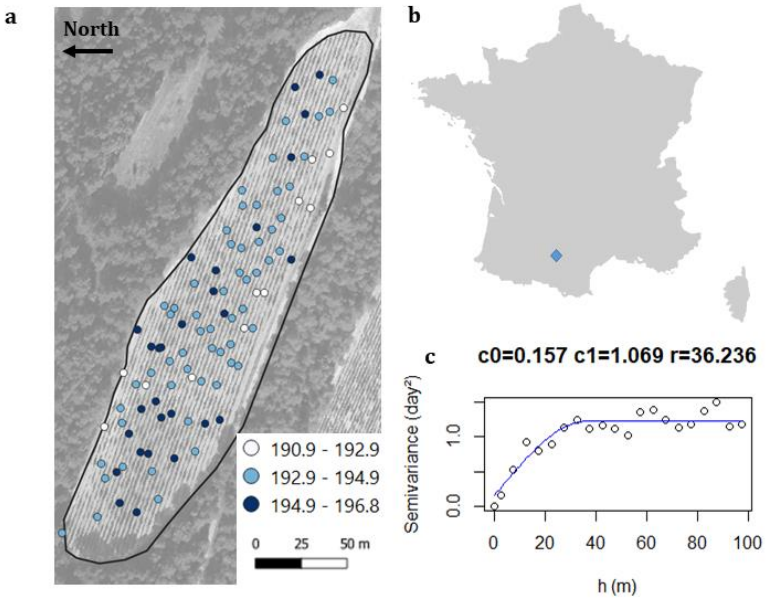

Figure 1. (a) Positions of the 97 data points coloured according to dG0.5, (b) the position of the vineyard in France and, (c) the SPN variogram fitted from the observed dG0.5 with nugget (c0) = 0.157, partial sill (c1) = 1.069, range (r)= 36 m.

Prior p.d.f. for nugget:sill ratio and range

Two sources of prior information were used and compared in this study. The first was based on simplistic assumptions of the vineyard. The variogram range was constrained to avoid too small (>5 m) or too large (<75 m) distances while it was also assumed that at least half of the observed variance was spatially correlated (nugget:sill < 0.5). Two uniform distributions (Unif) were used to describe these priors (Table 1).

The second source of prior information came from ancillary data. Similar to the iG-Apex index, NDVI is an indicator of vine vigour, and was considered as a relevant ancillary data source. There were 16 variograms fitted using NDVI imaged by the Sentinel-2 satellite at 16 dates during the growing season (the vineyard was covered by 81 pixels at each date). The gstat function *fit.variogram()* in R was used to estimate SPN variogram parameters of NDVI values at each date. A Triangular and a Normal distribution were used respectively to model the nugget:sill and the range (Table 1).

Design of the numerical experiment

Six data reduction schemes were tested that included different percentages of the original data: S1 (80%), S2 (70%), S3 (60%), S4 (50%), S5 (40%), and S6 (30%). In count number, S1 contained 79 data points, S2 - 68, S3 - 59, S4 - 48, S5 - 39, and S6 - 30 data points. Each reduction scheme was randomly generated 100 times from the full dataset, resulting in 600 reduced datasets. The SPN variogram parameters for each reduced dataset were estimated using no prior information, and the two sources of prior information outlined above.

Table 1. Distribution laws of priors generated from simplistic assumptions and from ancillary data analysis, as well as their discretization schemes. Unif stands for uniform distribution and N for the number of regular discretized values considered between the bounds.

	Source of prior information		Discretization	
	Simplistic prior	Ancillary prior	Bounds	N
Nugget:sill ratio	Unif(0, 0.5)	Triangle(0, 0.01, 0.5)	0 – 0.5	100
Range (m)	Unif(5, 75)	Normal(60, 15)	5 – 95	100

Evaluation of estimation performance

The Root Mean Squared Error (RMSE) criterion was used to assess the estimation of variogram parameters. For each data reduction scheme and each prior information source, a RMSE was calculated using the 100 estimations of each parameter and the reference that was obtained by fitting the variogram model using the full dataset.

Using the estimated parameters, maps were generated for all combinations of data reduction and prior information sources using ordinary kriging (Cressie, 1990) supported by the R function *gstat::krige()*. A reference map was made from the full dataset. A single interpolation grid was prepared in advance using R and QGIS, and was used for all kriging computations. Visual evaluation was carried out by comparing the kriged maps obtained using different prior information with the reference map. Contour lines were added to facilitate visual evaluation using the function *ggplot2::geom_contour()* (function setting: a fixed binwidth of 0.5 day). The variance of all dG0.5 values predicted by kriging was computed for each kriged map.

Results

As the size of the datasets diminished, the three distributions of possible estimated values exhibited an increasing dispersive trend indicating an increasing uncertainty associated with estimations (Fig. 2). The dispersion level was lowest for the ancillary prior approach (Fig. 2g-i) and highest when using no priors (Fig. 2a-c). The distributions for the no prior and simplistic prior approaches were visually very similar, indicating little advantage to the simplistic prior approach. For both approaches, the distributions of possible estimated values seemed to be globally centered on the reference values. However, the estimations of range derived from the simplistic prior were less dispersed than the no prior approach. When using ancillary data to generate the priors, the estimations of nugget effect out-performed those provided by the two other approaches across all data reduction schemes (Fig. 2g), although the nugget effect tended to be underestimated. For the estimations of range using ancillary priors, there was a clear bias of up to 15 m observed, and the

majority of range estimations fell between 36 m (the reference value) and 60 m (the mode of the prior Normal distribution). Although, the range distributions derived from the ancillary priors were much less dispersed than the two other approaches (Fig. 2i).
The above visual observations were confirmed by the RMSE (Table 2). As expected, errors showed an ascending trend as more data points were discarded, especially for the partial sill and range estimation. However, the RMSE for c0 and range estimated using the ancillary prior were less sensitive to different data reduction schemes. In general, the ancillary prior approach obtained the smallest RMSE, except for the range estimations that were affected by the bias identified in Figure 2i. For the estimations of partial sill with strong reduction schemes (S5, S6), the simplistic prior approach showed a marginal gain in RMSE (Table 2).
Representative examples of kriged maps were selected to visualise the effect of variogram model parameter estimation, using 48 (Fig. 3b-d) and 30 (Fig. 3e-g) data points. In general, fewer contour lines were generated when interpolating with fewer data points. The use of prior information increased the global resemblance to the reference map.
In the absence of prior information, the kriged outputs were very smooth for both cases, showing small variances of predicted values and few identified contour lines (Fig. 3b and 3e). The use of a simplistic prior allowed a better preservation of the spatial variability with the variances in the map values increasing from 0.332 to 0.375, and from 0.059 to 0.129 for the 48 and 30 observations respectively, and more contours of predicted values were present (Fig. 3c and 3f).

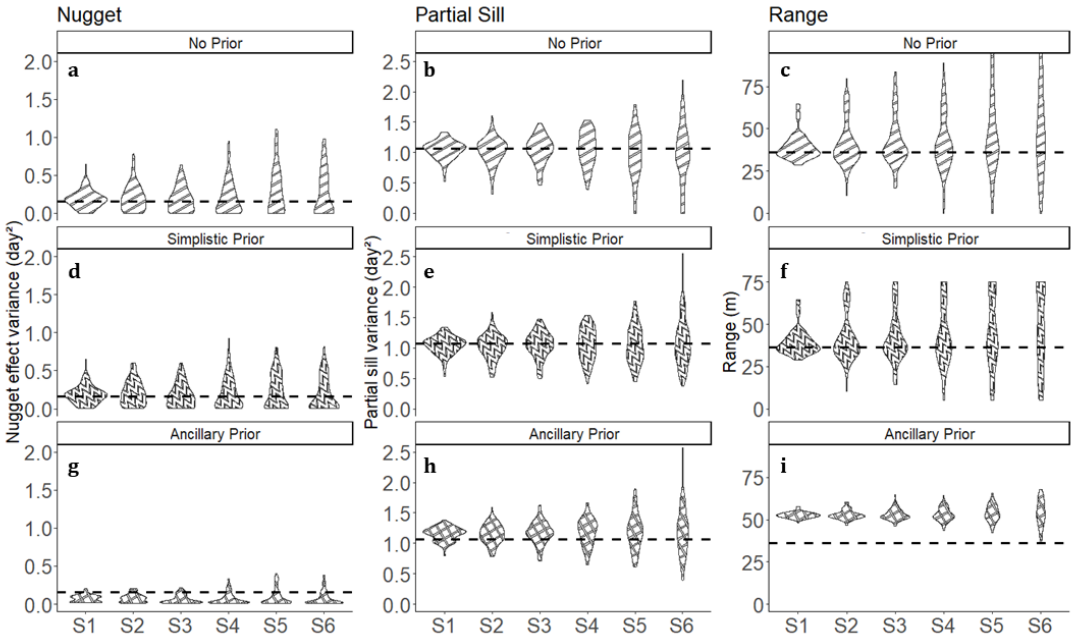

Figure 2. Distribution of estimated SPN variogram parameters using 1-no prior (a, b, c), 2-simplistic prior (d, e, f), and 3-ancillary prior (g, h, i) at the 6 data reduction schemes (S1-6) compared to the reference value represented by the dotted line.

Lastly, the ancillary prior had even larger variances of predicted values, and the resulting spatial structures were the most similar to the original reference map, especially when more data are used (Fig. 3d and 3g).

Table 2. RMSE of estimated variogram parameters (nugget, partial sill, and range) obtained under six data reduction schemes and three prior information treatments, compared to the reference value derived from the full dataset.

		Nugget	Partial Sill	Range
No Prior	S1	0.120	0.149	7.893
	S2	0.180	0.224	13.745
	S3	0.165	0.220	15.644*
	S4	0.219	0.270	17.845
	S5	0.332	0.404	24.695
	S6	0.307	0.477	27.459
Simplistic Prior	S1	0.119	0.148	7.891*
	S2	0.166	0.207	14.005
	S3	0.157	0.213	15.620*
	S4	0.211	0.258	18.874
	S5	0.235	0.298*	23.035
	S6	0.214	0.379*	25.329
Ancillary Prior	S1	0.094*	0.137*	16.518
	S2	0.094*	0.178*	16.804
	S3	0.107*	0.203*	17.056
	S4	0.113*	0.246*	17.286*
	S5	0.122*	0.305	18.306*
	S6	0.120*	0.395	19.022*

* represents the best value among the three prior information treatments.

Discussion

As expected, this study showed that stronger data reduction schemes introduced higher uncertainties in parameter estimation, as the certainty of estimation is linked to the availability of data (given a constancy in the data quality). However, it was possible to estimate semi-variogram parameters with fewer data points using prior information. The use of ancillary data can play an important role in accurately estimating the nugget:sill ratio, and to help to better preserve the spatial pattern of kriging predictions.

The slight improvement of nugget effect estimation brought by the simplistic prior was due to the upper boundary of nugget:sill ratio distribution being set at 0.5. Likewise, setting the upper boundary of the range at 75 m improved the RMSE in S5 and S6, showing that the uniform distribution was useful to rule out estimations with extreme values. However, the uniform distribution was relatively weakly informative (Hansen et al., 2016), estimations derived from the simplistic prior approach were mainly influenced by observations instead of the prior information, which explained the similar estimation performance compared to no prior approach. Triangular and Normal prior distributions permitted a reduction in the estimation uncertainty of the nugget and range, because both p.d.f. privileged parameter values centered around the statistical mode of the distribution (McElreath, 2016). They were informative priors because a reasonable level of confidence was attributed to them. Consequently, the resulting estimations were compromises between observations and prior information.

The bias in range estimation derived from the ancillary prior approach can be explained by two reasons. Firstly, the global spatial structure of the ancillary data was not (and is

highly unlikely to ever be) identical to that of the variable of interest, especially for the range. Secondly, the reduced datasets were not capable of reproducing the reference variogram parameters, maybe due to the completely randomly generated datasets and noise (stochastic error) in the measurement.

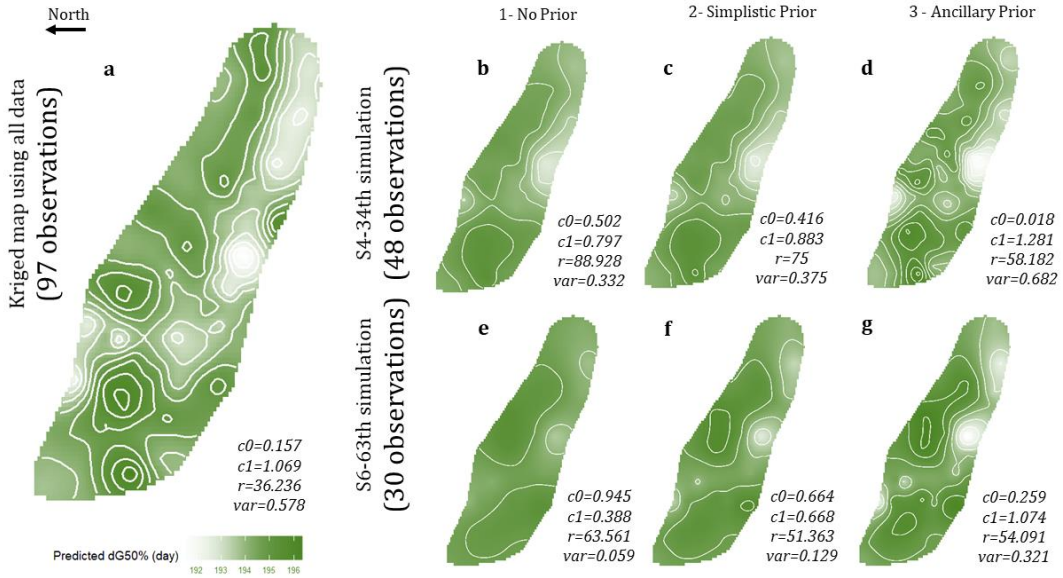

Figure 3. Kriged maps. With the full dataset (a). With a reduced dataset that contains 48 observations, and parameters estimated from no prior (b), simplistic prior (c), and ancillary prior (d). With a reduced dataset that contains 30 observations and parameters estimated from no prior (e), simplistic prior (f), and ancillary prior (g). Variances of all predicted dG0.5 are showed below each map.

The variability of kriging outputs was sensitive to the estimated nugget effect. However, a better preserved spatial pattern does not assure better predictions, as the estimated nugget and range values can both be biased.

For future improvements, NDVI observations collected during a shorter period, when the two agronomic variables are more correlated with each other, may improve the performance of the prior p.d.f.. Other vegetation indices, which are more sensitive to vine water status, will also be tested instead of NDVI (e.g. the Leaf Water Content Index (Ahamed et al., 2011)). Additionally, it would be interesting to see how a different sampling approach, for example a stratified sampling pattern instead of a random pattern, may affect this method (Kerry and Oliver, 2008). Lastly, the three variogram parameters were considered as independent of each other and analysed separately. It may be of interest to study the joint probability distribution in order to explore their interactions.

Conclusion

The Bayesian approach can combine prior information on semi-variogram model parameters and observed semi-variances. The performance of variogram modeling using fewer data points can be improved by introducing prior information extracted from appropriate ancillary data. Informative priors allow to reduce the uncertainty of variogram

parameter estimation. It is essential to identify ancillary data with a similar nugget:sill ratio compared to the property of interest, because accurately estimating this ratio can preserve the spatial variability for the kriging analysis.

Acknowledgements

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004. The ApeX-Vigne project was part of the DATI project, supported by the French National Research Agency under the Horizon 2020 PRIMA Program (ANR-21-PRIM-0001).

References

Ahamed, T., Tian, L., Zhang, Y. and Ting, K.C. (2011). A review of remote sensing methods for biomass feedstock production. Biomass and Bioenergy 35(7), 2455-2469. https://doi.org/10.1016/j.biombioe.2011.02.028

Cressie, N. (1990). The origins of kriging. Mathematical Geology *22*(3), 239-252. https://doi.org/10.1007/BF00889887

Hansen, T.M., Cordua, K.S., Zunino, A. and Mosegaard, K. (2016). Probabilistic Integration of Geo-Information. In Integrated Imaging of the Earth (pp. 93-116). American Geophysical Union (AGU). https://doi.org/10.1002/9781118929063.ch6

Kerry, R. and Oliver, M.A. (2004). Average variograms to guide soil sampling. International Journal of Applied Earth Observation and Geoinformation 5(4), 307-325. https://doi.org/10.1016/j.jag.2004.07.005

Kerry, R. and Oliver, M.A. (2008). Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Precision Agriculture 9(1), 33-56. https://doi.org/10.1007/s11119-008-9058-0

McElreath, R. (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Oliver, M.A. and Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 113, 56–69. https://doi.org/10.1016/j.catena.2013.09.006

Pichon, L., Bopp, O. and Tisseyre, B. (2021). 20. Characterising within-field variability of vine water status with simple visual observations of shoot growth. In: Precision agriculture 21 (pp. 179-186). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-916-9_20

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rajabi, M.M., Ataie-Ashtiani, B. and Simmons, C.T. (2018). Model-data interaction in groundwater studies: Review of methods, applications and future directions. Journal of Hydrology 567, 457-477. https://doi.org/10.1016/j.jhydrol.2018.09.053