Censored data in spatial statistics : Plug-In Method and Bayesian Method

M. Wieskotten
CEA, LMA Université d'Avignon

Supervisor(s): Dr. M. Crozet (CEA), Dr. B. Iooss (EDF), Prof. C. Lacaux (Université d'Avignon) and Dr. N. Pérot (CEA)

Ph.D. expected duration: Oct. 2019 - Sep. 2022

Adress: CETAMA, CEA, DES, ISEC, DMRC, Univ Montpelier, Marcoule, France
CEA, DES, IRESNE, DER, Cadarache, Saint-Paul-lez-Durance, France
L.M.A., Université d'Avignon, Campus Jean-Henri Fabre 301, rue Baruch de Spinoza, 21239, Avignon

Email: martin.wieskotten@gmail.com

Abstract: The spatial radioactive characterization (be it through contamination or irradiation) is one of the main issue in decommissioning and dismantling projects of nuclear facilities. Spatial statistics offer solutions for predicting the location of contamination. The use of spatial correlation in association with prediction allows for reliable contamination mapping, but only under specific conditions.

One of these conditions is that observation's values and positions have to be known. The modelling needs a quantified value for the observation to be taken into account for variogram evaluation or kriging.

Censored data do not correspond to quantified values, but only to an interval in which the true value lies (the measured value being unavailable). In environmental science, these censored data are often left-censored, meaning that they lie between $-\infty$ and a detection limit, which in practice means that the true value lies between 0 and the detection limit. When encountering censored data, the common method in spatial statistics is to simply replace them by an arbitrary value, be it 0 or the detection limit itself. As seen in [2], these methods introduce bias in the inference by altering the information contained in the censored data. The resulting contamination's map is either over-estimating the contaminated surface or under-estimating it.

Our goal was initially to find methods or algorithms that allowed the integration of censored data in geostatistical modelling without adding arbitrary information to the censored data and thus avoiding bias. We propose two main answers to this problem : a plug-in estimate methods called CensSpatial [3] and a Bayesian method called Data augmentation [1].

[3] uses a maximum-likelihood based method called CensSpatial to work with censored data without replacing them arbitrarily. To do so it estimates the censored data by sampling it conditionally to non-censored data. We test the method which is implemented in R with the CensSpatial package and show its advantages over replacement methods (cf. Figure 1). On one data set (provided by the CEA), we test the mean prediction error with varying proportion of censored data between the CensSpatial method and two replacement methods : one replacing the censored data by 0, the other replacing them by the detection limit. The CensSpatial method offers a great alternative to the replacement methods, but still has a major default. It uses plug-in parameters : even though the modelling parameters are estimated with the variogram, the uncertainties around the parameters are not taken into account.

Bayesian inferences were integrated during the late 90s to spatial statistics, and allowed to have better predictions with small data sets by allowing the introduction of an "expert knowledge"

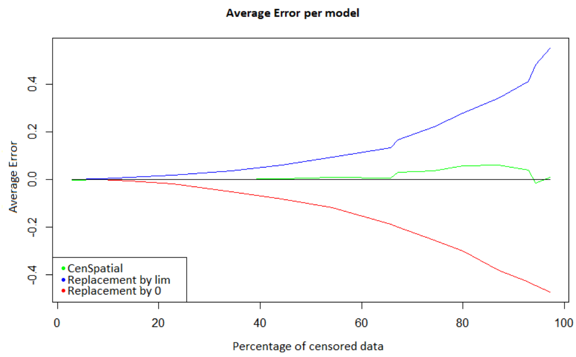

Figure 1: Results of the comparison

through the choice of a prior distribution. But the Bayesian inference also allows for the introduction of uncertainties in parameters estimation in Bayesian kriging where prediction and estimation are calculated simultaneously. These methods are based on a Markov Chain Monte Carlo algorithm to estimate the posterior distribution.

The Bayesian inference can take into account censored data by simply considering them as unknown parameters. [1] developed a method called data augmentation to do so and work with the censored data.

The comparison between these two approaches was the first step of our research. Despite many advantages, the Bayesian framework still has its limitation. Its calculations are computer-intensive, the kriging prediction depends on the prior and is also often more conservative than regular kriging (with plug-in estimations).

Our research aims at integrating more complex uncertainties modelling with a better robustness in the small-size sample cases while maintaining a balance with calculation time and estimation variance. Bayesian inference offers new possible development towards that goal.

References

[1] Brooke L. Fridley and Philip M. Dixon. Data augmentation for a bayesian spatial model involving censored observations. *Environmetrics*, 18(2):107–123, 2006.

[2] Dennis R. Helsel. *Statistics for Censored Environmental Data Using Minitab and R*. Wiley, 2012.

[3] José A. Ordoñez, Dipankar Bandyopadhyay, Victor H. Lachos, and Celso R.B. Cabral. Geostatistical estimation and prediction for censored responses. *Spatial Statistics*, 23:109–123, 2017.

Short biography – I am a first year PhD student with an engineering degree in geosciences. My research is on the relationship between metrology and geostatistical modelling. Currently I am working on censored data and their integration in geostatistical modelling. The main application is in dismantling and sanitizing projects of old nuclear facilities.