doi:10.20944/preprints202209.0082.v1

Article

On the Impacts of Hard Data Patterns on Bayesian Maximum Entropy Performance: Simulation-based Analysis

Ehnon Gongnet E. [1, 2*] , Agbangba Codjo E. [1] , Affossogbe Sédjro T[1], Vihotogbe R.[3] and Glèlè Kakaï R.[1]

[1] Laboratoire de Biomathématiques et d'Estimations Forestières ; agbangbacodjoemile@gmail.com
[2*] Institut Tchadien de Recherche Agronomique pour le Développement ; ehnon.gongnet@g-mail.com
[3] Ecole de Foresterie et Ingénierie du Bois ; rlvihotogbe@gmail.com

Abstract: Bayesian Maximum Entropy (BME) is increasingly used in predicting and mapping spatio-temporal data. However, studies that have fully evaluated its robustness empirically are rare. Therefore, this research examined empirically the effect of skewness, sample size and spatial dependency level using simulated data. We considered symmetric data, data positively skewed by 1, 3, 6 and 9, data with weak, moderate, and strong spatial dependency levels, and sample sizes from 100 to 500 at the interval length of 50. The results showed that the variation of sample sizes and spatial dependency levels do not affect the Mean Square Error (MSE) and bias of BME prediction. However, skewness affects the MSE of prediction but does not affect the bias. This result indicates that BME is robust to sample size and is unbiased. Despite the significant difference due to skewness, a graphical plot showed values of MSE close to zero, suggesting that BME can be considered robust to skewness.

Keywords: samples size; spatial dependency; skewness; Bayesian Maximum Entropy

1. Introduction

The Bayesian Maximum Entropy (BME) method was described as a robust technique that is not sensitive to skewed and biased data and can generate a posterior probability with minimum uncertainty from prior information [1]. However, BME has been mainly applied to Gaussian pdfs, i.e., non-skewed data [2]. A logarithmic transformation is often applied when the data is skewed to correct the skewness before using the BME method. The posterior probability derived from the transformed variable is transformed back to yield the estimates of the variable [3]. Using simulated data, Orthon and Lark [2] demonstrated that applying the standard BME method with logarithmic statistical moments included in the general knowledge base, gives a better result by reducing the MSE and bias of the prediction. This is applied to cases where the first and second-order moments represent the broad knowledge base, and the mean and covariance matrix of the transformed SRF are known. In BME analysis, non-Gaussian pdfs arise from several sources, which result mainly from either the prior or the posterior stages. At the prior stage, non-Gaussian pdfs can be due to experimental data from moments of order higher than two or directly from the physical laws that govern the physical process. In contrast, at the posterior stage, a non-Gaussian posterior pdf can also result from a non-Gaussian prior pdf [4] Also, environmental data, including soil data, often exhibit a strong positive skewness. In this case, the Gaussian assumption for the Spatial random field (SRF), may not be justified [5]. The skewed pdfs affect entropy measures and yield a smaller measure of entropy than a flatter pdf when a pdf has a greater degree of peakedness. A flat pdf shows a larger uncertainty about the variables thus, much information about this variable, while high peakedness shows more certainty, thus providing less information [2]. Also, when the data is strongly skewed, the derived variogram underestimates the spatial dependence due to few tremendous values contributing to many squared differences [6]. In addition, both sample size and sampling grids play a significant role in sampling natural resources.

The grid and sample size contribution to prediction accuracy, is an important parameter in any environmental modelling. It is crucial for better monitoring and sound decision making. The planning of any soil survey requires a sampling plan which specifies sample size and sampling grid. In most spatial prediction methods, the smaller resolution (grid size), the greater the precision in estimations [7]. The larger the sample size, the more significant the reduction of MSE obtained using the BME analysis method [5]. Several suggestions were made regarding sampling intensity for geostatistical analysis. Kerry and Oliver [8], proposed a minimum sample size of 100 locations, while they also suggested an optimum of 150 to 200 locations for a better variogram calculation. For soil analysis on a national scale, Sun et al. [9] recommended an optimal sample size to range from 500 to 1000 locations. As a result, it is necessary to numerically evaluate the robustness of BME to those factors. Moreover, most of the suggestions are made about classical geostatistics. Therefore, the present work seeks to reveal BME robustness to skewness with varying sample sizes and spatial dependency levels.

2. Methods

*2.1. Brief Bckground on Bayesian Maximum Entropy**

The Bayesian Maximum Entropy (BME) is a method for predicting spatial and temporal data [10]. BME is a knowledge-based approach with a rigorous mathematical background and inference schemes [11, 4]. It combines the maximum entropy theory with operational Bayesian statistics to construct its scientific mathematical framework [12]. Initially developed for continuous variables [12,13], BME was extended to categorical variables (ordinal or nominal variables) [14, 15], and then the combination of both categorical and continuous variables [16].

The BME analysis involves three steps namely the prior stage, the meta-prior stage and the posterior stage [4, 17, 18]. At the prior stage, data on the general knowledge is collected and processed to build the prior distribution. The prior G pdf $f_g(x_{map})$ is expressed in the general form:

$$f_g(x_{map}) = e^{(\mu_0+j)}, j = \sum_{\mu} \alpha\,(P_{map}) g_\alpha(x_{map}), \tag{1}$$

In the Meta-prior stage, the hard and soft data on the site are collected, arranged and transformed as possible to build the site-specific knowledge *S*. According to Christakos, **[13]** the quality and quantity of both data collected is a matter of experimental and computational investigations. In this step, Information from Prior and Metaprior are tied together using logical rules. Therefore, the posterior step uses the laws of probability to process the available site-specifc data (i.e. any measurement, hard or soft, recorded from the site) using the prior pdf and considering soft and hard data, the Bayes condition is applied to *fg*(*xmap*) to fnd the general posterior probability distribution function, *fk*(*xk*), expressed as follow:

$$f_x(x_k) = f_g(x_k|x_{data}) = \frac{f_g(x_k, x_{data})}{f_g(x_{data})}. \tag{2}$$

BME robustness to spatial prediction was evaluated by testing several sample sizes, skewness's and spatial dependency levels. The sample sizes considered were 100, 150, 200, 250, 300, 350, 400,450 and 500. Symmetric data (skewness=0) and data positively skewed by 1, 3,6 and 9 were considered. Data with strong (≤ 25%), moderate (25-77%) and weak (>75%) relative nugget effects (NE) [19].

$$NE(\%) = \frac{c_0}{c_0 + c_1} * 100. \tag{3}$$

Each spatial dependency levels and the ranges were repeated three times by spatial dependency levels shown in the table 1.

Table 1. simulation parameters.

Spatial dependency	Repetitions	Nugget	Psill	Range
Weak	1	0.9	0.1	6
	2	0.8	0.2	5
	3	0.7	0.3	4
Moderate	1	0.5	0.5	6
	2	0.4	0.6	5
	3	0.3	0.7	4
Strong	1	0.2	0.8	6
	2	0.1	0.05	5
	3	0.05	0.95	4

2.2. Simulation design

A random space field was simulated with a spherical covariance model having a sill $c_0 + c_1$ of 1 and ranges over three spatial dependency levels, as presented in table 1 [2, 20]. Two types of data were generated: hard and soft data. The hard data were generated using the unconditional sequential simulation method [21]. (See simulib from BMELIB package) to generate gaussian and non-gaussian data. For each spatial dependency level, fifteen thousand independent realizations of the standard gaussian variable were generated using a function built in R 3.6.2 (R Core Team 2019). A constant (range) was added to the value to yield a positive minimum value. Then, each value was raised to a given power, this exponentiation gives the data a positive coefficient of skewness. The data were then standardized to zero mean and unit variance [22]. The simulation function was based on the following arguments: Alpha (the exponentiation power to generate highly skewed values) and Skewness (the variable with this skewness is returned out of the 50,000 variables simulated: Sample size, Nugget effect, Effective range, Partial sill, and Variogram model).

Interval-type soft data was assumed to be randomly located over the unit size square. The width of the intervals was equal to 1.5, such that the simulated values were uniformly distributed over the intervals. In addition, a function was built to generate location parameters from a shapefile of 1ha extracted from the map of Benin republic by specifying the soft data size and the hard data size.

2.3. BME parameters specifcation and computation

BME method computation was performed using BME modes with the aim of predicting the best values at the nodes of grids. The computations were executed in Matlab 2016a with the library BMElib 2.0c.[23].

This research was carried out to compare 9 sample sizes, 5 degrees of skewness and 3 spatial dependency level in 3 repetitions, giving 405 ($9 * 5 * 3 * 3$) combinations of parameters levels to be compared. The sample sizes vary from 100 to 500, so the largest sample size was considered a reference sample. The prediction accuracy can be assessed using the Mean Square Error (MSE), Mean Absolute Error (MAE) or Mean Error (ME), and a good model is expected to have MSE, MAE or ME close to 0.24 The accuracy of each sample size (from 100 to 450) is compared to the reference sample (500) to find the optimum sample size. Then, firstly, the exploratory analysis was applied to assess the overall data variability in estimates and the MSE summaries [24]. The bias of prediction was also estimated. Secondly, the effect of sample size, skewness and spatial dependency level were graphically plotted using R 3.6.2 (R Core Team, 2019). Finally, the significance of the effect of the factors considered (skewness, sample size, and spatial dependency) on the MSE and bias were assessed using Analysis of Variance (ANOVA). The Materials and Methods should be described with sufficient details to allow others to replicate and build on the published results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 September 2022 doi:10.20944/preprints202209.0082.v1

3. Results

3.1. Effect of skewness, sample size and spatial dependency on BME performance

We notice from Table 2 that MSE and bias are significantly affected by the interaction between the Skewness and the spatial dependence (SD), with p-value, respectively 0.016 and 0.019. However, the sample size, and its interaction with the spatial dependency and skewness do not affect BME accuracy. This indicates that BME is robust to sample size variation and spatial dependence but sensitive to the combined effect of skewness and spatial dependency.

Table 2. Test of sample size, skewness and spatial dependency on MSE and bias of BME.

Source	MSE			Bias	
	DF	F-value	Pr (> F)	F-Value	Pr (> F)
Skewness (Sk)	4	116.94	< 0.001	1.91	0.110
SD	2	1.23	0.293	2.73	0.067
Sample size (Ss)	7	0.38	0.921	0.97	0.514
Sk:SD	8	2.40	0.016	2.36	0.019
Sk:Ss	28	0.64	0.921	0.97	0.514
SD:Ss	14	0.85	0.617	0.64	0.829
Sk:SD:Ss	56	0.89	0.689	1.19	0.185

Legend: SD: Spatial dependency; Sk: Skewness; Ss: Sample size

For the MSE, there is a significant variation in BME performance based on skewness and spatial dependency (Figure 1). For strong spatial dependency, no significant difference exists between the BME prediction accuracy of data skewed by 0 and 1; 0 and 3; and 1 and 3. However, there is a significant difference between data skewed by 0 and 6; 0 and 9; 1 and 6; 1 and 9; 3 and 6; and 6 and 9. For moderate spatial dependency, only data skewed by 9 is significantly different from 0. Also, no significant difference was observed between data skewed by 3 and 6. Only data skewed by 3, 6, and 9 are significantly different from 0 for weak spatial dependency. Also, no significant difference was observed between data skewed by 3 and 6. This means that skewness greater than 3 is highly significant from 0 in the MSE, regardless of spatial dependency. Also, strong spatial dependency is more accurate when data is skewed by 6 and 9.

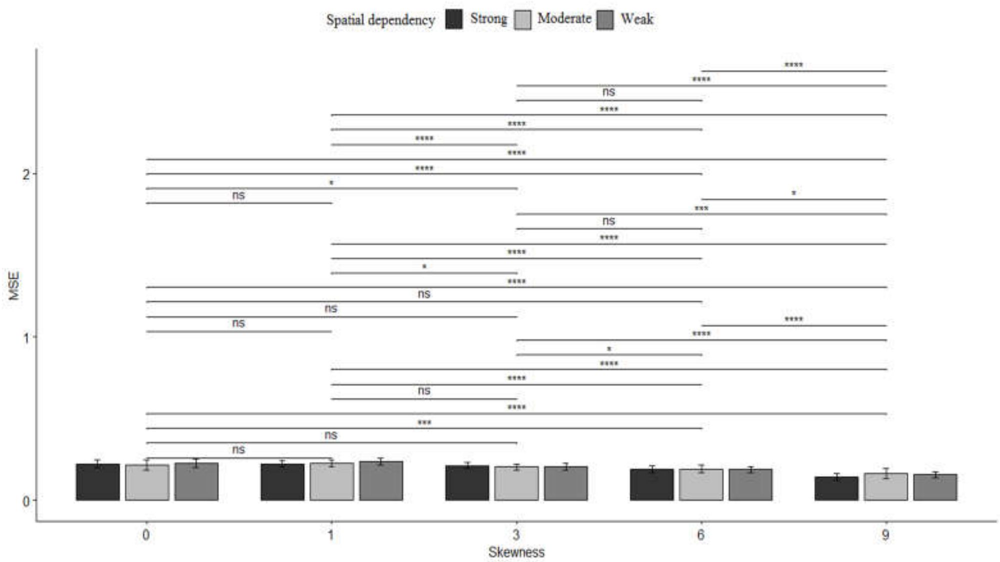

Figure 1. MSE of the effect of skewness and spatial dependence.

 doi:10.20944/preprints202209.0082.v1

The variation of bias of BME prediction between skewness and spatial dependency (Figure 2) reveals that despite the significance of the difference between all degree of skewness irrespective to spatial dependency, there is small variation with the bias very close to zero. No significant different between data skewed by 1 and 0 is observed. Data skewed by 3 is significantly different from data skewed by 0 only in weak spatial dependency. When data is skewed by 6, significant different from 0 is observed in strong and weak spatial dependency. There is a significant difference between data skewed by 9 and 0 irrespective of the spatial dependency level. This indicate that in the bias, skewness greater than 3 is highly significantly different from 0 irrespective of the Spatial dependency. Additionally, strong spatial dependency is more accurate when data is skewed by 6 and 9. Weak spatial dependency is more accurate when data is skewed by 0, 1 and 3.

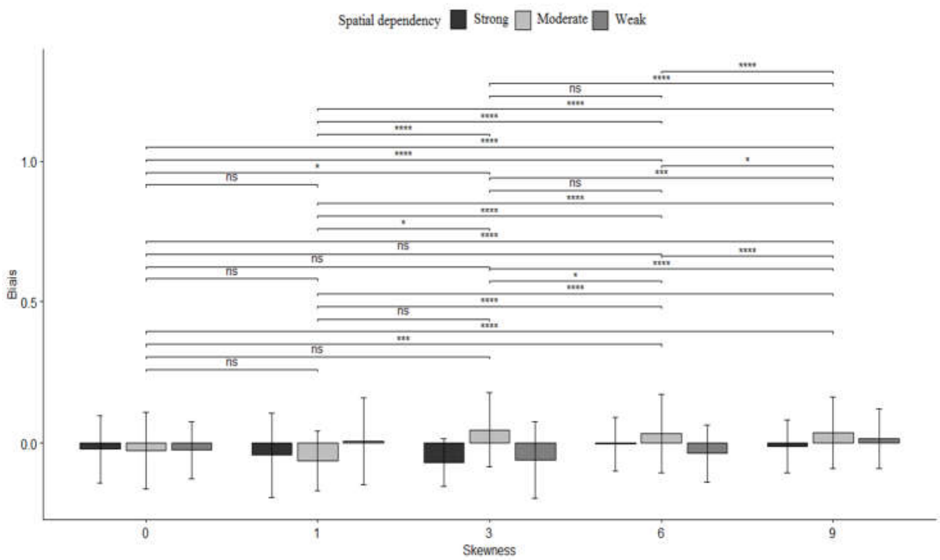

Figure 2. Bias of the effect of skewness and spatial dependence.

4. Discussion and conclusion

Bayesian Maximum Entropy (BME) is a spatiotemporal approach that provides a unique framework for combining theoretical concepts and physical knowledge beyond classical methods. According to Christakos [13] most spatial temporal techniques, such as regression methods, spline functions, basis functions, etc. have reached their limits. New scientifc methods capable of considering their shortages needed to be developed. He discussed four main issues related to classical methods: scientific content, indetermination thesis, spatiotemporal geometry, and sources of physical knowledge. Furthermore, he presented BME as a robust and flexible method for the researcher's choice of covariance, variogram, and generalized covariance models. Thus, this research was carried out to fully investigate Bayesian Maximum Entropy (BME) robustness empirically in spatial estimation. We specifically assessed the effect of skewness level, sample size and spatial dependency level on BME prediction accuracy. This study revealed that, firstly,

BME predictions are robust to sample size variations in terms of MSE and bias of prediction. BME accuracy is affected by neither sample size nor its interaction with spatial dependence or skewness, suggesting that BME provides accurate estimation without being affected by large samples requirement or limited by high soil sampling costs. According to Somarathna et al. [25], the sample size is the major driving factor of prediction accuracy. Additionally, larger sample sizes deliver equally good prediction accuracy despite the model type and when the sample size increases, the model's uncertainty related to the prediction decreases. The robustness to sample size and higher accuracy result from the use of a wide range of soft data. Lee et al. [26] demonstrated that the use of soft data in Urban Climate Research increased BME accuracy up to 35.28 percent over traditional linear kriging analysis and reduced the need for costly sampling protocols. This could also justify the large adoption of BME by many research felds [4]. According to Hengle et al.,[27] sampling cost is also used to evaluate map quality in pedology in terms of cost per area (eg. 2 US$ per km2). Secondly, the effect of skewed hard data (1,3, 6 and 9) on BME was evaluated. The results showed that skewness highly affects the MSE of prediction but does not affect the bias of prediction. Furthermore, the interaction between the skewness and the spatial dependence affected both the MSE and bias. These results indicate BME sensitivity to skewed data. Christakos [13] assessed the effect of incorporating skewness as prior knowledge into mapping numerically; he considered prior skewed by 0, 2, 3, and 3.5. His results indicated that the error pdf's varies with the skewness values. Despite the significance in MSE induced by the degree of skewness, we still observed that MSE remains very close to zero. Only data skewed by 6 or more are significantly different from symmetric data (skewness=0). This is an indication of BME robustness to skewness as compared to Kriging. The prediction bias was not significant in a variation of skewness and sample size which indicates BME unbiasedness. Thirdly, the variation due to spatial dependency level did not affect the prediction's MSE and bias. However, the interaction between the skewness and the spatial dependency affected both the MSE and bias of prediction.

These results are in accordance with Christakos [13, 11] findings on BME robustness in spatial estimation. It clearly showed that Bayesian Maximum Entropy (BME) is robust to sample size and spatial dependency. In fact, the prediction's bias was not significant in a variation of skewness and sample size which indicates BME unbiasedness. We conclude that BME is robust to sample size, spatial dependency, skewness (less than 6).

Supplementary Materials: Data used was simulated, it will be available on GitHub (https://github.com/) with the scripts if the manuscript is accepted.

Author Contributions: Conceptualization, methodology, software, funding acquisition, writing—original draft preparation, formal analysis.; Ehnon Gongnet E., Conceptualization, methodology and writing—original draft preparation, Agbangba Codjo E., formal analysis, data curing, Affossogbe S.T, validation, review and editing Vihotogbe R.[3], validation, review and editing, supervision, Glèlè Kakaï R.

Funding: This research was funded by DAAD In-Country/In-Region Scholarship Programme FSA/UAC, 2020, grant number 57546598.

Informed Consent Statement: "Not applicable." Data was simulated.

Data Availability Statement: Data and simulation code will be available on github (https://github.com/EHNONhub/BME-data.git).

Acknowledgments: We thank Mr Yassine Doudoua, Dr Nadjiaim Djirabaye and Dr Djondang Koye, the staff of ITRAD and LABEF for their technical support.

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

References

1. Hayunga, D. K.; Kolovos, A. Geostatistical space–time mapping of house prices using Bayesian maximum entropy. *International Journal of Geographical Information Science* 2016, 30(12), 2339-2354. https://doi.org/10.1080/13658816.2016.1165820
2. Orton, T.G.; Lark, R.M. Bayesian maximum entropy method for lognormal variables. *The Stochastic Environmental Research and Risk Assessment.* 2009; 23(3), 319-328. https://doi.org/10.1007/s00477-008-0217-7.
3. Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environmental science & technology* 2012, 46(5), 2772-2780. https://doi.org/10.1021/es203152a
4. He, J.; Kolovos, A. Bayesian maximum entropy approach and its applications: a review. *Stochastic Environmental Research and* Risk Assessment 2018, 32(4), 859-877. https://doi.org/10.1007/s00477-017-1419-7
5. Orton, T.; Rawlins, B.; Lark, R.M. Using measurements close to a detection limit in a geostatistical case study to predict selenium concentration in topsoil. Geoderma 2009, 152 269–282. https://doi.org/10.1016/j.geoderma.2009.06.014 .
6. Webster, R.; Oliver, M. A. Geostatistics for environmental scientists; John Wiley: Chichester, UK, 2001.
7. Hengl T. A Practical Guide to Geostatistical Mapping. European Communities. 2009. 22904, ISBN: 978-92-79-06904-8.

8. Kerry, R.; Oliver, M.A. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 2007, 140, 383–396. https://doi.org/10.1016/j.geoderma.2007.04.019
9. Sun, Y. X.; Wu, C.Z.; Zhu, K. B.; Cui, Z. L.; Chen, X. P.; Zhang, F. S. Influence of interpolation method and sampling number on spatial prediction accuracy of soil Olsen-P. Ying Yong Sheng tai xue bao= The Journal of Applied Ecology 2009, 20(3), 673-678.
10. Gao, S.; Zhu, Z.; Liu, S.; Jin, R.; Yang, G.; Tan, L. Estimating the spatial distribution of soil moisture based on Bayesian maximum entropy method with auxiliary data from remote sensing. Int. J. Appl. Earth Obs. Geoinf. 2014, 32, 54-66. https://doi.org/10.1016/j.jag.2014.03.003
11. Christakos, G. Spatiotemporal information systems in soil and environmental sciences. Geoderma 1998; 85(2-3), 141-179. https://doi.org/10.1016/S0016-7061(98)00018-4.
12. Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. Math Geol 22:763–777.
13. Christakos, G. Modern spatiotemporal geostatistics. Oxford University Press, New York. 2000;
14. P. Bogaert (2002), Spatial prediction of categorical variables: The Bayesian maximum entropy approach. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-002-0114-4
15. D'or, D., & Bogaert, P. (2004). Spatial prediction of categorical variables with the Bayesian maximum entropy approach: the Ooypolder case case study. European journal of soil science, 55(4),763-775.
16. Bogaert P. and Wibrin M. A. (2005). Combining categorical and continuous information using Bayesian Maximum Entropy. https://doi.org/10.1007/3-540-26535-X_2
17. Salman, R.; Nikoo, M.R.; Shojaeezadeh, S.A.; Beiglou, P. H. B.; Sadegh, M.; Adamowski, J.F.; Alamdari, N. A novel Bayesian maximum entropy-based approach for optimal design of water quality monitoring networks in rivers. Journal of Hydrology 2021, 603, 126822. https://doi.org/10.1016/j.jhydrol.2021.126822
18. Douaik, A.; Van Meirvenne M.; Toth, T.; Serre, M. Space-time mapping of soil salinity using probabilistic Bayesian maximum entropy. Stochastic Environmental Research and Risk Assessment 2004,18: 219-227. https://doi.org/10.1007/s00477-004-0177-5
19. Barbosa, I. C., Appel, E., Seidel, E. J., & Oliveira, M. S. D. (2017). Proposal of the spatial dependence evaluation from the power semivariogram model. Boletim de Ciências Geodésicas, 23, 461-475. https://doi.org/10.1590/S1982-21702017000200031
20. Lark, R.M. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of soil science 2000, 51(1), 137-157. https://doi.org/10.1046/j.1365-2389.2000.00280.x.
21. Christakos, G.; Bogaert, P.; Serre, L. Advanced functions of temporal GIS. Springer 2001, 978-3-642-56540-3
22. Lark, R.M. Estimating variograms of soil properties by the method-of moments and maximum likelihood. European Journal of Soil Science 2000, 51, 717–728. https://doi.org/10.1046/j.1365-2389.2000.00345.x.
23. Christakos, G.; Bogaert, P.; Serre, M. Temporal GIS: advanced functions for feld-based applications. Springer Science & Business Media. 2002;
24. John, K.; Abraham I.I.; Kebonye, N.M.; Agyeman, P.C.; Ayito, E.O.; Kudjo, A.S. Soil organic carbon prediction with terrain derivatives using geostatistics and sequential Gaussian simulation. Journal of the Saudi Society of Agricultural Sciences 2021, 20(6), 379-389. https://doi.org/10.1016/j.jssas.2021.04.005.
25. Somarathna P.D.S.N.; Minasny, B.; Malone, B.P. More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. Soil Science Society of America Journal. 2017; 81(6), 1413. https://doi.org/10.2136/sssaj2016.11.0376.
26. Lee, S.J.; Balling, R.; Gober, P. Bayesian maximum entropy mapping and the soft data problem in urban climate research. Annals of the Association of American Geographers 2008, 98(2), 309-322. https://doi.org/10.1080/00045600701851184.
27. Hengl, T.; Nikolić, M.; MacMillan, R.A. Mapping effciency and information content. International Journal of Applied Earth Observation and Geoinformation 2013; 22, 127–138. https://doi.org/10.1016/j.jag.2012.02.005.