A new local stochastic method for predicting data with spatial heterogeneity

Anderson Rodrigo da Silva[*] , **Ana Paula Alencastro Silva and Lauro Joaquim Tiago Neto**

Laboratório de Estatística e Geoprocessamento, Instituto Federal Goiano, Rodovia Geraldo Silva Nascimento, Km 2,5, 75790-000, Urutaí, Goiás, Brazil.
*Author for correspondence. E-mail: anderson.silva@ifgoiano.edu.br

ABSTRACT. Spatial data (e.g., phytopathogenic data) do not always meet assumptions such as stationarity, isotropy and Gaussian distribution, thereby requiring complex spatial methods and models. Some deterministic assumption-free methods such as the inverse distance weighting can also be applied to predict spatial data, but their output is limited for graphical solutions (mapping). We adapted a computer-based prediction method called Circular Variable Radius Moving Window (CVRMW) that is based on two others: moving window kriging (MWK) and inverse squared-distance weighting (ISDW). The algorithm is developed to meet an objective function that minimizes the index of variation of the spatial observations inside the moving window. A code in R language is presented and thoroughly described. The outputs include the range of the spatial dependence as the radius calculated at every target location and the standard error of the predicted values, mapped to provide a useful tool for spatial exploratory analysis. The method does not make any assumptions about the spatial process, and it is an alternative for dealing with spatial heterogeneity.

Keywords: moving window kriging; spatial prediction; soil nematodes.

Introduction

The knowledge about the within-field patterns of the spatial distributions of crop diseases, pests and their natural enemies is critical to planning control tactics, developing efficient sampling plans and predicting damage (Martins et al., 2018). The spatial distributions of many species of plant pathogens still remain understudied. Moreover, the spatial variation of the soil properties within fields offers various conditions for the diverse development of crops and many soil-borne pests (Patzold et al., 2008; Hbirkou et al., 2011). For instance, heterogeneous variables such as the occurrence of cyst nematodes require additional samples at shorter distances to ensure that the spatial variation is adequately addressed (Hbirkou et al., 2011). According to Freitas et al. (2017), soil chemical variables can also affect the spatial distribution of nematodes.

Studies of the spatial distributions of agricultural pests have been carried out using aggregation indices (Holguin, Mueller, Khalilian, & Agudelo, 2015) such as the Morisita index, the variance/average ratio, Taylor's Power law (Taylor et al., 2017) and others. However, these statistics are nonspatial, that is, they do not consider the spatial dependence of the data. Geostatistical methods based on variogram analyses have also been widely recommended and applied (Pimentel, Lopes, Mexia, & Mumford, 2017; Pulakkatu-Thodi, Reisig, Greene, Reay-Jones, & Toews, 2014; Rijal, Brewster, & Bergh, 2014; Masetti, Butturini, Lanzoni, De luigi, & Burgio, 2015; Martins et al., 2017; Yawson et al., 2017; Freitas et al., 2017) and should be preferred. In these methods, a common assumption is stationarity of first and/or second degree, when the mean and variance are constant over the area. Of course, each of the parameters may be independently stationary (Terrien, Germain, Marque, & Karlsson, 2013; Ching & Lin, 2014). Nonetheless, this assumption is rather an exception since the values of a spatial process over the landscape are conditioned by different factors, which causes them to lose their intrinsic nature (Siqueira et al., 2015).

Some spatial data such as the population density of phytopathogenic nematodes in a field have a local character. In this case of nonstationarity, a heterogeneous variogram can be obtained using a moving window centered on the location to be predicted, which is an approach known as Moving Window Kriging (MWK), as described by Hass (1990). Applications of MWK in soil data were carried out by Walter,

McBratney, Douaoui, and Minasny (2001) and Cattle, McBratney, and Minasny (2002). A local variogram is determined at every target location, in contrast with the conventional forms of kriging where there is only one global variogram (isotropy) produced for the entire dataset (Demšar & Harris, 2011).

The spatial and temporal heterogeneity of crops can also make it difficult to map pest or weed infestations. According to Baillod, Tscharntke, Clough, and Batáry (2017), the spatial heterogeneity of crops can be described both by their composition (e.g., crop diversity) and spatial arrangement (e.g., average field size). Temporal crop heterogeneity describes the changes in crop patterns due to the annual succession of crops. For instance, Bertrand, Burel, and Baudry (2016) found that some beetle species with high dispersal power such as *Trechus quadristriatus* were more abundant in landscapes with high spatial heterogeneity, whereas the abundance of less mobile species such as *Poecilus cupreus* was only positively influenced by the temporal crop dynamics.

Infestation maps obtained by kriging are often based on semivariance models for continuous variables that follow a known distribution. Ordinary kriging implicitly assumes a Gaussian-like distribution. It is not rare to find works which data are firstly transformed to approximate normality or to remove variance trends. According to Yamamoto and Landin (2013), a semivariogram is quite sensitive to the statistical distribution. Models that incorporate information from discrete variables (Diggle, Tawn, & Moyeed, 1998) are complex and have little appeal. On the other hand, simpler models are, in general, limited by spatial assumptions. Moreover, some variograms can show a common feature known as a "pure nugget effect", suggesting to avoid kriging.

Deterministic interpolation approaches, such as the inverse squared-distance weighting method (ISDW), can also be used to build infestation maps. This method does not make any assumptions about the data and is not model-based. However, unlike the geostatistical approach, the essential parameters in crop pest studies such as the degree of spatial dependence and the practical range are not estimated, thus making it impracticable to carry out studies on sampling for pest monitoring purposes or studies on within-field insect dispersion/mobility.

In this paper we present a heuristic method based on the MWK approach, called the Circular Variable Radius Moving Window method (CVRMW), to perform interpolation/prediction of spatial phytopathogenic data as an alternative to kriging and other interpolation methods that can deal with spatial heterogeneity. A code in R language is also presented.

Material and methods

The dataset used to illustrate the method consists of 20 spatial observations of nematode density (individuals per dm^3 of soil). A 50 × 50 prediction grid was used for the predictions. Three methods were applied: ordinary kriging, ISDW and CVRMW. In order to evaluate and compare the prediction accuracies obtained with ISDW and CVRMW, we have calculated the percentual absolute mean error (PAME) through the difference between the observed and the predicted values. All the statistical analyses and computations were performed in R (http://www.R-project.org/). Additionally, a video illustrating the CVRMW method is available from: https://youtu.be/mmg5jRti8gs.

Calculation

The proposed method is basically a mixture of MWK and ISDW. In the latter, spatial predictions are performed through the weighted mean of all spatial observations via the inverse of the squared distances to the predicted point. In addition, the CVRMW method consists of predicting spatial points using only the data within a circular window. This is done through the mean weighted by the inverse of the squared-distances between a prediction location and the spatial points within a radius h that varies according to the location and is determined to meet the objective function of minimizing the index of variation (the ratio between the coefficient of variation and the square root of the sample size). The window is then moved to the next prediction location, and the procedure is repeated for all prediction points. Thus, both of the base methods of CVRMW meet Tobler's first law.

A mathematical formulation is given as follows. Take $Z(s_i)$ as the spatial observation at the coordinates s_i (i = 1, 2, ..., N) and $\hat{Z}(s_k)$ as the k-th point to be predicted. Then,

$$\hat{Z}(s_k) = \left[\sum_{i=1}^{N(h_k)} \frac{Z(s_i)}{d_{i,k}^2} \right] \times \left[\sum_{i=1}^{N(h_k)} \frac{1}{d_{i,k}^2} \right]^{-1}$$

where: $N(h_k)$ is the number of obsevations limited by the radius h_k and $d_{i,k}^2$ is the squared-distance between s_i and s_k.

Algorithm

The first step of the algorithm (Figure 1) for implementing CVRMW in R is to define the inputs: i) the sampling data, ii) the geographical coordinates of the sampling points, and iii) the prediction grid.

```
Algorithm: CVRMW
Input: data, coords, grid
Output: a "data.frame" containing spatial predictions, standard errors, the radius and the number
of observations used in each prediction over the grid
1. y <- data
   n <- nrow(coords)
   N <- nrow(grid)
   pred <- data.frame(grid, pred = NA, SE = NA, radius = NA, n = NA)
2. i <- 1
   repeat {
2.1   s <- sweep(coords, 2, as.matrix(grid[i, ]), FUN = "-")
      d <- sqrt(apply(s^2, 1, sum))
      w <- 1/d^2
      max.w <- max(w[is.finite(w)])
      w[!is.finite(w)] <- max.w
2.2.  iv <- function(x) sd(x)/( mean(x) * sqrt(length(x)) )
2.3   seq.d <- seq(min(d), max(d), length.out = 200)
      vari <- c()
      for(j in 1:length(seq.d)) {
         vari[j] <- iv( y[which(d <= seq.d[j])] )
      }
      pred$radius[i] <- seq.d[which.min(vari)]
2.4   id <- which(d <= pred$radius[i])
      pred$pred[i] <- weighted.mean(y[id], w[id])
      pred$n[i] <- length(y[id])
      pred$SE[i] <- sd(y[id])/sqrt(pred$n[i])
      i <- i + 1
      if (i > N) break()
   }
3. if (all(coords == grid)) {
      pame <- 100*mean(abs(data - pred$pred)/data)
      cat("PAME:", round(pame, 2), "%\n")
   }
4. Return
```

Figure 1. Pseudocode of CVRMW.

In Step 2, the loop process is properly given by defining an index variable 'i', a sequence ending at 'N', and the number of prediction points. In Step 2.1, the Euclidian distances between every sampling point (coords) and every point of the prediction grid are calculated and stored in the object 'd', which is a vector of dimension n (the number of sampling points). The prediction weights, w, are defined as the inverse squared-distances, and a restriction given in case 'd' is zero. In this case, w is assumed to be the maximum calculated value. In Step 2.2, an auxiliary function is written in order to calculate the index of variation. This makes the code faster and easier to debug. In Step 2.3, a sequence of length 200 is created to determine the radius to be used for predicting the i-th spatial point in 'grid'. The index of variation of the spatial observations within each radius is calculated and stored in 'vari'. The radius value is a value in 'seq.d' corresponding to the minimum value of 'vari'. In Step 2.4, those sampling points within the i-th radius or less are computed and stored in the object 'id', which is a list of characters indicating the sampling points. The i-th predicted value is the weighted mean of the spatial observations in 'id' considering their weights stored in 'w'. The prediction is stored as a column of the data frame 'pred'. The number of sampling points used for the i-th prediction and the standard error are stored as columns in 'pred' as well. The loop is broken when the last location in 'grid' is taken.

In Step 3, if the user passes the same input objects to 'coords' and 'grid', the percentual absolute mean error (PAME) is calculated and printed (a side effect).

Features

The function accepts a numeric vector containing the sampling data as the input for the argument 'data'. The argument 'coords' must be a numeric matrix or data.frame containing the geographical coordinates of the sampling points. The argument 'grid' must be a numeric matrix or data.frame with the same number of columns as 'coord'.

The output is an object of the 'data.frame' class with the same number of rows as 'grid'. The first 2 or 3 columns (in general) correspond to the input grid. Then, the following columns are used: 'pred' – the spatial predictions, 'SE' – the standard error of the predicted value, 'radius' – the length of the radius used for each prediction, and 'n' – the number of sampling points used for each prediction.

Results and discussion

The graphical solution of the spatial interpolation is given in Figure 2. With the three methods, an East-West gradient is observed, although the level of detail is lower and a smoother solution occurs with kriging. Using CVRMW, the predicted values do not represent unbiased predictions as in kriging methods. However, the PAME in map (C) for ISDW is 35.08% compared with 30.04% in map (D) for CVRMW. In addition, an analysis of the spatial dependence can be done through the frequency distribution of the calculated radii (Figure 3A). In this case, there is a pattern of spatial dependence within 0.0004 degrees. This is not as easily observed using a classical omnidirectional variogram (Figure 3B), where a wave model was fit. Nonetheless, we found the estimate $\phi = 0.0004$ for the range parameter to be quite close to the upper limit of the most frequent radius interval (0.0005 degrees) observed with CVRMW. The mean of the prediction standard errors using the CVRMW method was 365 ($\pm$150) nematodes dm^{-3} compared with 1,070 ($\pm$24) nematodes dm^{-3} using ordinary kriging.
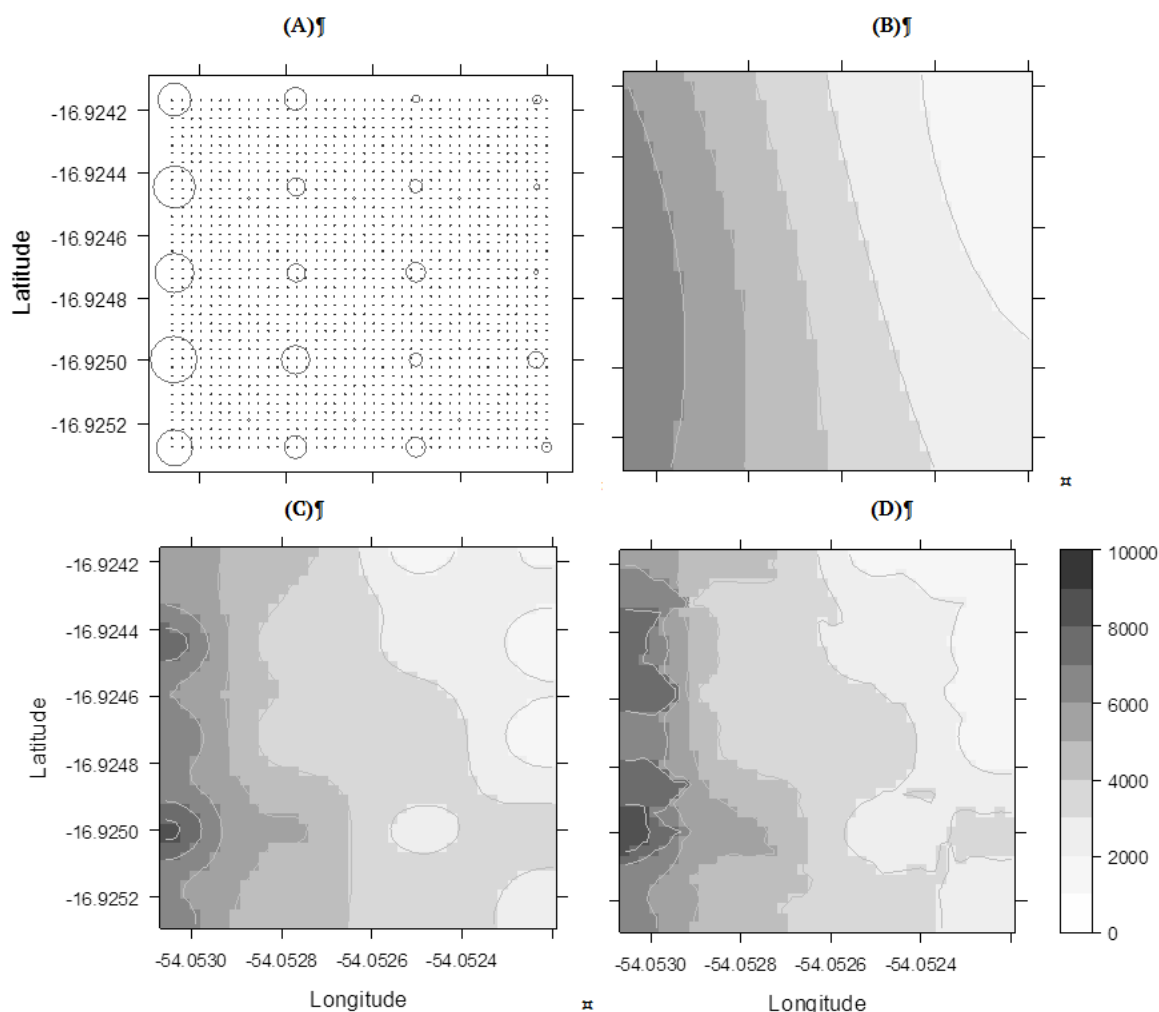

Figure 2. Infestation (nematodes dm^{-3} soil) maps. (A) The sampling points are in blue, the prediction grid is in red, and the point size represents the relative infestation; (B) heat map via ordinary kriging; (C) heat map via the ISDW method; and (D) heat map via the CVRMW method.

Yawson et al. (2017) did not find substantial differences in mapping the spatial distribution of the nematode population in organic and conventional fields using ISDW and ordinary kriging. Souza, Bazzi, Khosla, Uribe-Opazo, and Reich (2016) observed that ISDW performed slightly better at predicting crop yields than ordinary kriging. Rhodes, Liburd, and Grunwald (2011) obtained flower thrip infestation maps with similar accuracy using inverse distance weighting and ordinary kriging. On the other hand, Pimentel et al. (2017) stated that inverse distance weighting and ordinary kriging did not perform so well at predicting *Ceratitis capitata* infestations due to the spatial heterogeneity and topographical conditions. These authors recommended using a geographic weighted regression (GWR) instead.
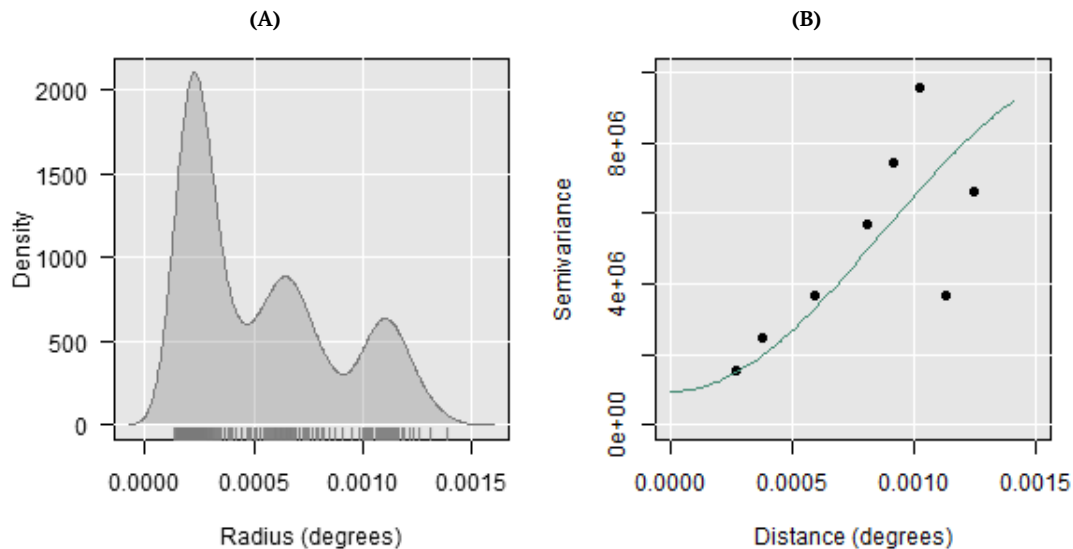
(A)   **(B)**

Figure 3. (A) Kernel density of the calculated radii of CVRMW and (B) omnidirectional classical variogram.

Spatial dependence is commonly studied through the analysis of variograms, such as that presented in Figure 3B. Nevertheless, using CVRMW allows one to actually see "where" the spatial dependence is larger in the field by building a map of the calculated radii.

Since predictions are performed using a loop, large datasets may increase the processing time. The timings for several analyses can be seen in Table 1. The larger the size of the prediction grid is, the more time used by the process. Nevertheless, note that the number of spatial observations (sampling points) only slightly affects the processing time.

Table 1. Timing* (in seconds) for the analyses performed with various data sizes.

Sampling Points	Prediction grid		
	20 × 20	50 × 50	100 × 100
20	5.68	44.09	223.06
60	5.71	47.59	223.28
200	5.75	46.76	226.53

*Based on an Intel® Core™ i5 1.40 GHz with 3.71 GB of RAM.

Conclusion

An alternative computer-based heuristic method to perform spatial predictions was developed and implemented in R. The method consists of a local interpolator with stochastic features. It allows one to build effective detailed maps and to estimate the spatial dependence without any assumptions on the spatial process.

Acknowledgements

References

Baillod, A. B., Tscharntke, T., Clough, Y., & Batáry, P. (2017). Landscape-scale interactions of spatial and temporal cropland heterogeneity drive biological control of cereal aphids. *Journal of Applied Ecology*, *54*(6), 1804-1813. DOI: 10.1111/1365-2664.12910

Bertrand, C., Burel, F., & Baudry, J. (2016). Spatial and temporal heterogeneity of the crop mosaic influences carabid beetles in agricultural landscapes. *Landscape Ecol*ogy, *31*(2), 451-466. DOI: 10.1007/s10980-015-0259-4

Cattle, J. A., McBratney, A. B., & Minasny, B. (2002). Kriging methods evaluation for assessing the spatial distribution of urban soil lead contamination. *Journal of Environmental Quality*, *31*(5), 1576-1588. DOI: 10.2134/jeq2002.1576

Ching, J., & Lin, C. J. (2014). Probability distribution for mobilized shear strengths of saturated undrained clays modeled by 2-D Stationary Gaussian Random Field-A 1-D stochastic process view. *Journal of Mechanics*, *30*(3), 229-239. DOI: 10.1017/jmech.2014.9

Demšar, U., & Harris, P. (2011). Visual comparison of Moving Window Kriging Models. *Cartographica*, *46*(4), 211-226. DOI: 10.3138/carto.46.4.211

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, *47*(3), 299-350. DOI: 10.1007/978-0-387-48536-2

Freitas, J. R. B., Moitinho, M. R., De Bortoli, D. T., Silva, E. B., Silva, J. F., Siqueira, D. S., & Pereira, G. T. (2017). Soil factors influencing nematode spatial variability in soybean. *Agronomy Journal*, *109*(2), 610-619. DOI: 10.2134/agronj2016.03.0160

Haas, T. C. (1990). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, *85*(412), 950-963. DOI: 10.2307/2289592

Hbirkou, C., Welp, G., Rehbein, K., Hillnhütter, C., Daub, M., Oliver, M. A., & Pätzold, S. (2011). The effect of soil heterogeneity on the spatial distribution of Heterodera schachtii within sugar beet fields. *Applied Soil Ecology*, *51(1)*, 25-34. DOI: 10.1016/j.apsoil.2011.08.008

Holguin, C. M., Mueller, J. D., Khalilian, A., & Agudelo, P. (2015). Population dynamics and spatial distribution of Columbia lance nematode in cotton. *Applied Soil Ecology*, *95*(1), 107-114. DOI: 10.1016/j.apsoil.2015.06.004

Martins, J. C., Picanço, M. C., Silva, R. S., Gonring, A. H., Galdino, T. V., & Guedes, R. N. (2018). Assessing the spatial distribution of *Tuta absoluta* (Lepidoptera: Gelechiidae) eggs in open-field tomato cultivation through geostatistical analysis. *Pest Management Science*, *74*(1), 30-36. DOI: 10.1002/ps.4664

Masetti, A., Butturini, A., Lanzoni, A., De luigi, V., Burgio, G. (2015). Area-wide monitoring of potato tuberworm (*Phthorimaea operculella*) by pheromone trapping in Northern Italy: phenology, spatial distribution and relationships between catches and tuber damage. *Agricultural and Forest Entomology*, *17*(2), 138-145. DOI: 10.1111/afe.12089

Patzold, S., Mertens, F. M., Bornemann, L., Koleczek, B., Franke, J., Feilhauer, H., & Welp, G. (2008). Soil heterogeneity at the field scale: a challenge for precision crop protection. *Precision Agriculture*, *9*(6), 367-390. DOI: 10.1007/s11119-008-9077-x

Pimentel, P., Lopes, D. J. H., Mexia, A. M. M., & Mumford, J. D. (2017). Validation of a geographic weighted regression analysis as a tool for area-wide integrated pest management programs for Ceratitis capitata Wiedemann (Diptera: Tephritidae) on Terceira Island, Azores. *International Journal of Pest Management*, *63*(2), 172-184. DOI: 10.1080/09670874.2016.1256512

Pulakkatu-Thodi, I., Reisig, D. D., Greene, J. K., Reay-Jones, F. P., & Toews, M. D. (2014). Within-field spatial distribution of Stink Bug (Hemiptera: Pentatomidae)-induced boll injury in commercial cotton fields of the southeastern United States. *Environmental Entomology*, *43*(3), 744-752. DOI: 10.1603/EN13332

Rijal, J. P., Brewster, C. C., & Bergh, J. C. (2014). Spatial distribution of Grape Root Borer (Lepidoptera: Sesiidae) infestations in Virginia vineyards and implications for sampling. *Environmental Entomology*, *43*(3), 716-728. DOI: 10.1603/EN13285

Rhodes, E. M., Liburd, O. E., & Grunwald, S. (2011). Examining the spatial distribution of flower thrips in southern highbush blueberries by utilizing geostatistical methods. *Environmental Entomology*, *40*(4), 893-903. DOI: 10.1603/EN10312

Siqueira, G. M., Silva, J. S., Bezerra, J. M., Silva, E. F. F., Dafonte, J., & Melo, R. F. (2015). Estacionariedade do conteúdo de água de um Espodossolo Humilúvico. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *19*(5), 439-448. DOI: 10.1590/1807-1929/agriambi.v19n5p439-448

Souza, E. G., Bazzi, C. L., Khosla, R., Uribe-Opazo, M. A., & Reich, R. M. (2016). Interpolation type and data computation of crop yield maps is important for precision crop production. *Journal of Plant Nutrition*, *39*(4), 531-538. DOI: 10.1080/01904167.2015.1124893

Taylor, R. A. J., Park, S. J., & Grewal, P. S. (2017). Nematode spatial distribution and the frequency of zeros in samples. *Nematology*, *19*(3), 263-270. DOI: 10.1163/15685411-00003045

Terrien, J., Germain, G., Marque, C., & Karlsson, B. (2013). Bivariate piecewise stationary segmentation; improved pre-treatment for synchronization measures used on non-stationary biological signals. *Medical Engineering & Physics*, *35*(8), 1188-1196. DOI: 10.1016/j.medengphy.2012.12.010

Walter, C., McBratney, A. B., Douaoui, A., & Minasny, B. (2001). Spatial prediction of topsoil salinity in the Chelif Valley, Algeria, using local ordinary kriging with local variograms versus whole-area variogram. *Australian Journal of Soil Research*, *39*(2), 259-272. DOI: 10.1071/SR99114

Yamamoto, J. K., & Landin, P. M. B. (2013). *Geoestatística*: conceitos e aplicações. São Paulo, SP: Oficina de Textos.

Yawson, D. O., Adu, M. O., Vanderpuije, G., Arthur, W. S., Aggrey, K. T., & Boateng, E. (2017). Spatial distribution of nematodes at organic and conventional crop fields in Cape Coast, Ghana. *West African Journal of Applied Ecology*, *25*(1), 57-66.