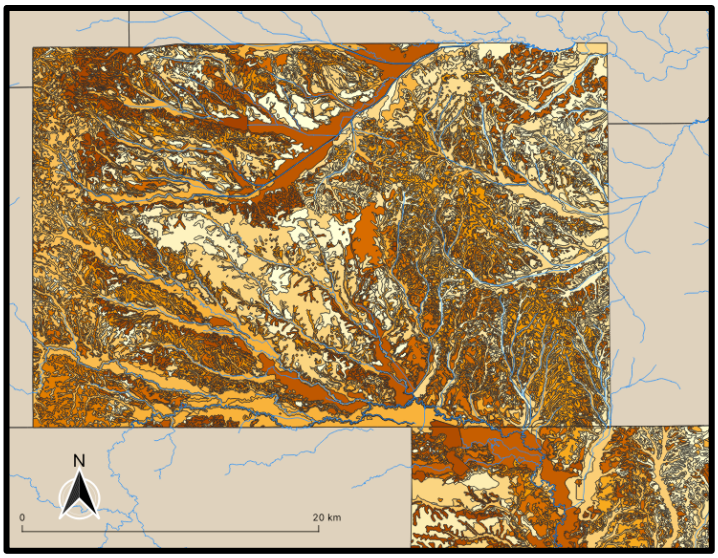
Landscape context influences accuracy of predicted georeferenced soil data: Implications for research



Adrián Lázaro-Lobo[1], Gary N. Ervin[2]

[1] Biodiversity Research Institute IMIB (University of Oviedo-CSIC-Principado de Asturias), Mieres, Spain; adrianlalobo@gmail.com

[2] Department of Biological Sciences, Mississippi State University, MS 39762, USA; gary.ervin@msstate.edu

Geosystems Research Institute Report #5104

May 2024

MISSISSIPPI STATE UNIVERSITY
DEPARTMENT OF BIOLOGICAL SCIENCES

GRI

Citation: Lázaro-Lobo, A. and G. N. Ervin. 2024. Landscape context influences accuracy of predicted georeferenced soil data: Implications for research. Geosystems Research Institute Report 5104, Geosystems Research Institute, Mississippi State University, MS.

Landscape context influences accuracy of predicted georeferenced soil data: Implications for research

Adrián Lázaro-Lobo[1], Gary N. Ervin[2]

[1] Biodiversity Research Institute IMIB (University of Oviedo-CSIC-Principado de Asturias), Mieres, Spain; adrianlalobo@gmail.com

[2] Department of Biological Sciences, Mississippi State University, MS 39762, USA; gary.ervin@msstate.edu

ABSTRACT

Background and Objective: Large soil databases built with prediction methods can impose multiple limitations associated with prediction uncertainty. The objective of this study was to evaluate the accuracy of predicting soil properties with a polygon-based prediction approach and to assess the influence of the edaphic and topographic landscape context on that prediction accuracy. **Materials and Methods:** Ground-verified soil data (sand, silt, and clay, organic matter, and pH) were collected from 443 sample sites throughout Mississippi, and GIS predicted soil data were downloaded from the SSURGO database. The influence of the surrounding landscape at different spatial scales (0-300 m and 0-3000 m) on the absolute differences between ground-verified and GIS predicted values was evaluated using generalized linear models (GLMs). **Results:** Landscapes with high variability in the evaluated edaphic attributes showed higher differences between ground-verified and GIS predicted data, which suggested that the prediction accuracy of soil properties with GIS techniques decreases in landscapes with more variable edaphic attributes However, differences between ground-verified and GIS predicted data were generally lower in landscapes where edaphic and topographic data were spatially more heterogeneous. This could be the result of there being greater samples taken to develop the SSURGO database from areas with more heterogeneous soils or topography. Furthermore, differences between ground-verified and GIS predicted soil data were higher when the sample sites were nearer transportation routes and/or utility ROWs. **Conclusion:** Results showed that the surrounding land use and edaphic and topographic landscape highly influence the prediction accuracy of soil attributes with GIS techniques.

INTRODUCTION

Mapping soil properties at large scales can be challenging because soil attributes change continuously in space and time[1,2]. However, there is an increasing demand for high-resolution spatial soil maps from land managers, farmers, and researchers to refine management practices and improve research[2,3]. Therefore, multiple entities and initiatives, such as the United States Department of Agriculture–Natural Resources Conservation Service (USDA-NRCS) or the GlobalSoilMap, have developed digital soil maps based on geographic information systems (GIS) data layers at large spatial scales. Such databases are generated by sampling representative spatial points and then predicting the values of the remaining area with different techniques, including spatial interpolation[4,5] and predictions from soil polygons[4,6]. The use of soil polygons across large spatial scales to predict soil data from a limited number of representative sample points is a common approach used by the USDA-NRCS[3].

Although the use of large soil databases built with prediction methods can be beneficial for research on ecological patterns, it can impose multiple limitations associated with prediction uncertainty. For example, deterministic models used to predict soil property values involve some inherent degree of uncertainty, because they cannot capture the full extent of soil variation[2]. Likewise, considering the high spatial and temporal variation of soil properties[1,2], another important source of uncertainty comes from the generation of polygon map units with a single value per soil property, or the prediction of soil values with different interpolation techniques[3].

To the best of our knowledge, an evaluation of the possible influence of different landscape variables on data prediction accuracy of soil databases is lacking in the literature. In this regard, it may be that areas with certain landscape characteristics will negatively affect the prediction accuracy of resulting ecological models. For example, it may be more likely that the GIS predicted data differ from ground-verified data if the area around the sample sites has a high diversity of soil types or high variability of a given edaphic attribute. Also, local topography or topographic heterogeneity could cause differences between ground-verified data and GIS predicted data. Lastly, road construction and maintenance activities move large amounts of soil, which could increase the spatial and temporal variation of soil properties in nearby areas. Therefore, proximity to roads could be an important factor affecting the prediction accuracy of GIS models.

The objectives of this study were a) to evaluate the accuracy of predicting soil properties with a polygon-based prediction approach (GIS predicted data), b) to assess the influence of the edaphic and topographic landscape context on that prediction accuracy, and c) to examine possible road effects on prediction accuracy of soil properties. For the first objective, ground-verified data collected in multiple sample sites was compared with GIS predicted data generated by the US Soil Survey Geographic (SSURGO) database. For the second objective, different edaphic and topographic properties were analyzed within two spatial scales around the sample sites. The prediction for this second objective was that landscapes with high edaphic and topographic heterogeneity and variability would have lower prediction accuracy of soil properties than landscapes with the opposite characteristics. The prediction for the third objective was that sample sites near roads would have higher absolute differences between ground-verified data and GIS predicted data than sample sites far from roads.

MATERIALS AND METHODS

The study area for ground soil data collection: Soil data collection was carried out in 443 sample sites throughout Mississippi (USA), spanning an area of 125,430 km^2, from May 2006 to September 2009. The sample sites were randomly distributed throughout multiple soil association units and geographic regions present in the state, where permission to collect soil samples could be readily obtained.

Edaphic and topographic attributes: At each sample site, soil samples from the upper 10 to 20 cm of the soil (below the unconsolidated organic layer, where present) were collected to analyze soil particle size composition (sand, silt, and clay; measured in percentage), organic matter (percent by weight), and pH (ranges from 0 to 14) in the laboratory. Soil particle size analysis was performed by the hydrometer method, with corrections for temperature, as needed[7,8], at Biological Sciences Department, Plant Ecology Lab, Mississippi State University, USA from May 2006 to September 2009. Analyses of pH and organic matter were carried out at Department of Plant and Soil Sciences, Soil Testing Lab, Mississippi State University, USA from May 2006 to September 2009. Soil pH was measured in water in a 1:2 soil: water slurry and organic matter was determined by the loss of mass after ignition[9].

Predicted soil data based on geographic information systems (GIS) data layers were downloaded from the SSURGO database, which contains multiple soil attributes collected by the

National Cooperative Soil Survey over a century at scales ranging from 1:12,000 to 1:63,360. This large database is divided into polygon map units, which include soils and other components that have unique properties, interpretations, and productivity. Those polygon map units are based on tacit soil-landscape and slope models, and their soil property values have been assigned from a low number of representative pedons sampled within them[3]. Soil data were downloaded on October 23rd, 2019. Then the downloaded shapefile layers were rasterized to a 30-m resolution grid and selected the soil attributes of interest (percent of sand, silt, and clay; organic matter (percent by weight); and pH). Lastly, the blank rows of the raster layers were reclassified as "NoData" values.

Topographic variables were downloaded from the U.S. Geological Survey (www.nationalmap.gov). This database contains digital elevation data available at 100-m resolution from 2013. Using the bilinear resample technique, the elevation raster layer was resampled to a grid of 30 m pixel size. This technique calculates the value of each pixel by averaging (weighted for distance) the values of the four nearest pixels[10]. Then, the aspect and slope were calculated from the elevation map using ArcGIS 10.7.1.

Landscape data collection: Two spatial scales were considered by creating buffers around the sample sites at medium (0-300 m) and long (0-3000 m) distances using "rgeos" package of the program R. Those two spatial scales were chosen to include ~10 and ~100 raster cells of 30-m resolution in every direction from the center of the sample site (0-300 m and 0-3000 m buffers, respectively). Within each of those spatial scales and for each soil and topographic variable, different metrics were evaluated: patch density (number of patches in the landscape, divided by total landscape area), patch heterogeneity (number of different values that are included in the corresponding buffer), range (subtraction of the lowest value from the highest value) and variance (average squared deviation from the mean) of independent values. Distance from the sample sites to the nearest transportation route (e.g., highways, county roads, and streets) and/or utility right-of-way (ROW; e.g., pipelines and transmission lines) using ArcGIS 10.7.1.

Statistical analyses: The R program was used to conduct all the statistical analyses for this study. First, the topographic and predicted edaphic information was extracted from the raster cells of the different GIS layers. Second, the absolute difference between the values obtained from ground-verified data and GIS predicted data was calculated for each edaphic variable (sand, clay, silt,

organic matter, and pH). Lastly, the influence of the surrounding landscape at different spatial scales on the absolute differences between ground-verified and GIS values was evaluated using generalized linear models (GLMs).

Before conducting the GLMs, the values of the predictor variables were standardized using the function "scale" of the program R, to make the interpretation of their effect sizes more comparable. GLMs with gamma errors were used because the variables were distributed with different degrees of positive skewness and, therefore, the residuals of the models did not follow a normal distribution[11]. In addition, lower AIC (Akaike information criterion) values were obtained using the gamma family, which makes the GLMs with gamma errors more parsimonious. To avoid collinearity problems, only non-correlated variables were included in the GLMs. Heterogeneity and patch density of the different edaphic variables were highly correlated in both spatial landscape scales (Pearson´s $r > 0.4$). The same applied to the range and variance of the evaluated edaphic variables (Pearson´s $r > 0.64$). Heterogeneity and range also were highly correlated in some edaphic variables within the different spatial scales. Therefore, heterogeneity and variance of the edaphic variables were included to build the models. As for topographic variables, only slope heterogeneity of the topographic profile was included in the models because a) elevation, slope, and aspect variables were highly correlated (Pearson´s $r > 0.8$), b) models including slope had lower AIC than those including elevation and aspect, and c) slope heterogeneity and variance in the landscape was highly correlated (Pearson´s $r > 0.95$). Lastly, distance to linear infrastructures (transportation systems and utility ROWs) was included in the models to evaluate the influence of road construction and maintenance on differences between ground-verified and GIS predicted values. Thus, the following equation was used to build the GLMs:

$$|\textit{Difference ground vs. GIS data values}|_{(a)} \sim \textit{Edaphic heterogeneity}_{(a,\,c)} + \textit{Edaphic variance}_{(a,\,c)} + \textit{Topographic heterogeneity}_{(b,\,c)} + \textit{Distance to linear infrastructures}$$

where "a" is the corresponding soil variable (sand, clay, silt, organic matter, or pH), "b" is the slope of the topographic profile, and "c" is the spatial scale (buffer size of 0-300 m, or 0-3000 m).

RESULTS

There was high variability in absolute differences between ground-verified data and GIS predicted data values (Fig. 1a-e). In some instances, the soil values obtained from the field-

collected samples were very different from those in the SSURGO database. For example, sand values differed ~ 40 % between ground verified data and GIS predicted data in 35 % of the sample sites (Fig. 1a). On the other hand, there also were some edaphic values that were very similar in both databases, as indicated by the near-zero first quartile values (Fig. 1b, d, and e).

The variance of the edaphic variables obtained from GIS generally had a positive relationship with the absolute differences between our ground-verified data and the GIS predicted data values (Table 1). That is a higher variance among grid cell values in the SSURGO database generally correlated with greater disagreement between our field-collected data and the SSURGO data.

In contrast, landscape heterogeneity (i.e., spatial patchiness in the GIS data) in edaphic and topographic predictor variables was generally negatively correlated with absolute differences between our ground-verified data and the GIS predicted data values. For the edaphic GIS data, this pattern was more prevalent at the larger spatial scale (0-3000 m). However, the effect size of edaphic variance, mentioned above, was generally stronger than that found for edaphic and topographic heterogeneity.

Furthermore, the absolute differences between ground-verified data and GIS predicted data in sand and pH were lower when the sample sites were farther from the nearest transportation route and/or utility ROW (Table 1). Around half of the sample sites (n = 206) were within 50 m of the nearest linear infrastructure (transportation, pipelines, electricity transmission lines), making them susceptible to road construction and other ROW maintenance activities.

DISCUSSION

The results from this study demonstrate that ground soil data collected at a given spatial sample point can be very different from estimated soil properties produced by large databases as SSURGO, which collect soil data at scales ranging from 1:12,000 to 1:63,360 and use a polygon-based prediction approach to generate polygon map units[3]. Therefore, while the production of soil maps at very large spatial scales enables more research opportunities, it could also be prejudicial for data accuracy because soil attributes such as soil texture, organic matter, and pH can be highly variable at small spatial scales[1,2].

The surrounding edaphic and topographic landscape context were strongly correlated with the observed differences between ground-verified data and GIS predicted data at different spatial scales. Generally, landscapes with high variability in the evaluated edaphic attributes showed higher differences between ground and GIS data, as was initially predicted. Thus, the higher the variability of a soil variable in the landscape, the higher the chances to have a mismatch between field-collected data and data from GIS databases, such as SSURGO.

Contrary to the initial expectations, there were generally lower differences between our field-collected data and SSURGO data in landscapes where GIS-based soil and topographic data were spatially more heterogeneous. This could be the result of there being greater samples taken to develop the SSURGO database from areas with more heterogeneous soils or topography. The SSURGO database is divided into polygon map units, which include soils and other components that have unique properties, interpretations, and productivity[3]. Thus, heterogeneous landscapes are more likely to be represented by a greater number of polygon map units, which should correspond with an increased number of soil samples taken in the area.

Along these lines, it is expected that areas with greater topographic heterogeneity generally will have steeper or more heterogeneous slopes. Previous studies aimed at evaluating the spatial relationships between topography and soil attributes have found significant relationships between topography and soil characteristics[12,13,14]. For example, a study examined soil-terrain relationships within a GIS framework and found that variance in the thickness of both the A horizon and loess layers tended to be substantially higher (more than double) in areas with lower slopes, such as level ground or foot slopes of rolling terrain[15]. Such study also found that relationships among soil and topographic variables were highly dependent upon the pixel size of the GIS window used to calculate parameters such as topographic curvature, with the strongest correlations found at the smallest window sizes (3x3 grid cells). In other words, the study found that, as the area represented by each discrete sample increased, the relationships among parameters being estimated tended to weaken. This latter finding again suggests that areas with greater sampling density in the development of digital soil databases ought to have greater correspondence, on average, with randomly collected field data.

In our study, proximity to roads also had a great effect on the absolute differences between ground-verified and GIS predicted soil pH and sand content data. Such differences were generally

higher when the sample sites were nearer transportation routes and/or utility ROWs. This finding could suggest that road construction and maintenance activities increase the spatial variation of soil properties due to soil disturbance, soil movement, or import of large amounts of soil[16,17], which, in turn, decrease the prediction accuracy of soil properties using GIS polygon map units, such as those provided by the SSURGO database.

CONCLUSION

This study shows that the surrounding land use and edaphic and topographic landscape highly influence the degree of similarity between ground-verified data and GIS predicted data for commonly used soil variables (particle size composition, organic matter content, and pH). The prediction accuracy of soil properties with GIS techniques decreased in landscapes with more variable edaphic attributes, whereas accuracy unexpectedly increased in more topographically and edaphically heterogeneous landscapes. These findings may have important implications for models that incorporate georeferenced soil data. Thus, this study will improve research conducted at multiple spatial scales.

ACKNOWLEDGMENTS

The authors thank Dr. Alison Paulson for writing the script with guidelines to collect data from the SSURGO database. The authors also are grateful to the following, who contributed to the collection of soil samples and data used in these analyses: Chris Doffitt, Chris Holly, Steven Hughes, Lucas Majure, Rima Lucardi, Taylor Sawyer, and Nathan Sonderman. This work was supported by grants from the US Department of Agriculture (2006-03613, 2008-35320-18679, and 19-DG11083150-006) and the US Geological Survey (04HQAG0135 and 08HQAG0139) to GNE.

REFERENCES

1. Cambardella, C.A., T.B. Moorman, T.B. Parkin, D.L. Karlen, J.M. Novak, R.F. Turco and A.E. Konopka, 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci. Soc. Am. J., 58: 1501–1511. DOI: 10.2136/sssaj1994.03615995005800050033x.
2. Heuvelink, G.B.M. and R. Webster, 2001. Modelling soil variation: past, present, and future. Geoderma, 100: 269–301. DOI: https://doi.org/10.1016/S0016-7061(01)00025-8.
3. Libohova, Z., C. Seybold, K. Adhikari, S. Wills, D. Beaudette, S. Peaslee, D. Lindbo and P.R. Owens, 2019. The anatomy of uncertainty for soil pH measurements and predictions:

Implications for modellers and practitioners. Eur. J. Soil Sci., 70: 185–199. DOI: https://doi.org/10.1111/ejss.12770.

4. McBratney, A.B., M.L. Mendonca Santos and B. Minasny, 2003. On digital soil mapping. Geoderma, 117: 3–52. DOI: https://doi.org/10.1016/S0016-7061(03)00223-4.
5. Robinson, T.P. and G. Metternicht, 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Comput. Electron. Agric., 50: 97–108. DOI: https://doi.org/10.1016/j.compag.2005.07.003.
6. Webster, L. and M.A. Oliver, 2007. Geostatistics for Environmental Scientists, 2nd edition. Hoboken, NJ: John Wiley & Sons, Ltd. DOI: https://doi.org/10.1002/9780470517277.
7. Ashworth, J., D. Keyes, R. Kirk and R. Lessard, 2001. Standard procedure in the hydrometer method for particle size analysis. Commun. Soil Sci. Plant Anal., 32: 633–642. DOI: https://doi.org/10.1081/CSS-100103897.
8. Scott, H.D. 2000. Soil Physics: Agricultural and Environmental Applications. Ames, IA: Iowa State University Press. DOI: https://doi.org/10.1097/00010694-200110000-00007.
9. Roper, W.R., W.P. Robarge, D.L. Osmond and J.L. Heitman, 2019. Comparing four methods of measuring soil organic matter in North Carolina soils. Soil Sci. Soc. Am. J., 83: 466–474. DOI: 10.2136/sssaj2018.03.0105
10. Lázaro-Lobo, A., K.O. Evans and G.N. Ervin, 2020. Evaluating landscape characteristics of predicted hotspots for plant invasions. Invasive Plant Sci. Manag., 13: 163–175. DOI: https://doi.org/10.1017/inp.2020.21.
11. Ng, V.K. and R.A. Cribbie, 2017. Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Curr. Psychol., 36: 225–235. DOI: 10.1007/s12144-015-9404-0.
12. Li, X., G.W. McCarty, L. Du and S. Lee, 2020. Use of topographic models for mapping soil properties and processes. Soil Syst., 4: 32. DOI: https://doi.org/10.3390/soilsystems4020032
13. Alijani, Z. and F. Sarmadian, 2014. The role of topography in changing of soil carbonate content. India J. Sci. Res., 6: 263–271. URL: https://www.researchgate.net/profile/Zohreh-Alijani/publication/280237568_THE_ROLE_OF_TOPOGRAPHY_IN_CHANGING_OF_SOIL_CARBONATE_CONTENT/links/55ae96c608ae98e661a6eceb/THE-ROLE-OF-TOPOGRAPHY-IN-CHANGING-OF-SOIL-CARBONATE-CONTENT.pdf

14. Florinsky, I.V., S. McMahon and D.L. Burton, 2004. Topographic control of soil microbial activity: A case study of denitrifiers. Geoderma, 119: 33–53. DOI: https://doi.org/10.1016/S0016-7061(03)00224-6
15. Park, S.J., K. McSweeney and B. Lowery, 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma, 103: 249–272. DOI: https://doi.org/10.1016/S0016-7061(01)00042-8.
16. Deljouei, A., S.M.M. Sadeghi, E. Abdi, M. Bernhardt-Römermann, E. L. Pascoe and M. Marcantonio, 2018. The impact of road disturbance on vegetation and soil properties in a beech stand, Hyrcanian forest. Eur. J. For. Res., 137: 759-770. DOI: https://doi.org/10.1007/s10342-018-1138-8.
17. Hacisalihoğlu, S., S. Gümüş, U. Kezik and H. Karadag, 2019. Impact of forest road construction on topsoil erosion and hydro-physical soil properties in a semi-arid mountainous ecosystem in Turkey. Pol. J. Environ. Stud., 28: 113-121. DOI: 10.15244/pjoes/81615

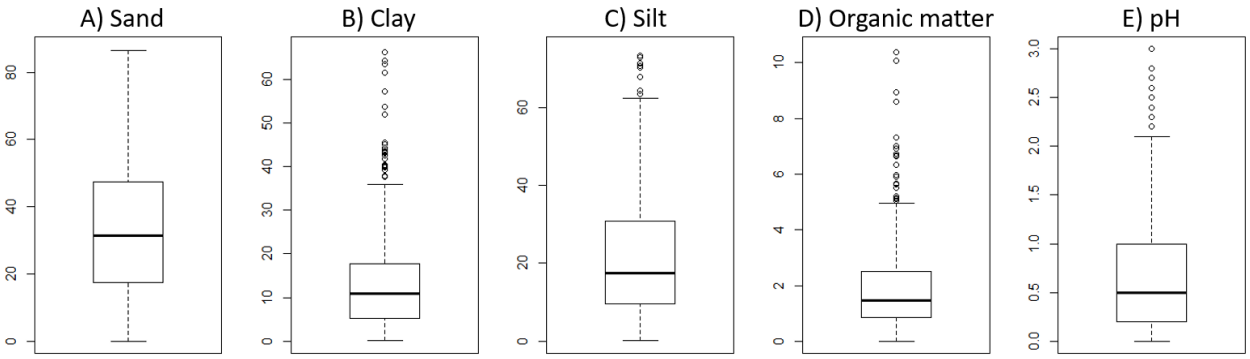


FIGURE 1a-e. Boxplots indicating the absolute differences between ground-verified data and GIS predicted data values for each edaphic variable.

Table 1. Generalized linear model outputs showing the correlation of the surrounding landscape on absolute differences between ground-verified data and GIS predicted data values.

Spatial scale	GIS-based landscape variables	Model coefficients	Soil attributes				
			Sand	Clay	Silt	Organic matter	pH
0-300 m	Soil attribute heterogeneity	Estimate	-0.02	-0.14	+0.05	-0.03	-0.02
		Std. Error	0.03	0.04	0.04	0.04	0.06
		P-value	0.53	0.001	0.16	0.56	0.75
	Soil attribute variance	Estimate	-0.03	+0.26	+0.09	-0.08	+0.09
		Std. Error	0.03	0.05	0.04	0.04	0.05
		P-value	0.32	<0.001	0.02	0.06	0.04
	Topographic heterogeneity	Estimate	-0.10	-0.05	-0.16	-0.07	+0.06
		Std. Error	0.03	0.04	0.04	0.04	0.04
		P-value	0.001	0.23	<0.001	0.12	0.18
	Distance to transportation systems and utility ROWs	Estimate	-0.06	+0.01	-0.07	+0.02	-0.26
		Std. Error	0.03	0.04	0.04	0.04	0.048
		P-value	0.03	0.80	0.07	0.63	<0.001
0-3000 m	Soil attribute heterogeneity	Estimate	-0.10	-0.09	+0.002	-0.04	-0.11
		Std. Error	0.03	0.04	0.04	0.04	0.05
		P-value	<0.001	0.02	0.95	0.28	0.01
	Soil attribute variance	Estimate	+0.03	+0.28	+0.18	-0.18	+0.15
		Std. Error	0.03	0.05	0.04	0.04	0.05
		P-value	0.33	<0.001	<0.001	<0.001	0.001
	Topographic heterogeneity	Estimate	-0.08	-0.02	-0.11	-0.17	+0.12
		Std. Error	0.03	0.04	0.04	0.04	0.04
		P-value	0.006	0.72	0.004	<0.001	0.002
	Distance to transportation systems and utility ROWs	Estimate	-0.04	-0.003	-0.05	+0.03	-0.24
		Std. Error	0.03	0.04	0.03	0.04	0.04
		P-value	0.14	0.95	0.15	0.47	<0.001

Footnote: Positive and negative estimates with p-value < 0.05 are shaded light and dark grey, respectively. Each soil attribute was tested with edaphic landscape data of the corresponding attribute, topographic landscape data, and distance to transportation systems and utility ROWs, at two spatial scales (0-300 and 0-3000 m). Topographic heterogeneity refers to the slope heterogeneity of the topographic profile.