A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review

Luca Patelli, Michela Cameletti, Natalia Golini, Rosaria Ignaccolo

Taxonomy Category	Strategy	Reference	Title	From
Pre	RF with SI	Behrens et al. (2018)	*Spatial modelling with Euclidean distance fields and machine learning*	Q
		Dhara et al. (2018)	*Machine learning based methods for estimation and stochastic simulation*	Q
		Hengl et al. (2018)	*Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables*	S
		Møller et al. (2020)	*Oblique geographic coordinates as covariates for digital soil mapping*	S
		Sekulić et al. (2020)	*Random forest spatial interpolation*	S
		Còrdoba & Balzarini (2021)	*A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping*	S
		Hu et al. (2022)	*Incorporating spatial autocorrelation into house sale price prediction using random forest model*	Q
		Santiago-Rosario et al. (2022)	*Contrasts among cationic phytochemical landscapes in the southern United States*	Q
		Talebi et al. (2022)	*A truly spatial random forests algorithm for geoscience data analysis and modelling*	Q
	RF with FFS	Meyer et al. (2019)	*Importance of spatial predictor variable selection in machine learning applications*	S
Post	RF-RK	Guo et al. (2015)	*Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach*	S
		Hengl et al. (2015)	*Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions*	S
		Fayad et al. (2016)	*Regional scale rain-forest height mapping using regression-kriging of spaceborne and airborne LiDAR data: application on French Guiana*	S
		Ahmed et al. (2017)	*Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks*	S
		Garcìa-Tomillo et al. (2017)	*Estimating soil organic matter using interpolation methods with a electromagnetic induction sensor and topographic parameters: a case study in a humid region*	Q
		Vaysse & Lagacherie (2017)	*Using quantile regression forest to estimate uncertainty of digital soil mapping products*	S
		dos Reis et al. (2018)	*Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods*	Q
		Liu et al. (2018)	*Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach*	Q
		Fox et al. (2020)	*Comparing spatial regression to random forests for large environmental data sets*	S
		Còrdoba et al. (2021)	*A spatially based quantile regression forest model for mapping rural land values*	Q
		Paccioretti et al. (2021)	*Statistical models of yield in on-farm precision experimentation*	Q
		Makungwe et al. (2021)	*Performance of linear mixed models and random forests for spatial prediction of soil Ph*	Q
		Mammadov et al. (2021)	*Estimation and mapping of surface soil properties in the Caucasus Mountains, Azerbaijan using high-resolution remote sensing data*	Q
		Szatmàri et al. (2021)	*Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics*	Q
		da Silva-Sangoi et al. (2022)	*Soil organic matter and clay predictions by laboratory spectroscopy: data spatial correlation*	Q
		Smith et al. (2022)	*Spatial variability and uncertainty of soil nitrogen across the conterminous United States at different depths*	Q
	RF-sGs	Koch et al. (2019)	*Modeling depth of the redox interface at high resolution at national scale using random forest and residual Gaussian simulation*	Q
In-Post	RF-GLS-RK	Saha et al. (2021)	*Random forests for spatially dependent data*	Q
Pre-Post	RF-RK with SI	Li et al. (2011)	*Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation*	S
		Kurina et al. (2019)	*Spatial predictive modelling essential to assess the environmental impacts of herbicides*	Q
	RF-RK with SB	Viscarra Rossel et al. (2014)	*Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging*	S
		Szatmàri & Pàsztor (2019)	*Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms*	S

Table 1 List of the 32 documents identified through the PRISMA methodology and categorised by the proposed taxonomy.
Adopted strategy: RF with SI: Random Forest with Spatial Information; RF with FFS: Random Forest with Forward Feature Selection; RF-RK: Random Forest Residual Kriging; RF-sGs: Random Forest sequential Gaussian simulation; RF-RK with SI: Random Forest Residual Kriging with Spatial Information; RF-RK with SB: Random Forest Residual Kriging with Spatial Bootstrap; RF-GLS-RK: Random Forest based on GLS Residual Kriging.
From: Q: query; S: backward snowballing.

References

Ahmed, Z. U., Woodbury, P. B., Sanderman, J., Hawke, B., Jauss, V., Solomon, D. & Lehmann, J. (2017), 'Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks', *Journal of Geophysical Research: Biogeosciences* **122**(2), 354–369.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T. & MacMillan, R. A. (2018), 'Spatial modelling with Euclidean distance fields and machine learning', *European journal of soil science* **69**(5), 757–770.

Còrdoba, M. & Balzarini, M. (2021), 'A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping', *Computers and Electronics in Agriculture* **184**, 106094.

Còrdoba, M., Carranza, J. P., Piumetto, M., Monzani, F. & Balzarini, M. (2021), 'A spatially based quantile regression forest model for mapping rural land values', *Journal of Environmental Management* **289**, 112509.

da Silva-Sangoi, D. V., Horst, T. Z., Moura-Bueno, J. M., Dalmolin, R. S. D., Sebem, E., Gebler, L. & da Silva Santos, M. (2022), 'Soil organic matter and clay predictions by laboratory spectroscopy: Data spatial correlation', *Geoderma Regional* **28**, e00486.

Dhara, A., Trainor-Guitton, W. & Tura, A. (2018), Machine-learning-based methods for estimation and stochastic simulation, *in* 'SEG Technical Program Expanded Abstracts 2018', pp. 2261–2265.

dos Reis, A. A., Carvalho, M. C., de Mello, J. M., Gomide, L. R., Ferraz Filho, A. C. & Acerbi Junior, F. W. (2018), 'Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods', *New Zealand Journal of Forestry Science* **48**(1), 1–17.

Fayad, I., Baghdadi, N., Bailly, J.-S., Barbier, N., Gond, V., Hèrault, B., El Hajj, M., Fabre, F. & Perrin, J. (2016), 'Regional scale rain-forest height mapping using regression-kriging of spaceborne and airborne LiDAR data: Application on French Guiana', *Remote Sensing* **8**(3), 240.

Fox, E. W., Ver Hoef, J. M. & Olsen, A. R. (2020), 'Comparing spatial regression to random forests for large environmental data sets', *PloS one* **15**(3), e0229509.

Garcìa-Tomillo, A., Miràs-Avalos, J. M., Dafonte-Dafonte, J. & Paz-Gonzàlez, A. (2017), 'Estimating soil organic matter using interpolation methods with a electromagnetic induction sensor and topographic parameters: a case study in a humid region', *Precision Agriculture* **18**(5), 882–897.

Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W. & Lin, Z.-M. (2015), 'Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach', *Geoderma* **237**, 49–59.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L. et al. (2015), 'Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions', *PloS one* **10**(6), e0125814.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. & Gräler, B. (2018), 'Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables', *PeerJ* **6:e5518**.

Hu, L., Chun, Y. & Griffith, D. A. (2022), 'Incorporating spatial autocorrelation into house sale price prediction using random forest model', *Transactions in GIS* **26**(5), 2123–2144.

Koch, J., Stisen, S., Refsgaard, J. C., Ernstsen, V., Jakobsen, P. R. & Højberg, A. L. (2019), 'Modeling depth of the redox interface at high resolution at national scale using random forest and residual gaussian simulation', *Water Resources Research* **55**(2), 1451–1469.

Kurina, F. G., Hang, S., Macchiavelli, R. & Balzarini, M. (2019), 'Spatial predictive modelling essential to assess the environmental impacts of herbicides', *Geoderma* **354**, 113874.

Li, J., Heap, A. D., Potter, A., Huang, Z. & Daniell, J. J. (2011), 'Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin', *Continental Shelf Research* **31**(13), 1365–1376.

Liu, Y., Cao, G., Zhao, N., Mulligan, K. & Ye, X. (2018), 'Improve ground-level PM2.5 concentration mapping using a random forests-based geostatistical approach', *Environmental Pollution* **235**, 272–282.

Makungwe, M., Chabala, L. M., Chishala, B. H. & Lark, R. M. (2021), 'Performance of linear mixed models and random forests for spatial prediction of soil pH', *Geoderma* **397**, 115079.

Mammadov, E., Nowosad, J. & Glaesser, C. (2021), 'Estimation and mapping of surface soil properties in the Caucasus Mountains, Azerbaijan using high-resolution remote sensing data', *Geoderma Regional* **26**, e00411.

Meyer, H., Reudenbach, C., Wöllauer, S. & Nauss, T. (2019), 'Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction', *Ecological Modelling* **411**, 108815.

Møller, A. B., Beucher, A. M., Pouladi, N. & Greve, M. H. (2020), 'Oblique geographic coordinates as covariates for digital soil mapping', *Soil* **6**(2), 269–289.

Paccioretti, P., Bruno, C., Gianinni Kurina, F., Còrdoba, M., Bullock, D. & Balzarini, M. (2021), 'Statistical models of yield in on-farm precision experimentation', *Agronomy Journal* **113**(6), 4916–4929.

Saha, A., Basu, S. & Datta, A. (2021), 'Random forests for spatially dependent data', *Journal of the American Statistical Association* pp. 1–19.

Santiago-Rosario, L. Y., Harms, K. E. & Craven, D. (2022), 'Contrasts among cationic phytochemical landscapes in the southern United States', *Plant-Environment Interactions* **3**(5), 226–241.

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolič, M. & Bajat, B. (2020), 'Random forest spatial interpolation', *Remote Sensing* **12**(10), 1687.

Smith, E. M., Vargas, R., Guevara, M., Tarin, T. & Pouyat, R. V. (2022), 'Spatial variability and uncertainty of soil nitrogen across the conterminous United States at different depths', *Ecosphere* **13**(7), e4170.

Szatmàri, G. & Pàsztor, L. (2019), 'Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms', *Geoderma* **337**, 1329–1340.

Szatmàri, G., Pàsztor, L. & Heuvelink, G. B. (2021), 'Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics', *Geoderma* **403**, 115356.

Talebi, H., Peeters, L. J., Otto, A. & Tolosana-Delgado, R. (2022), 'A truly spatial random forests algorithm for geoscience data analysis and modelling', *Mathematical Geosciences* **54**(1), 1–22.

Vaysse, K. & Lagacherie, P. (2017), 'Using quantile regression forest to estimate uncertainty of digital soil mapping products', *Geoderma* **291**, 55–64.

Viscarra Rossel, R. A., Webster, R. & Kidd, D. (2014), 'Mapping gamma radiation and its uncertainty from weathering products in a tasmanian landscape with a proximal sensor and random forest kriging', *Earth Surface Processes and Landforms* **39**(6), 735–748.