Mapping soil thickness by accounting for right-censored data with machine learning

Stephan van der Westhuizen[a,b,c,*], Gerard B. M. Heuvelink[b,c], David P. Hofmeyr[d], Laura Poggio[c], Madlene Nussbaum[e], Colby Brungard[f]

[a] *Dept. of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa*
[b] *Soil Geography and Landscape group, Wageningen University, Wageningen, The Netherlands*
[c] *ISRIC-World Soil Information, Wageningen, The Netherlands*
[d] *Dept. of Mathematics and Statistics, Lancaster University, Lancaster, UK*
[e] *Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands*
[f] *Dept. of Plant and Environmental Sciences, New Mexico State University, Las Cruces, USA*

Appendix A. Synthetic simulation studies

Appendix A.1. More details on the methodology of the simulation study

In each experiment we simulated a synthetic data set on unit square discretised into a 100×100 grid comprising $N = 10\,000$ response values. First, we generated values, $z(\mathbf{s}_k)$, for $k = 1 \ldots N$, from a Gaussian process with a covariance structure governed by an isotropic spherical variogram model with a nugget of 0.1, a partial sill of 0.9, and a range of $a = 0.2$. The covariate values were calculated using

*Corresponding author
Email address: stephanvdw@sun.ac.za (Stephan van der Westhuizen)

$$x_1(\mathbf{s}_k) = \begin{cases} 2z(\mathbf{s}_k) + r(\mathbf{s}_k), & z(\mathbf{s}_k) \geq 0.5, \\ -2z(\mathbf{s}_k) + r(\mathbf{s}_k), & z(\mathbf{s}_k) < 0.5, \end{cases} \tag{A.1}$$

$$x_2(\mathbf{s}_k) = -\frac{1}{2}z(\mathbf{s}_k) + r(\mathbf{s}_k), \tag{A.2}$$

$$x_3(\mathbf{s}_k) = x_1(\mathbf{s}_k) \cdot x_2(\mathbf{s}_k), \tag{A.3}$$

for $k = 1, \ldots, N$, and where the residual values $r(\mathbf{s}_k)$, which are i.i.d., were generated from a standard Gaussian random variable. This is similar to the simulation study performed in van der Westhuizen et al. (2022). We then transformed the values, $z(\mathbf{s}_k)$, to follow a log-normal distribution which would then represent soil thickness, $d(\mathbf{s}_k)$. The transformation was done with $e^{z(\mathbf{s}_k)\sigma_z + \mu_z}$, where

$$\mu_z = \ln\left(\frac{\mu_d^2}{\sqrt{\sigma_d^2 + \mu_d^2}}\right), \tag{A.4}$$

$$\sigma_z = \sqrt{\ln(1 + \frac{\sigma_d^2}{\mu_d^2})}. \tag{A.5}$$

The log-normal mean, μ_d, and standard deviation, σ_d, were set equal to 60 and 40, respectively. We used this standard approach to generate draws from a log-normal distribution to ensure control of the mean and variance of the log-normal variable. In addition, we simulated the synthetic data in this way to ensure that the simulation parameters which regulate the censoring scenarios, are directly comparable to the response values, $d(\mathbf{s}_k)$. In addition, the purpose of this analysis is to compare the various modelling approaches given a set of covariates, and for this reason it is satisfactory to just have a few covariates to enable us to simulate the censoring scenarios and to fit the various models. Similar to van der Westhuizen et al. (2022), the covariates in Eqs. (A.1), (A.2)

and (A.3), on average, describe about 40% of the variation in $d(\mathbf{s}_k)$. For the entire grid we therefore have a set values $d(\mathbf{s}_k)$, and values for the three covariates, $x_1(\mathbf{s}_k)$, $x_2(\mathbf{s}_k)$ and $x_3(\mathbf{s}_k)$. Finally, a sample of observations was obtained by randomly selecting either $n = 400, 800, 1\,600$ of the N grid values (in case of the non-informative censoring mechanism).

Appendix A.2. Simulation results for the mean error

Table A.1: ME results for the synthetic simulation study for both censoring scenarios. Results are shown for the non-informative mechanism, censoring proportion $\rho = \{0, 0.3, 0.6, 0.9, 1\}$, censoring depth $\delta = \{60, 90, 120\}$ and sample size $n = \{400, 800\}$. Model results include RF fitted on the true soil thickness, and ARF, HRF, SRF, weighted random forest (IPC-RF), fitted on censored soil thickness.

Parameters			$n = 400$					$n = 800$				
Scenario	λ	ρ	RF	ARF	HRF	SRF	IPC-RF	RF	ARF	HRF	SRF	IPC-RF
1	60.0	0.0	-0.3	-0.2	-0.2	-0.2	-0.4	-0.1	-0.2	-0.2	-0.7	-0.2
		0.1	-0.4	-1.8	-1.1	-0.5	-1.0	-0.1	-1.5	-0.8	-0.6	-0.6
		0.3	-0.3	-4.5	-2.7	0.0	-2.2	-0.2	-4.5	-2.6	0.0	-1.9
		0.6	-0.4	-8.7	-7.1	1.3	-6.1	-0.2	-8.4	-6.3	2.3	-5.3
		0.9	-0.4	-12.9	-16.4	8.3	-15.9	-0.2	-12.5	-15.0	12.1	-14.5
		1.0	-0.3	-14.5	-26.8	-14.5	-14.5	-0.2	-14.7	-27.8	-15.0	-14.7
	90	0.0	-0.3	-0.3	-0.3	-0.2	-0.5	-0.2	-0.2	-0.2	-1.0	-0.3
		0.1	-0.3	-0.9	-0.6	-0.8	-0.6	-0.1	-0.8	-0.4	-0.7	-0.4
		0.3	-0.2	-2.1	-1.3	-0.6	-1.1	-0.2	-2.1	-1.2	-0.8	-0.9
		0.6	-0.3	-4.1	-3.3	-0.3	-2.7	-0.2	-4.1	-2.8	-0.3	-2.3
		0.9	-0.1	-5.9	-8.2	1.6	-7.6	-0.2	-6.1	-7.6	2.2	-7.1
		1.0	-0.3	-6.9	-14.4	-7.2	-6.9	-0.2	-6.8	-15.3	-7.7	-6.8
	120	0.0	-0.3	-0.3	-0.3	0.0	-0.4	-0.2	-0.2	-0.2	-0.7	-0.3
		0.1	-0.4	-0.6	-0.5	-0.5	-0.6	-0.2	-0.5	-0.3	-1.0	-0.3
		0.3	-0.2	-1.1	-0.9	-0.9	-0.8	-0.2	-1.0	-0.6	-1.0	-0.5
		0.6	-0.2	-2.0	-1.9	-0.7	-1.5	-0.2	-2.0	-1.5	-0.7	-1.2
		0.9	-0.4	-3.1	-4.3	-0.9	-3.7	-0.2	-2.8	-3.3	0.0	-2.8
		1.0	-0.3	-3.3	-5.9	-3.8	-3.3	-0.3	-3.2	-5.7	-3.7	-3.2
2	60	0.0	-0.2	-0.3	-0.2	-0.2	-0.4	-0.1	-0.2	-0.2	-0.7	-0.2
		0.1	-0.4	-0.9	-1.1	-0.3	-1.0	-0.1	-0.6	-0.8	-0.3	-0.7
		0.3	-0.3	-1.7	-2.6	1.9	-2.2	-0.2	-1.6	-2.4	1.7	-2.0
		0.6	-0.5	-3.2	-6.6	7.3	-5.8	-0.2	-2.8	-5.8	7.7	-5.1
		0.9	-0.4	-4.5	-14.3	17.2	-10.7	-0.2	-4.2	-13.0	19.3	-10.6
		1.0	-0.3	-4.8	-19.8	6.1	-13.0	-0.3	-4.9	-19.4	7.8	-13.5
	90	0.0	-0.3	-0.3	-0.3	-0.1	-0.4	-0.2	-0.2	-0.2	-1.0	-0.3
		0.1	-0.3	-0.5	-0.6	-0.5	-0.6	-0.1	-0.4	-0.4	-0.6	-0.4
		0.3	-0.1	-0.9	-1.3	0.5	-1.2	-0.2	-0.9	-1.1	0.3	-1.0
		0.6	-0.3	-1.5	-3.2	3.0	-2.7	-0.2	-1.5	-2.6	3.2	-2.3
		0.9	-0.1	-2.1	-7.4	5.4	-4.6	-0.2	-2.1	-6.4	7.1	-4.6
		1.0	-0.3	-2.5	-10.8	0.2	-6.2	-0.2	-2.4	-9.8	1.7	-5.9
	120	0.0	-0.3	-0.3	-0.3	0.0	-0.4	-0.2	-0.2	-0.2	-0.7	-0.3
		0.1	-0.3	-0.5	-0.6	-0.6	-0.6	-0.2	-0.3	-0.3	-0.8	-0.3
		0.3	-0.2	-0.6	-0.9	-0.3	-0.8	-0.2	-0.5	-0.7	-0.5	-0.5
		0.6	-0.2	-0.9	-1.9	0.6	-1.5	-0.2	-0.8	-1.5	0.7	-1.2
		0.9	-0.4	-1.4	-4.2	0.2	-2.3	-0.2	-1.0	-3.0	1.7	-1.6
		1.0	-0.3	-1.3	-5.1	-1.5	-2.5	-0.3	-1.3	-4.6	-0.7	-2.2

Appendix A.3. Informative censoring mechanism results

The mean square error (RMSE) results for the censoring scenarios (constant versus non-constant) for the informative censoring mechanism are presented in Table A.2. The mechanism is informative as the selection of observations to be censored were inversely proportional to the values of $x_1(\mathbf{s}_k)$. In this table we also present the results for a random forest (RF) model that was fitted with the true soil thickness which can be used as a baseline for comparing the results of the rest of the models. Prediction results are also highlighted in terms of the mean error (ME), with blue indicating underestimation and red indicating overestimation (a darker shade represents a larger value for the ME).

Table A.2: RMSE results for the synthetic simulation study for both censoring scenarios (constant versus non-constant censoring depths) and for the informative censoring mechanism. The results are shown for the censoring proportion $\rho = \{0, 0.3, 0.6, 0.9, 1\}$, censoring depth $\delta = \{60, 90, 120\}$ and sample size, $n = \{400, 800\}$. Model results include RF fitted on the true soil thickness, and random forest with all data (ARF), random forest with only hard data (HRF), random survival forest (SRF), IPC-RF, fitted on censored soil thickness. Model results are also highlighted in terms of bias, i.e., ME, with blue indicating underestimation and red indicating overestimation.

Parameters			$n = 400$					$n = 800$				
Scenario	λ	ρ	RF	ARF	HRF	SRF	IPC-RF	RF	ARF	HRF	SRF	IPC-RF
1	60	0.0	23.8	23.8	23.8	26.2	23.5	22.3	22.3	22.3	23.8	22.1
	60	0.1	24.0	24.3	24.2	26.2	23.8	22.3	22.8	22.6	24.1	22.4
	60	0.3	24.0	25.1	24.3	26.7	23.9	22.3	23.6	22.8	24.5	22.6
	60	0.6	24.0	28.1	26.0	29.8	25.2	22.3	27.0	24.5	31.5	24.1
	60	0.9	24.0	34.5	32.8	43.6	32.3	22.3	33.2	30.3	49.2	30.3
	60	1.0	24.0	38.9	51.2	38.9	38.9	22.3	38.3	51.8	39.5	38.3
	90	0.0	24.0	23.9	23.8	26.4	23.5	22.3	22.4	22.4	24.1	22.3
	90	0.1	24.0	24.2	24.1	26.5	23.7	22.3	22.5	22.4	23.9	22.2
	90	0.3	24.0	24.9	24.5	26.3	23.9	22.3	23.5	22.9	24.1	22.6
	90	0.6	24.0	26.4	25.7	26.6	24.7	22.3	25.6	24.0	24.4	23.5
	90	0.9	24.0	29.6	30.3	29.1	29.8	22.3	28.8	28.0	29.9	28.1
	90	1.0	24.0	31.9	42.2	31.9	32.0	22.3	30.9	43.7	32.7	30.9
	120	0.0	24.0	23.9	23.9	26.2	23.6	22.3	22.3	22.3	23.9	22.1
	120	0.1	24.0	24.1	24.1	26.3	23.7	22.3	22.6	22.5	24.2	22.3
	120	0.3	24.0	24.3	24.3	25.9	23.8	22.3	23.1	22.8	24.2	22.4
	120	0.6	24.0	25.6	25.4	26.3	24.5	22.3	24.3	23.6	24.4	22.9
	120	0.9	24.0	27.2	28.0	26.7	27.1	22.3	25.9	25.6	25.6	25.2
	120	1.0	24.0	28.4	33.0	29.0	28.4	22.3	27.2	31.9	27.7	27.3
2	60	0.0	24.0	23.8	23.8	26.2	23.5	22.3	22.3	22.3	23.8	22.1
	60	0.1	24.0	24.6	24.8	26.1	24.1	22.3	23.2	23.2	24.1	22.7
	60	0.3	24.0	25.4	26.2	26.6	25.1	22.3	24.1	24.4	24.4	23.7
	60	0.6	24.0	26.2	29.6	28.8	27.3	22.3	25.5	27.7	27.9	26.2
	60	0.9	24.0	27.4	37.0	32.3	29.6	22.3	26.0	34.5	33.5	28.8
	60	1.0	24.0	27.6	41.4	31.2	31.3	22.3	26.5	39.7	30.9	30.5
	90	0.0	24.0	23.9	23.9	26.4	23.5	22.3	22.4	22.4	24.1	22.3
	90	0.1	24.0	24.2	24.4	26.1	23.8	22.3	22.6	22.7	23.9	22.3
	90	0.3	24.0	24.6	25.3	26.2	24.5	22.3	23.3	23.7	24.2	23.1
	90	0.6	24.0	25.1	27.5	26.6	25.9	22.3	24.1	25.9	25.3	24.8
	90	0.9	24.0	25.4	31.7	28.1	27.1	22.3	24.5	30.3	26.9	26.2
	90	1.0	24.0	25.8	35.4	27.6	28.2	22.3	24.3	34.0	26.2	27.3
	120	0.0	24.0	23.9	23.9	26.1	23.6	22.3	22.3	22.3	23.9	22.1
	120	0.1	24.0	24.1	24.3	26.2	23.9	22.3	22.6	22.7	24.0	22.4
	120	0.3	24.0	24.1	24.7	25.7	24.1	22.3	22.9	23.3	24.0	22.8
	120	0.6	24.0	24.6	26.4	25.9	25.2	22.3	23.4	24.6	24.2	23.7
	120	0.9	24.0	24.9	29.2	26.5	25.7	22.3	23.6	27.4	24.8	24.2
	120	1.0	24.0	25.1	31.3	27.1	26.2	22.3	23.7	29.2	24.8	24.7

Appendix A.4. Investigation of SRF results

In Figure A.1 we present the estimated survival functions of the SRF model for the first censoring scenario (left column) as well as the second scenario (right column). Each row panel represents a value for ρ, a proportion of censored data. Note that for these illustrattions, teh sample size and censoring depth was kept constant at 400 and 60, respectively. The proportions that were considered were $\rho = \{0.1, 0.3, 0.6, 0.9, 1.0\}$, shown in the first, second, third, fourth, and fifth row. In each title, information of the distribution of soil thickness is provided, as well as the number of distinct soil thickness values.

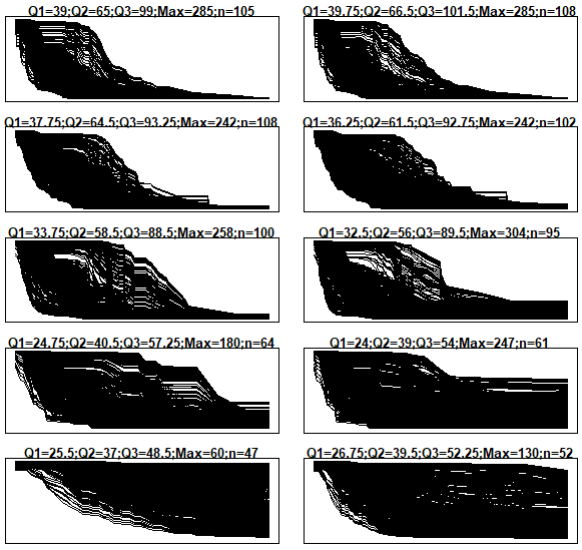

Figure A.1: Estimated survival functions from which predictions were obtained for the simulation study.

Appendix B. Additional information for the Maine case study

Appendix B.1. Covariate information

Covariates for the Maine case study is presented in Table B.3.

Table B.3: Overview of geodata sets and derived covariates for the Maine case study. r: pixel size for raster datasets or scale for vector datasets, p: number of covariates per dataset, NDVI: normalized differenced vegetation index, SWI: Saga wetness index.

geodata set	r	p	covariate examples
Morphometry			
Convergence index	1:100	1	maximum curvature, cross-sectional
Curvature and derivatives	1:100 000	2	curvature, mean elevation
Elevation and derivatives	1:10	1	
Hydrology			
Catchment and SWI	1:1 000	2	catchment slope, SWI, valley depth, ridge
Depth	1:10	3	height
Relative landscape position	1:100	1	

Appendix B.2. Additional modelling results

In Figure B.2, density plots of the predictions (non-truncated) are shown, together with the densities of soil thickness from the calibration data. As expected the densities of ARF and IPC-RF were very similar, because all data (censored and non-censored) were taken into account in both models. These two models also produced predictions that were skewed to the left, mostly accounting for the non-censored data, and somewhat for the lower soil thickness data that were not censored. The densities of HRF and SRF were also similar, but SRF showed somewhat larger predictions overall (influenced by the large censored soil thickness data).

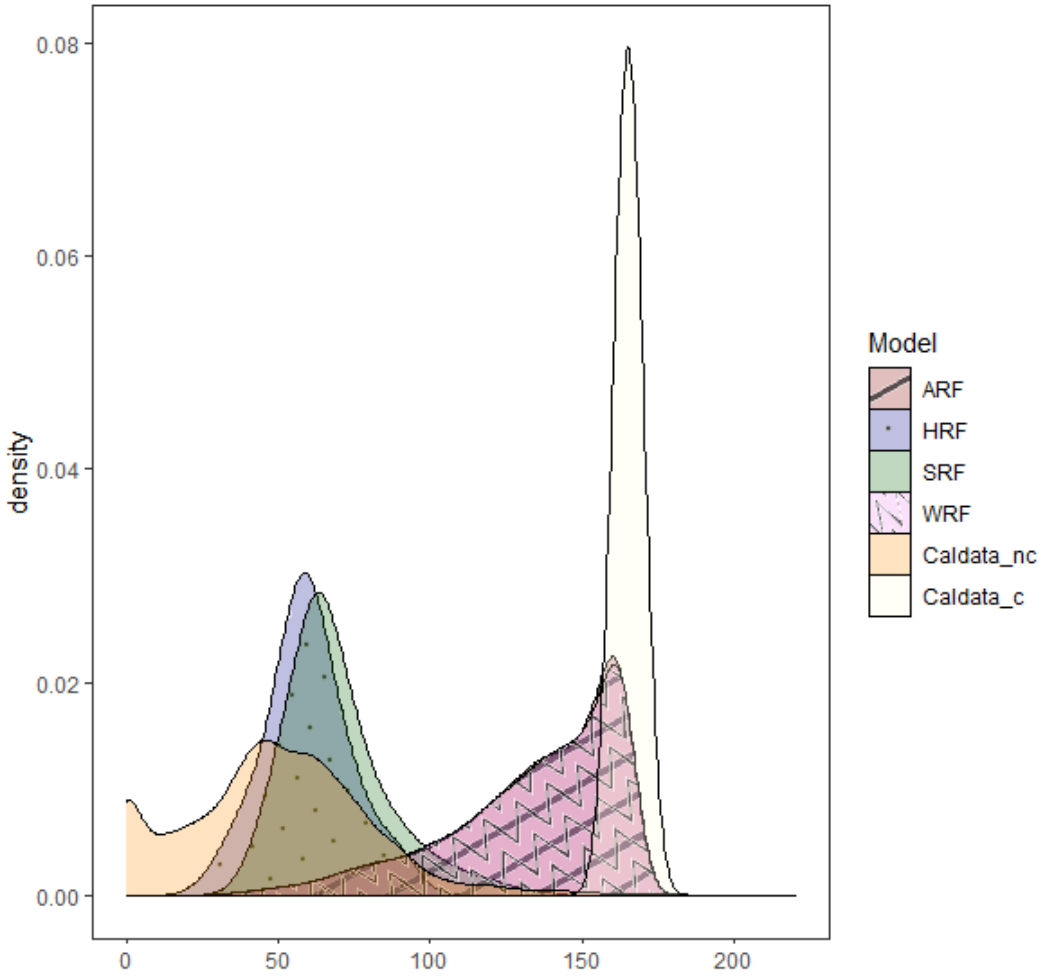

Figure B.2: Density plots of the predictions for the Maine case study. Densities of soil thickness from the calibration data, non-censored (Caldata_nc) and censored (Caldata_c), is also shown.

Appendix C. Additional information for the Switzerland case study

Appendix C.1. Covariate information

Covariates for the Switzerland case study is presented in Table C.4.

Table C.4: Overview of geodata sets and derived covariates for the Switzerland case study. r: pixel size for raster datasets or scale for vector datasets, p: number of covariates per dataset, NDVI: normalized differenced vegetation index, TPI: topographic position index, TWI: topographic wetness index, MRVBF: multi-resolution valley bottom flatness).

geodata set	r	p	covariate examples
Soil			
Soil overview map (FSO, 2000)	1:200 000	6	physiographic units, surface of historic wetlands, presence of drainage networks or soil amelioration
Wetlands Wild maps (ALN, 2002)	1:50 000	1	
Wetlands Siegfried maps (Wüst-Galley et al., 2015)	1:25 000	1	
Artificial drainage networks (ALN, 2014b)	1:5 000	2	
Parent material			
Map of last glacial maximum (Swisstopo, 2009)	1:500 000	1	(aggregated) geological units, distance to end moraines, occurrence of calcerous parent material, ice level during last glaciation, aquifers
Geotechnical map (Swisstopo, 2007; BAFU and GRID-Europe, 2010)	1:200 000	2	
Geological map (ALN, 2014a)	1:50 000	7	
Groundwater occurrence (AWEL, 2014)	1:25 000	2	
Climate			
Climatic means 1981–1990 (Frehner et al., 2011)	100 m	17	mean annual/monthly temperature and precipitation, temperature variation index
Topography			
Digital terrain model (Swisstopo, 2020)	2 m	50	elevation, slope, slope length and position, curvature, northness, elevation above channel network, catchment area, TWI, TPI, MRVBF (various radii/resolutions)

Appendix C.2. Additional modelling results

The densities of the predictions (non-truncated) are shown in Figure C.3. The densities of HRF and SRF are very similar and overlap more with the density of the non-censored calibration data. ARF and IPC-RF is somewhat in

between the two densities of the non-censored and censored data. It is important to note that IPC-RF produced a longer tail to the right to account for the larger non-censored soil thickness data.

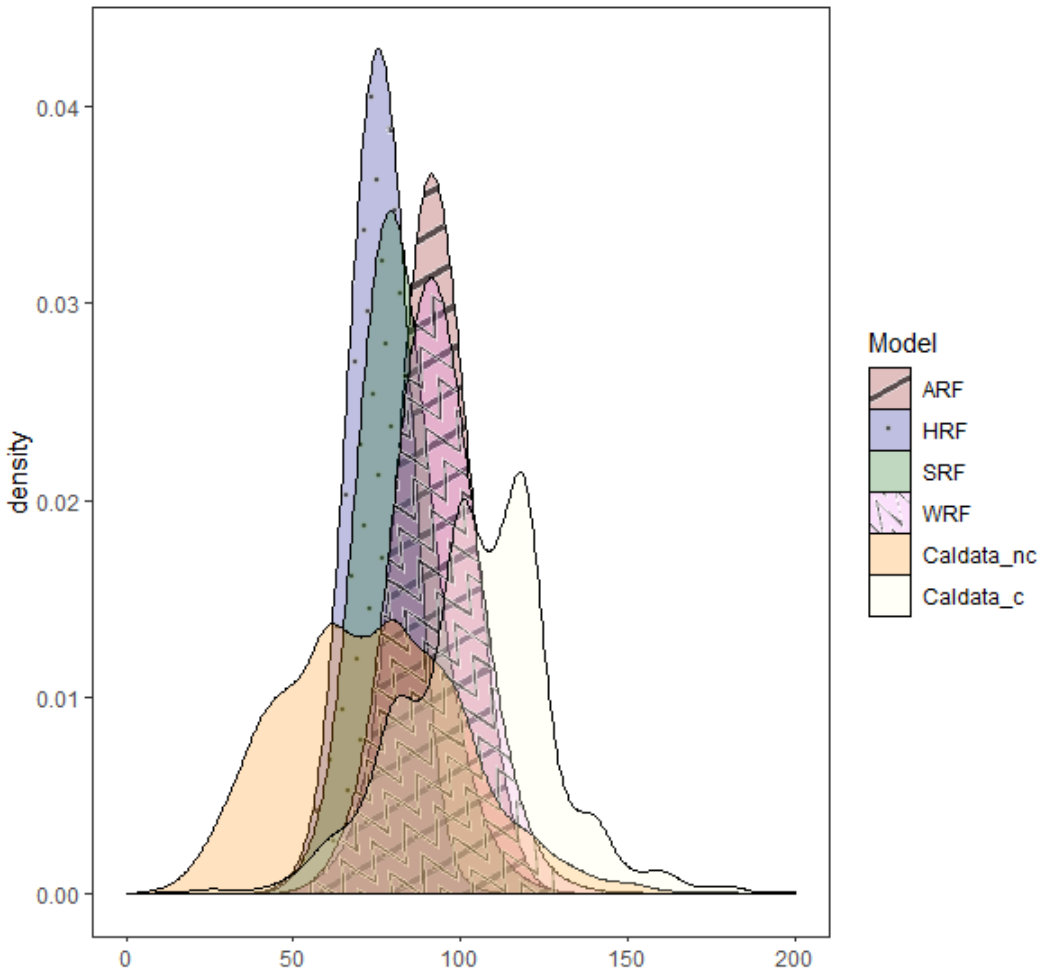

Figure C.3: Density plots of the predictions for the Switzerland case study. Densities of soil thickness from the calibration data, non-censored (Caldata_nc) and censored (Caldata_c), are also shown.

References

ALN (2002). Historische Feuchtgebiete der Wildkarte 1850. Amt für Landschaft und Natur des Kantons Zürich. URL: http://www.aln.zh.ch/internet/baudirektion/aln/de/naturschutz/naturschutzdaten/geodaten.html last access 29.03.2017.

ALN (2014a). Geologische Karte des Kantons Zürich nach Hantke et. al 1967, GIS-ZH Nr. 41. Amt für Landschaft und Natur des Kantons Zürich. URL: http://www.gis.zh.ch/Dokus/Geolion/gds_41.pdf last access: 15.02.2015.

ALN (2014b). Meliorationskataster des Kantons Zürich, GIS-ZH Nr. 148. Amt für Landschaft und Natur des Kantons Zürich. URL: http://www.geolion.zh.ch/geodatensatz/show?nbid=387 last access 29.03.2017.

AWEL (2014). Grundwasservorkommen, GIS-ZH Nr. 327. Amt für Abfall, Wasser, Energie und Luft des Kanton Zürich. URL: http://www.geolion.zh.ch/geodatensatz/show?nbid=723 last access 29.03.2017.

BAFU and GRID-Europe (2010). *Swiss Environmental Domains. A new spatial framework for reporting on the environment*. Environmental studies 1024 Federal Office for the Environment FOEN, Berne. URL: https://www.bafu.admin.ch/bafu/en/home/topics/landscape/publications-studies/publications/swiss-environmental-domains.html last access 15.09.2023.

Frehner, M., Remund, J., Walthert, L., Kägi, M., Rihm, B., & Brang, P. (2011). *Schätzung standortspezifischer Trockenstressrisiken in Schweizer Wäldern. Schlussbericht / Version 2.3*. Report Eidg. Forschungsanstalt für Wald, Schnee und Landschaft WSL Birmensdorf. doi:doi: 10.3929/ethz-a-010693256.

FSO (2000). Swiss soil suitability map 1:20000, published 1980. BFS GEOSTAT. Swiss Federal Statistical Office. URL: https://www.bfs.admin.ch/bfs/en/home/services/geostat/swiss-federal-statistics-geodata/land-use-cover-suitability/derivative-complementary-data/swiss-soil-suitability-map.html last accessed 15.09.2023.

Swisstopo (2007). Lithological-petrographic map of Switzerland 1:200000 , original name: Geotechnical Map of Switzerland 1:200000, published 1967. URL: https://opendata.swiss/de/dataset/geotechnische-karte-der-schweiz-1-200000 last access: 15.09.2023.

Swisstopo (2009). Switzerland during the Last Glacial Maximum 1:500 000. URL: https://www.swisstopo.admin.ch/en/geodata/geology/maps/gk500/raster.html last access: 15.09.2023.

Swisstopo (2020). swissAlti3D. The high precision digital elevation model of Switzerlandd. URL: https://www.swisstopo.admin.ch/en/geodata/height/alti3d.html last access: 08.09.2023.

van der Westhuizen, S., Heuvelink, G. B. M., Hofmeyr, D. P., & Poggio, L. (2022). Measurement error-filtered machine learning in digital soil mapping. *Spatial Statistics*, *47*, 100572. doi:doi: https://doi.org/10.1016/j.spasta.2021.100572.

Wüst-Galley, C., Grünig, A., & Leifeld, J. (2015). *Locating organic soils for the Swiss greenhouse gas inventory*. Agroscope Science 26 Agroscope Zurich. URL: https://www.bafu.admin.ch/dam/bafu/en/dokumente/klima/klima-climatereporting-referenzen-cp2/wuest-galley_c_gruenigaleifeldj2015.pdf.download.pdf last access: 08.09.2023.