Supplementary Material for:

Article

Sample Size Optimization for Digital Soil Mapping: An Empirical Example

Daniel D Saurette [1,2], Richard J. Heck [1], Adam W. Gillespie [1], Aaron A. Berg [3] and Asim Biswas [1,*]

[1] School of Environmental Sciences, University of Guelph, 50 Stone Rd East, Guelph, Ontario, Canada, N1G 2W1; dsaurett@uoguelph.ca (D.D.S.); rheck@uoguelph.ca (R.J.H.); agilles@uoguelph.ca (A.W.G.); biswas@uoguelph.ca (A.B.)
[2] Ontario Ministry of Agriculture, Food and Rural Affairs, 1 Stone Rd West, Guelph, Ontario, Canada, N1G 2Y4; daniel.saurette@ontario.ca (D.D.S.)
[3] Department of Geography, Environment & Geomatics, University of Guelph, 50 Stone Rd East, Guelph, Ontario, Canada, N1G 2W1; aberg@uoguelph.ca (A.A.B.)
* Correspondence: biswas@uoguelph.ca

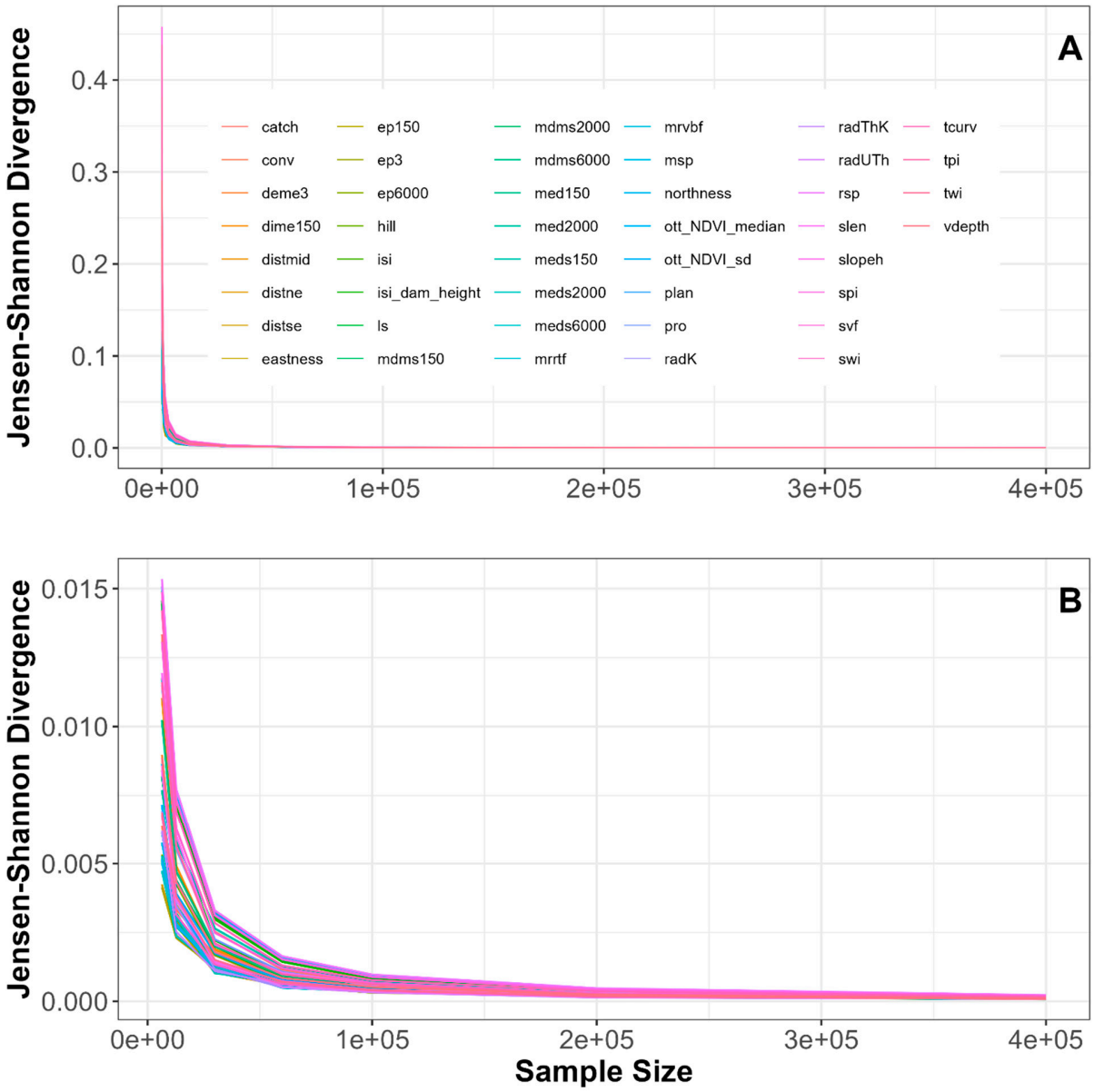

Figure S1. Jensen-Shannon Divergence (D_{JS}) as a function of sample size for the 44 continuous covariates retained after the variance inflation factor covariate reduction. Sample plans were generated from the full covariate rasters using simple random sampling. Subplot A shows the D_{JS} across all sample sizes, while subplot B highlights the D_{JS} for sample sizes ≥6400. Codes for the covariates are provided for completeness despite the overlap, which prevents identifying them individually.

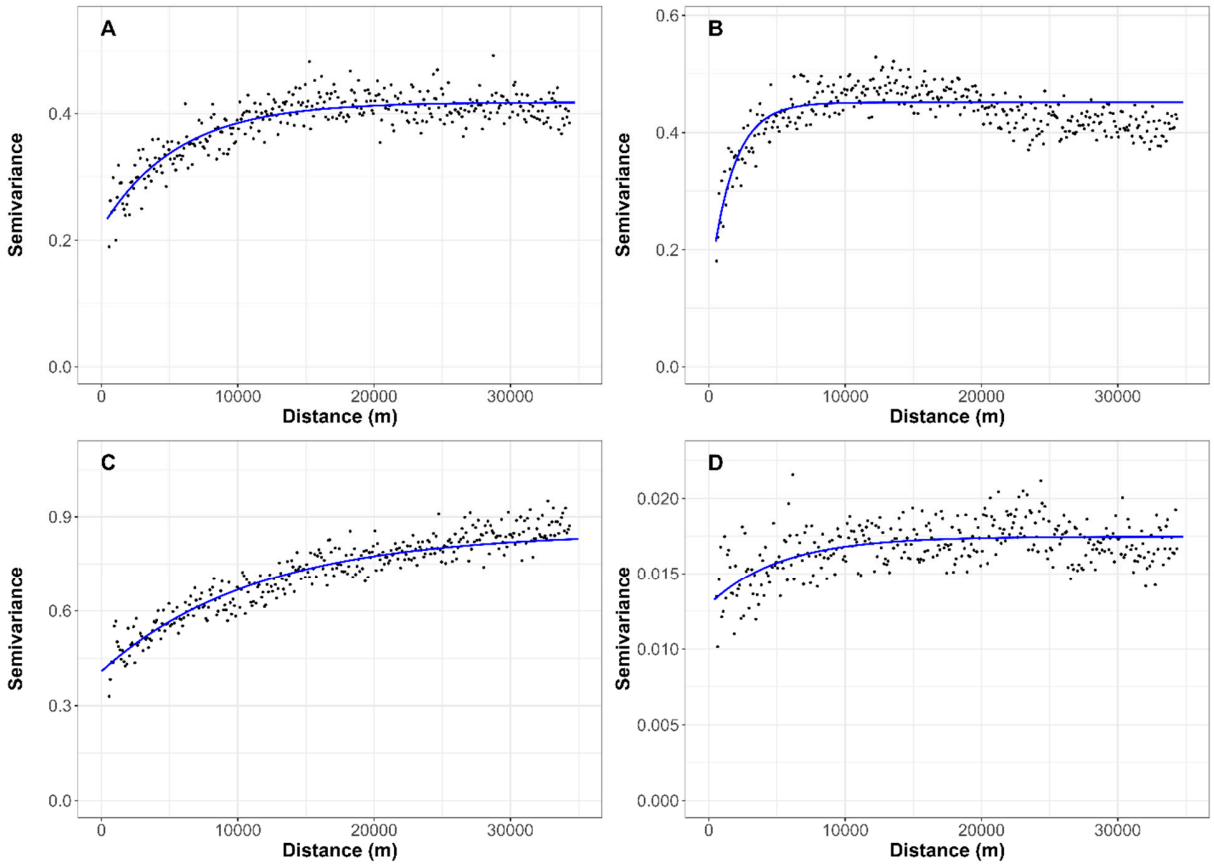

Figure S2. Experimental semivariograms with a fitted exponential model (solid line) for the four target soil properties: cation exchange capacity (A), clay content (B), soil pH (C), and soil organic carbon (D). Note that the kriging was performed on transformed values for cation exchange capacity, clay, and soil organic carbon; therefore, the semivariances presented along the ordinate are transformed units.

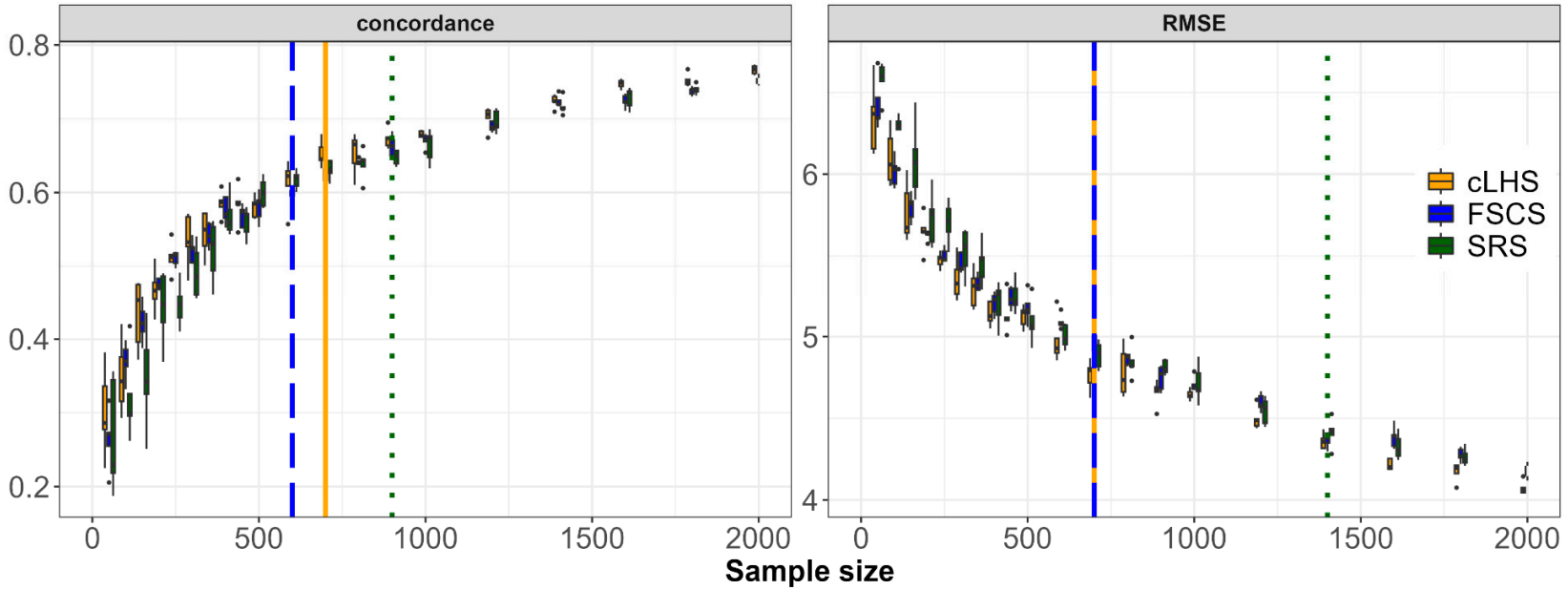

Figure S3. Change in the concordance and root mean square error (RMSE) with increasing sample size from the external validation of the random forest models trained with sample plans developed using conditioned Latin hypercube sampling (cLHS), feature space coverage sampling (FSCS), and simple random sampling (SRS) for clay content. The solid (orange) vertical line, dashed (blue) vertical line, and dotted (green) vertical line identify the optimal sample size based on the unit invariant knee for the cLHS, FSCS, and SRS sampling algorithms, respectively.

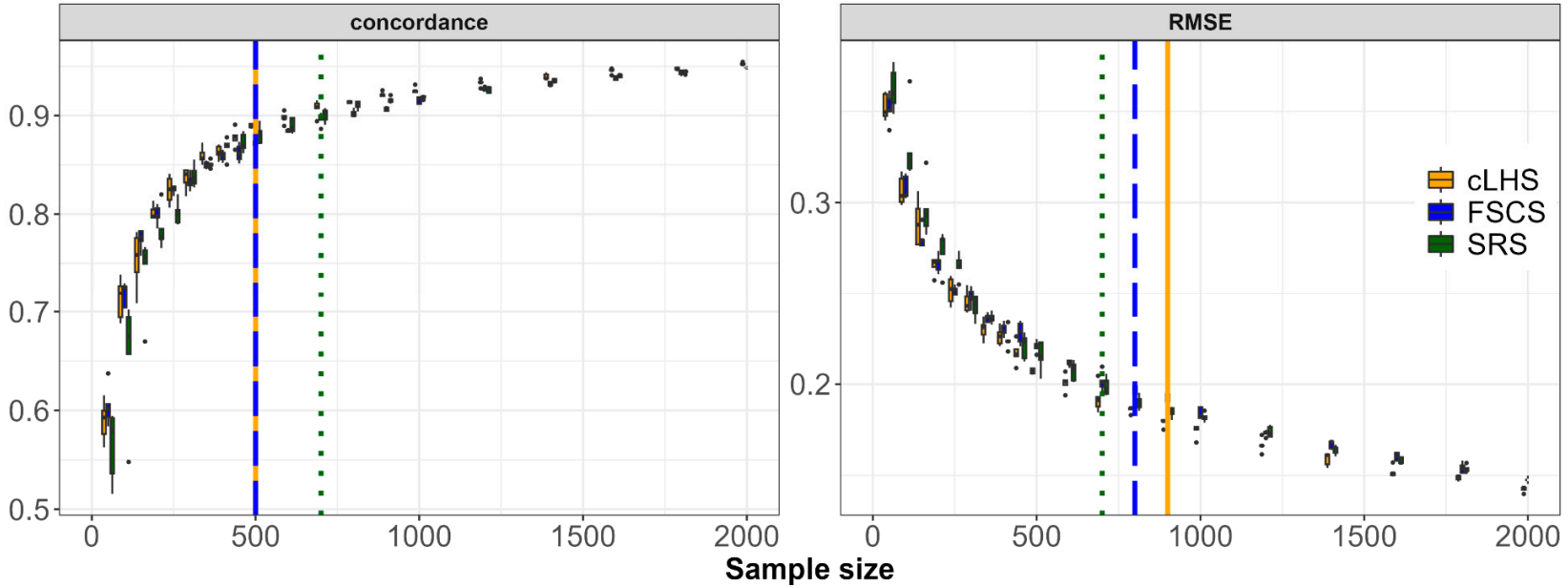

Figure S4. Change in the concordance and root mean square error (RMSE) with increasing sample size from the external validation of the random forest models trained with sample plans developed using conditioned Latin hypercube sampling (cLHS), feature space coverage sampling (FSCS), and simple random sampling (SRS) for soil pH. The solid (orange) vertical line, dashed (blue) vertical line, and dotted (green) vertical line identify the optimal sample size based on the unit invariant knee for the cLHS, FSCS, and SRS sampling algorithms, respectively.

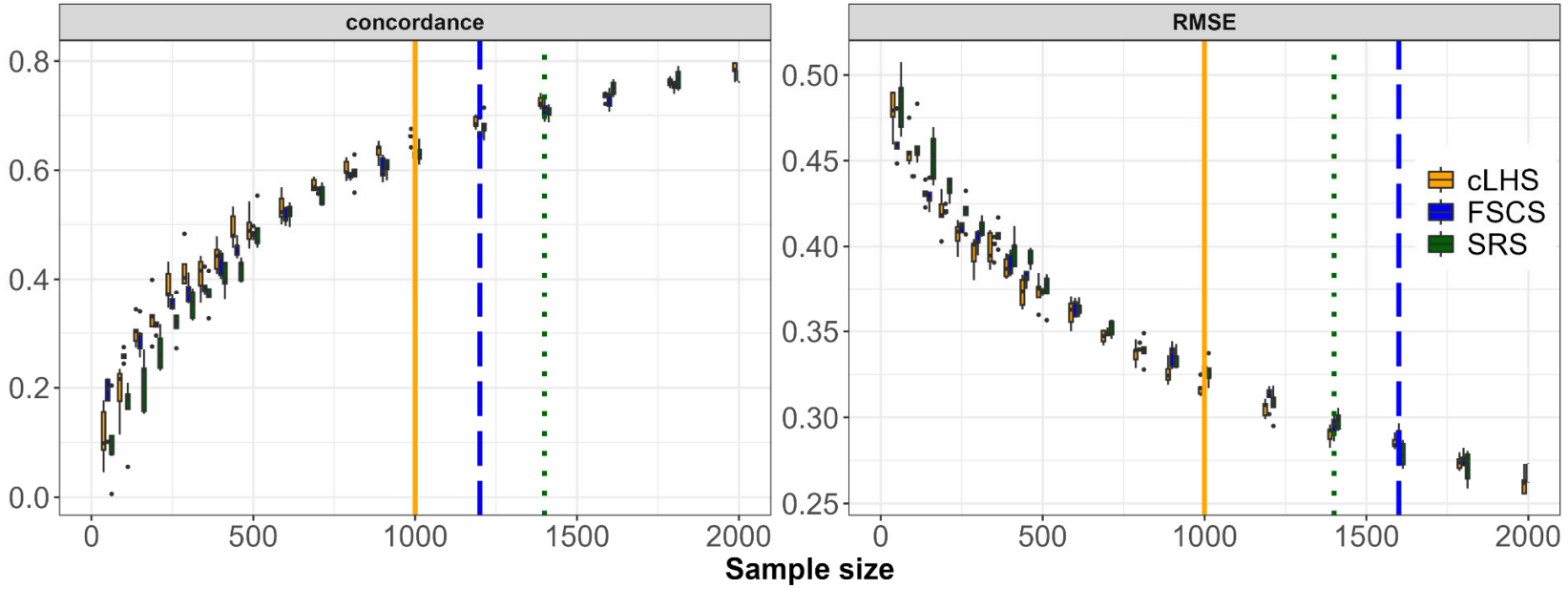

Figure S5. Change in the concordance and root mean square error (RMSE) with increasing sample size from the external validation of the random forest models trained with sample plans developed using conditioned Latin hypercube sampling (cLHS), feature space coverage sampling (FSCS), and simple random sampling (SRS) for soil organic carbon. The solid (orange) vertical line, dashed (blue) vertical line, and dotted (green) vertical line identify the optimal sample size based on the unit invariant knee for the cLHS, FSCS, and SRS sampling algorithms, respectively.

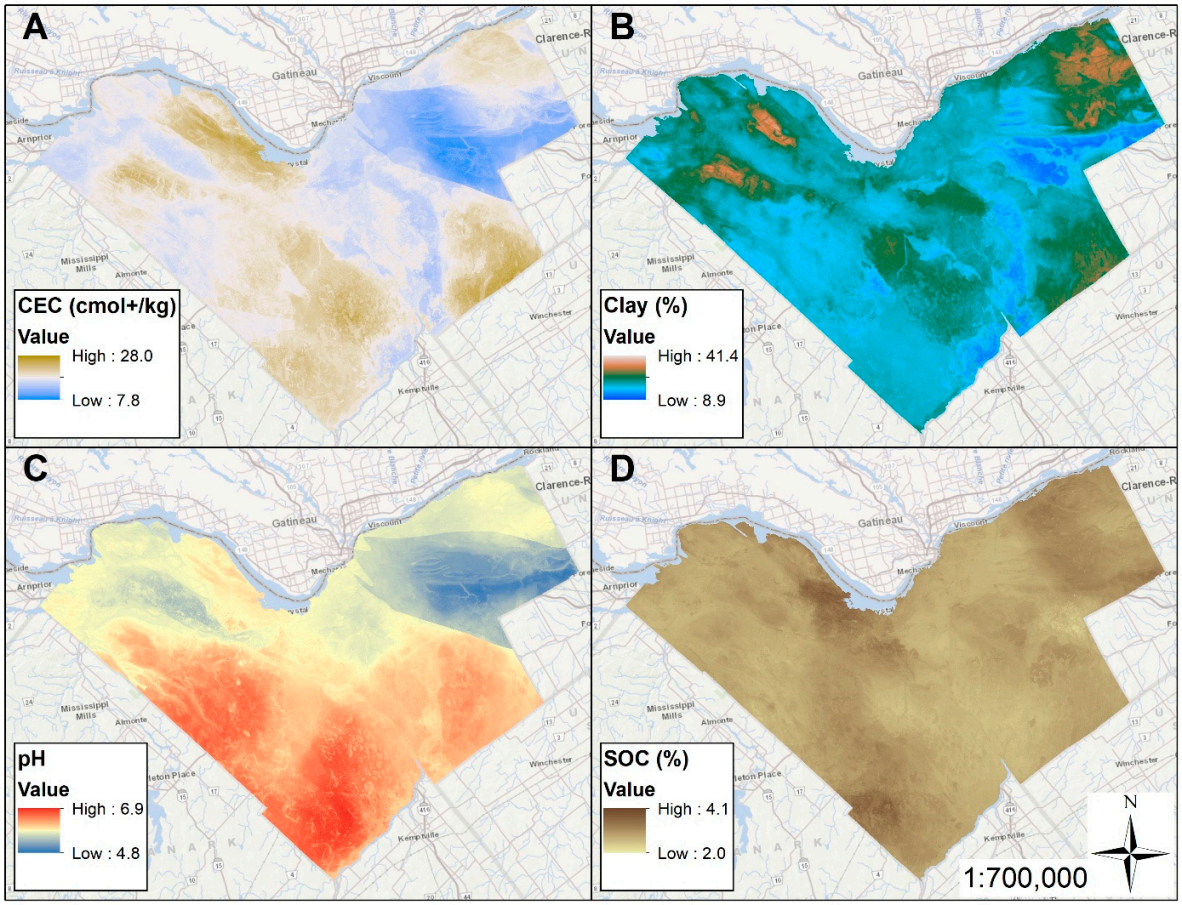

Figure S6. Random forest predictions of (A) cation exchange capacity (CEC), (B) clay content, (C) soil pH, and (D) soil organic carbon (SOC) for the Ottawa Study area using a sample plan created with feature space coverage sampling and the overall optimal sample size of 874 sample locations.

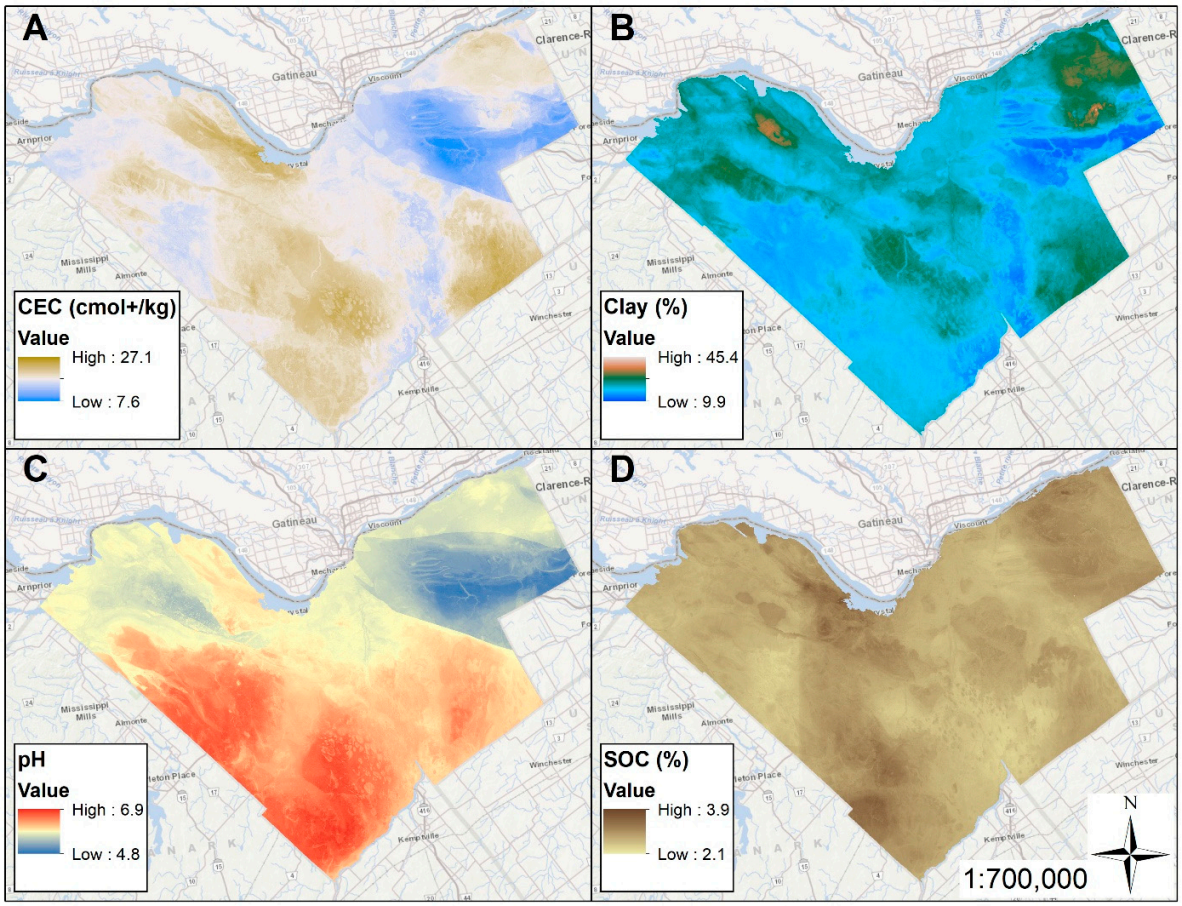

Figure S7. Random forest predictions of (A) cation exchange capacity (CEC), (B) clay content, (C) soil pH, and (D) soil organic carbon (SOC) for the Ottawa Study area using a sample plan created with simple random sampling and the overall optimal sample size of 869 sample locations.

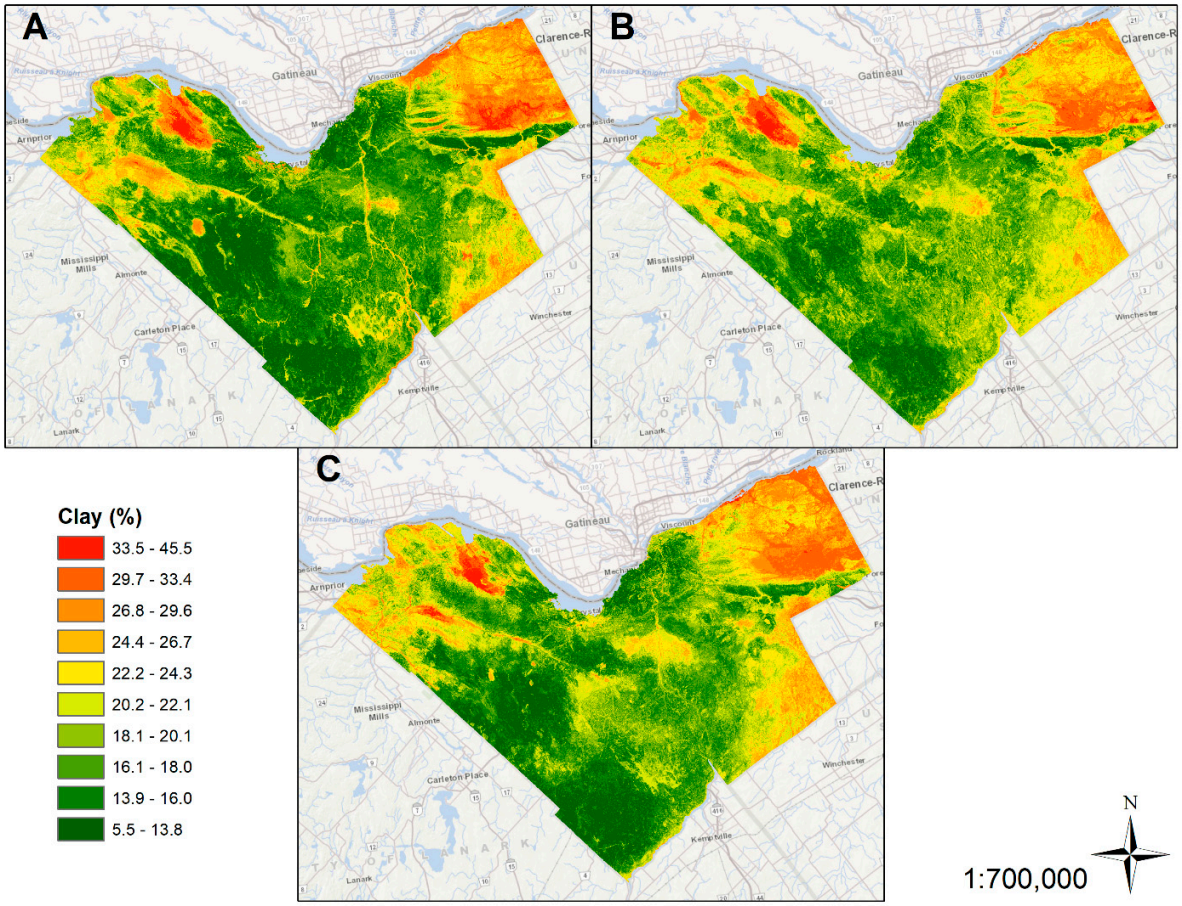

Figure S8. Prediction interval maps (90%) for clay content generated using quantile regression forest for the optimal sample sizes based on the Jenson–Shannon Divergence for conditioned Latin hypercube sampling (A), feature space coverage sampling (B), and simple random sampling (C) algorithms.

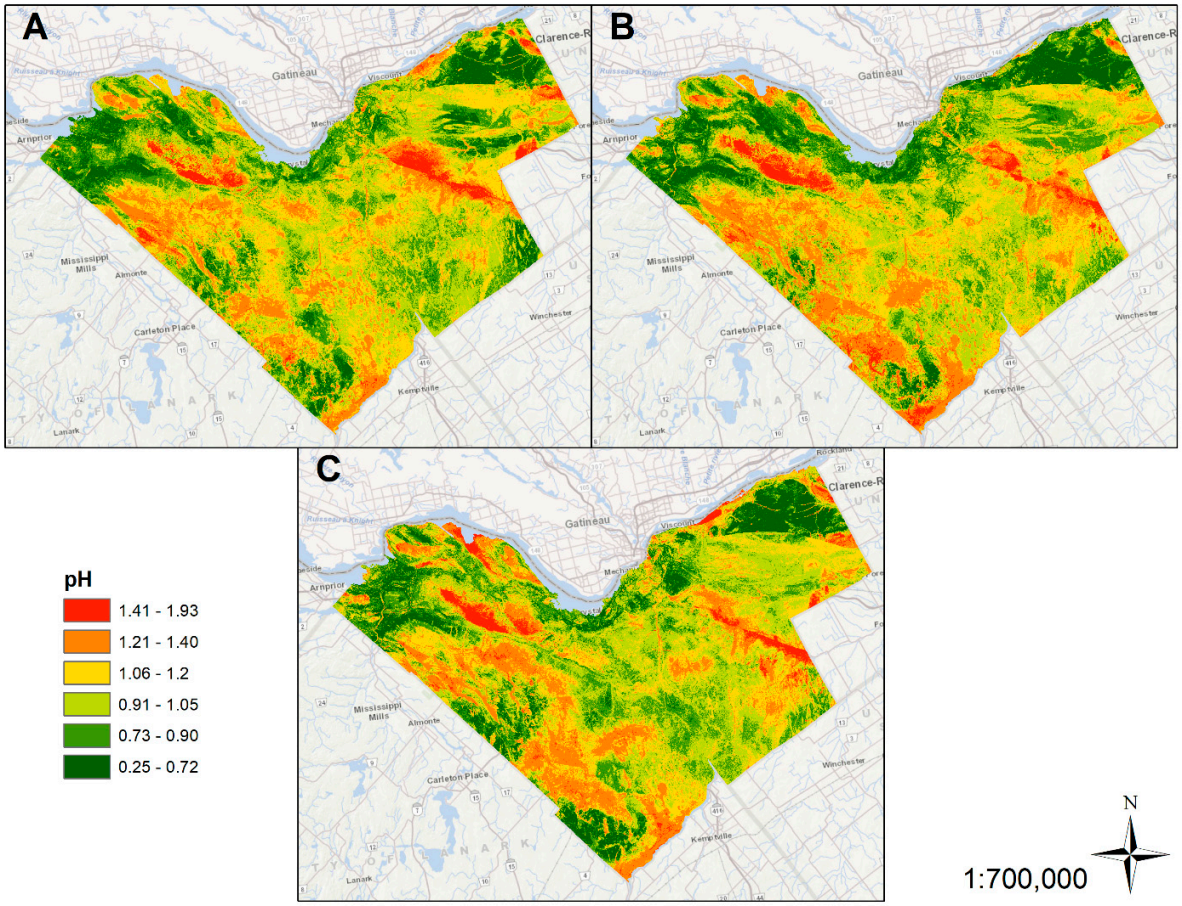

Figure S9. Prediction interval maps (90%) for soil pH generated using quantile regression forest for the optimal sample sizes based on the Jenson–Shannon Divergence for the conditioned Latin hypercube sampling (A), feature space coverage sampling (B), and simple random sampling (C) algorithms.

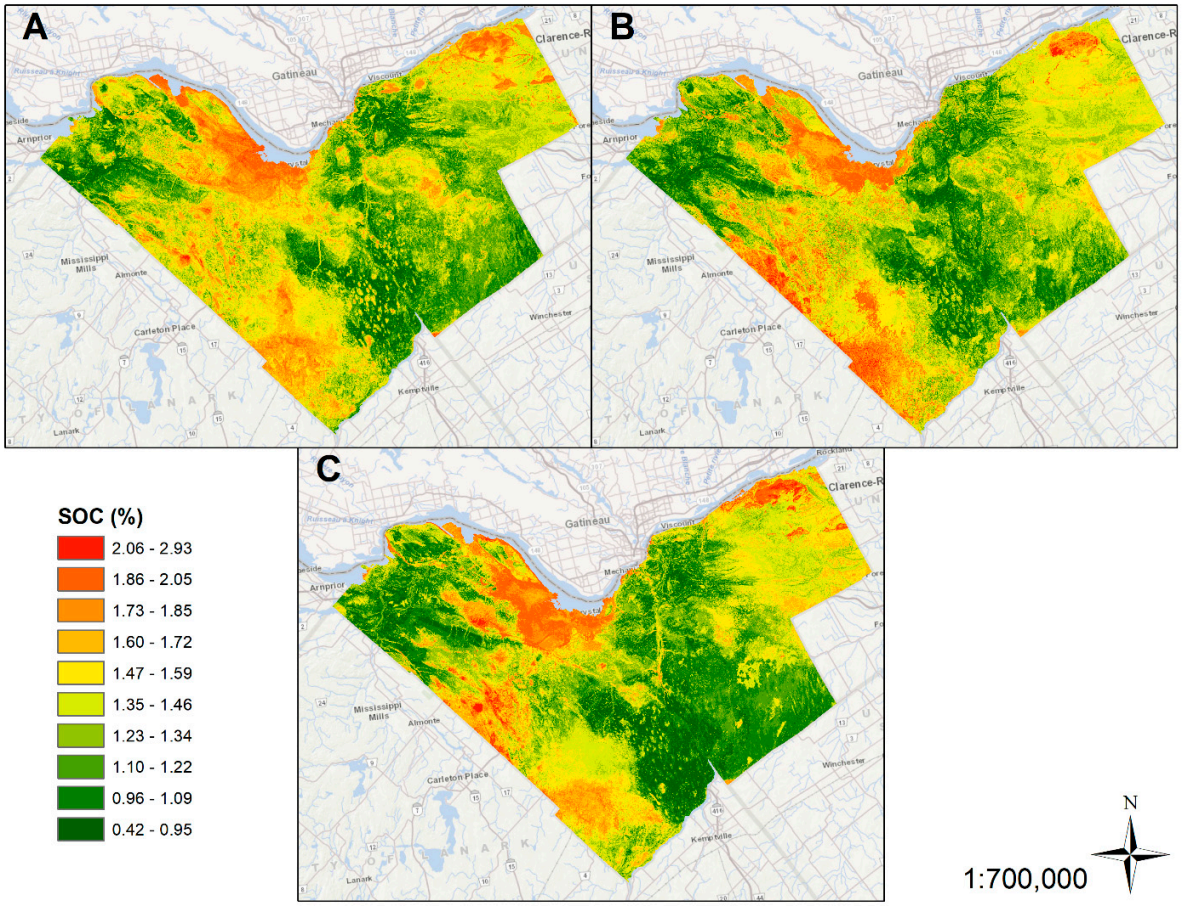

Figure S10. Prediction interval maps (90%) for soil organic carbon (SOC) generated using quantile regression forest for the optimal sample sizes based on the Jenson–Shannon Divergence for conditioned Latin hypercube sampling (A), feature space coverage sampling (B), and simple random sampling (C) algorithms.