Where to Sample Efficiently for More?

-- Combining uncertainties from the feature domain and spatial domain for digital soil mapping

A-Xing Zhu[1,2], Yan Li[3], Zhou Shi[3]

[1] *Department of Geography,* ***University of Wisconsin-Madison****, Madison, WI 53706, USA*

[2] *School of Geography,* ***Nanjing Normal University****, 210023, China*

[3] *Institute of Land Science and Property Management, School of Public Affairs,* ***Zhejiang University****, Hangzhou 310058, China*

Background – where to sample efficiently for more

Existing soil samples, which haven collected during previous soil and land resource surveys, serve as an important resource for digital soil mapping.

For many reasons we often need to add additional samples to the existing set. Sample for more (Arrouays et al., 2014).

Naturally, people would ask “Where to sample for more?”, but a more interesting question is “Where to sample **efficiently** for more?”.

Background – where to sample efficiently for more

"Efficiently" here refers to as maximizing the reduction in uncertainty in the final product through the selection of a new sample, in hope to achieve maximum increase in accuracy.

Referred to as "Uncertainty Directed Sampling"

In DSM, efforts in this area can be grouped into two general categories: uncertainty in the spatial domain and uncertainty in feature domain.

Background – where to sample efficiently for more

Uncertainty directed sampling in the spatial domain draws additional samples based on their effects on the prediction variance from a spatial model (kriging) (McBratney and Webster, 1981; Delmelle and Goovaerts, 2009).

It achieves sampling “efficiency” through spatial configuration of samples.

Uncertainty directed sampling in the feature domain adds additional samples based on the prediction uncertainty from a feature model (iPSM) (Zhu et al., 2015; Zhang et al, 2016).

It achieves sampling “efficiency” through sample representativeness in the feature domain.

The Question – Can and how we combine the two?

It seems that uncertainty in the spatial domain and uncertainty in the feature domain are covering the two sides of the same coin (soil formation).

Would it be more "efficient" if we combine the two uncertainties in designing for additional samples?

The Idea – Combining the lists

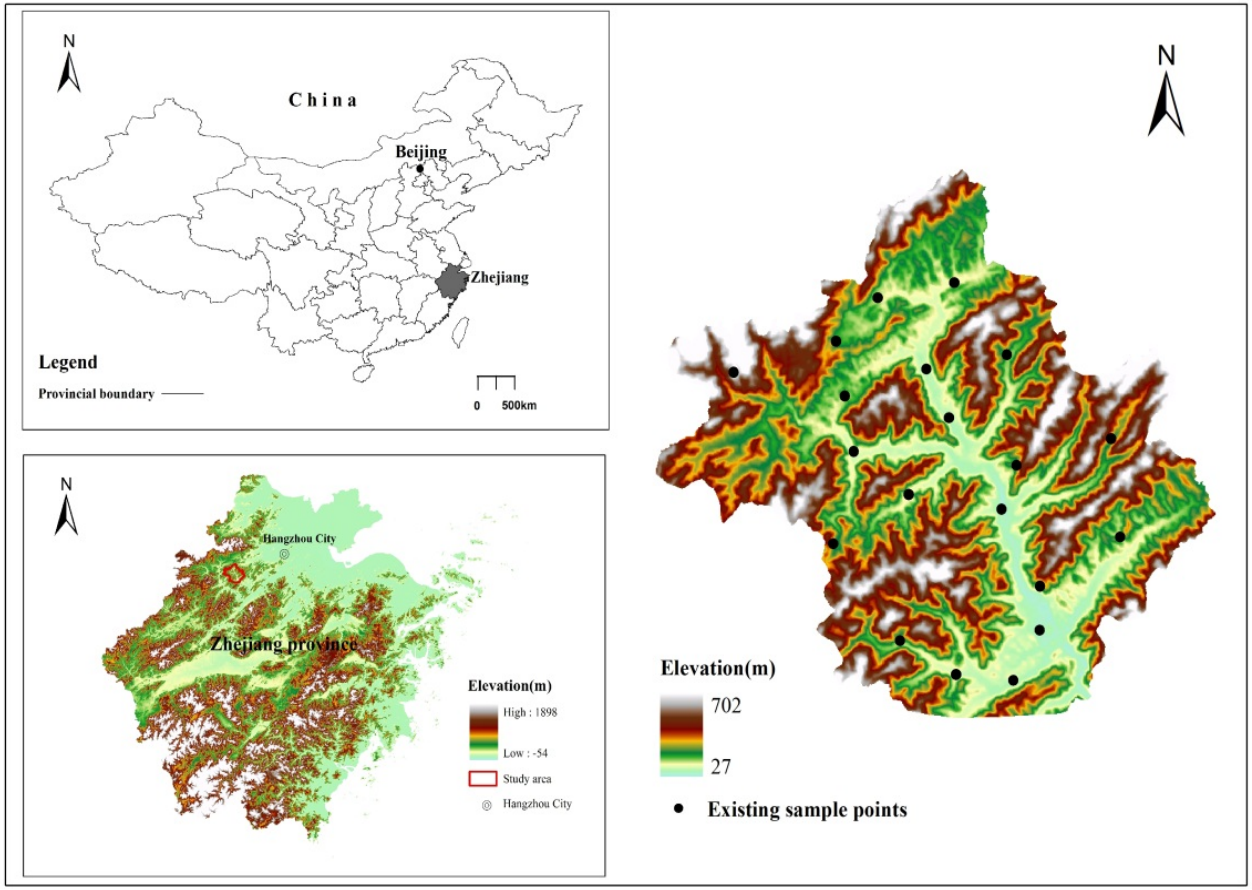
Ordered list of samples based on the uncertainty from the spatial domain

Ordered list of samples based on the uncertainty from the feature domain

Combine the top half from each of the two ordered lists to form a third ordered list representing both domains

Compare and evaluate the soil maps based on the three lists.

Case Study – Fuyang, Zhejiang, China



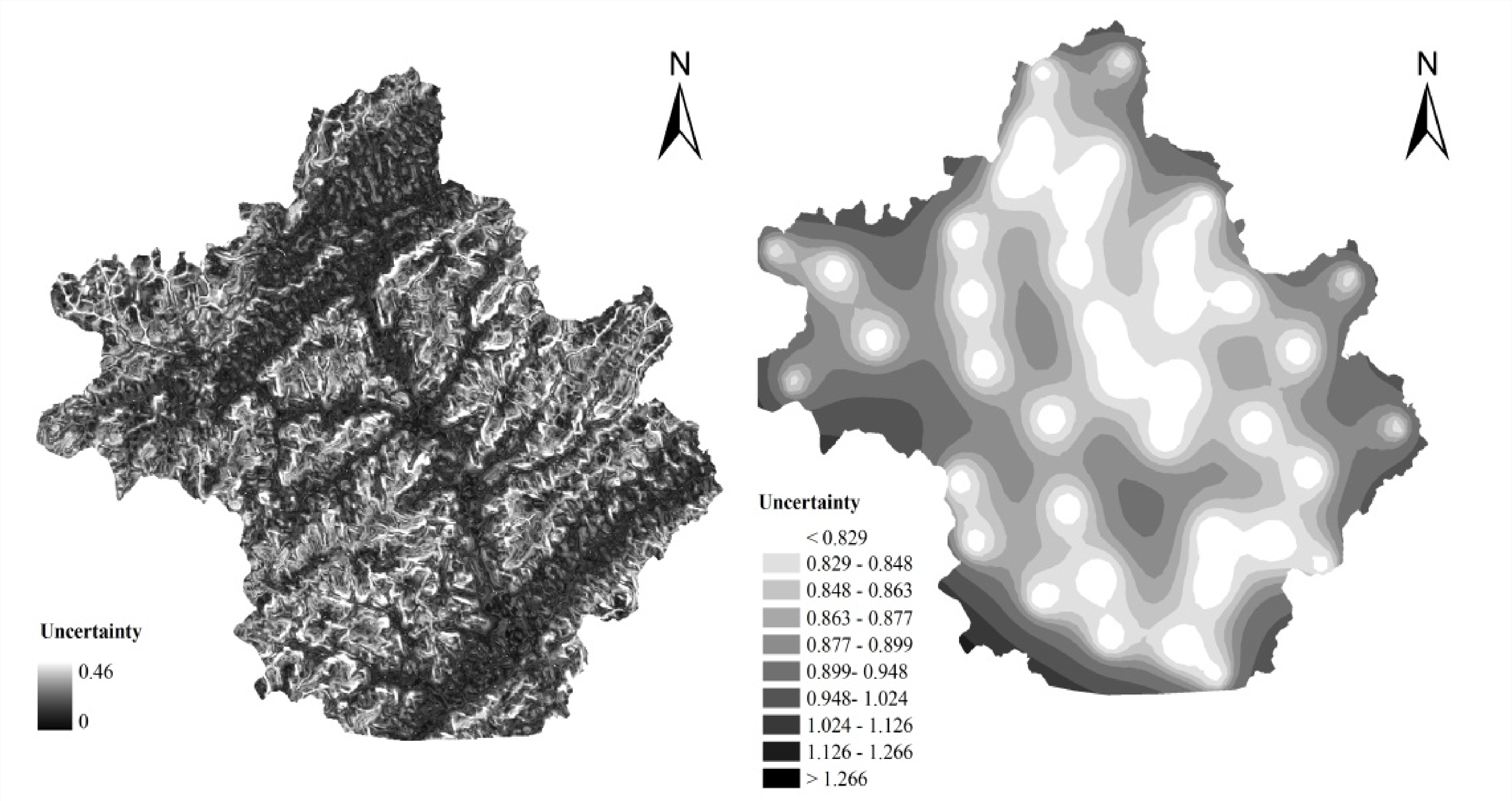
Area size:
299.14km^2

Existing samples:
50

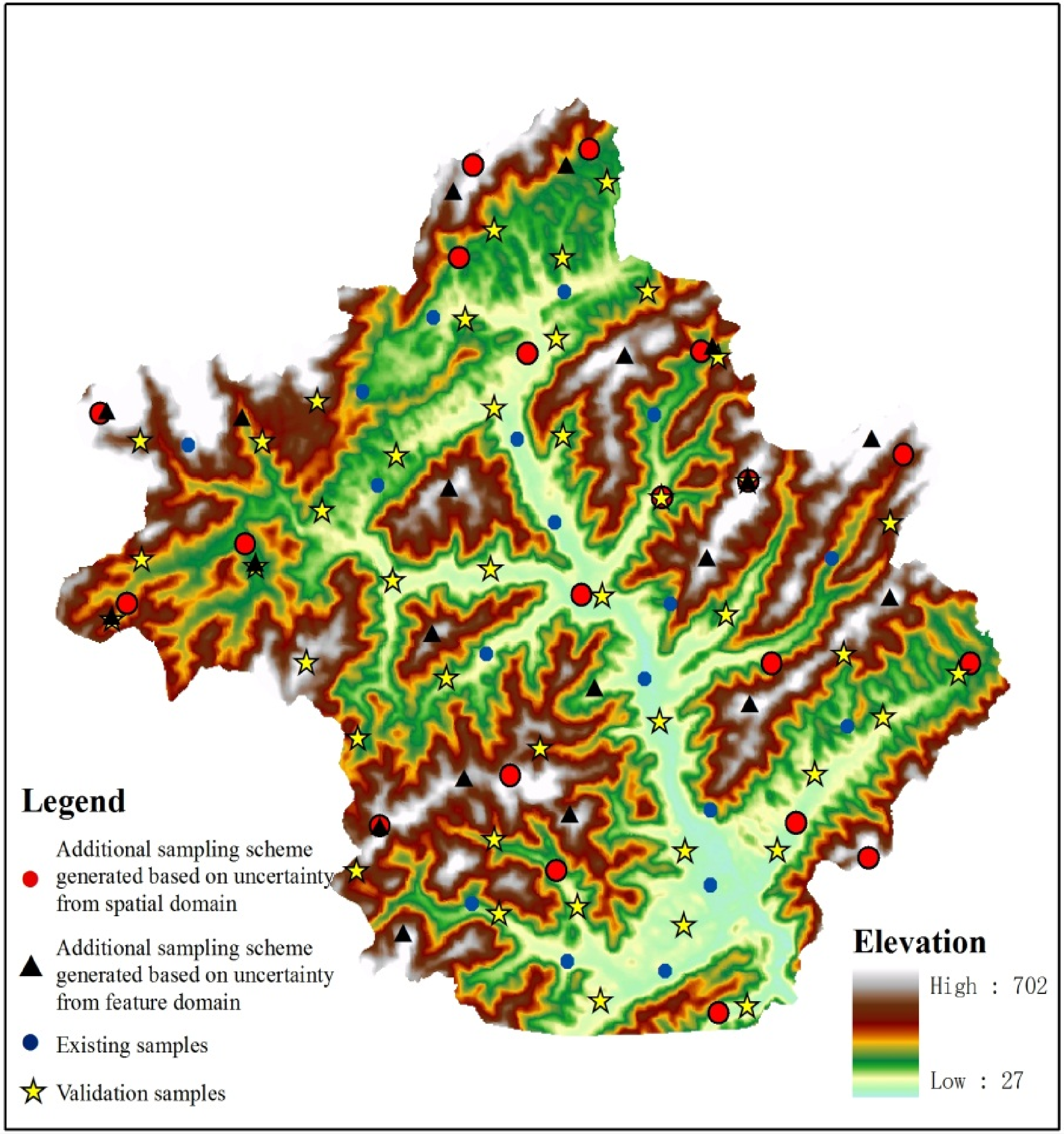
Validation samples:
42

Property to be predicted:
SOM

Comparison of two uncertainty maps



Prediction uncertainty maps based on 20 existing samples plus 20 combined samples: using iPSM method with the uncertainty threshold equal to 0.3 (left) and using ordinary kriging (right)



Spatial distribution of additional samples:

Black triangles: based on uncertainty from feature domain (iPSM, Zhang et al., 2016)

Red dots: based on uncertainty from spatial domain (kriging, spatial simulated annealing, Van Groenigen, 1999)

Two sets of samples were collected:

20 samples based on uncertainty from the feature domain

20 samples based on uncertainty from the spatial domain.

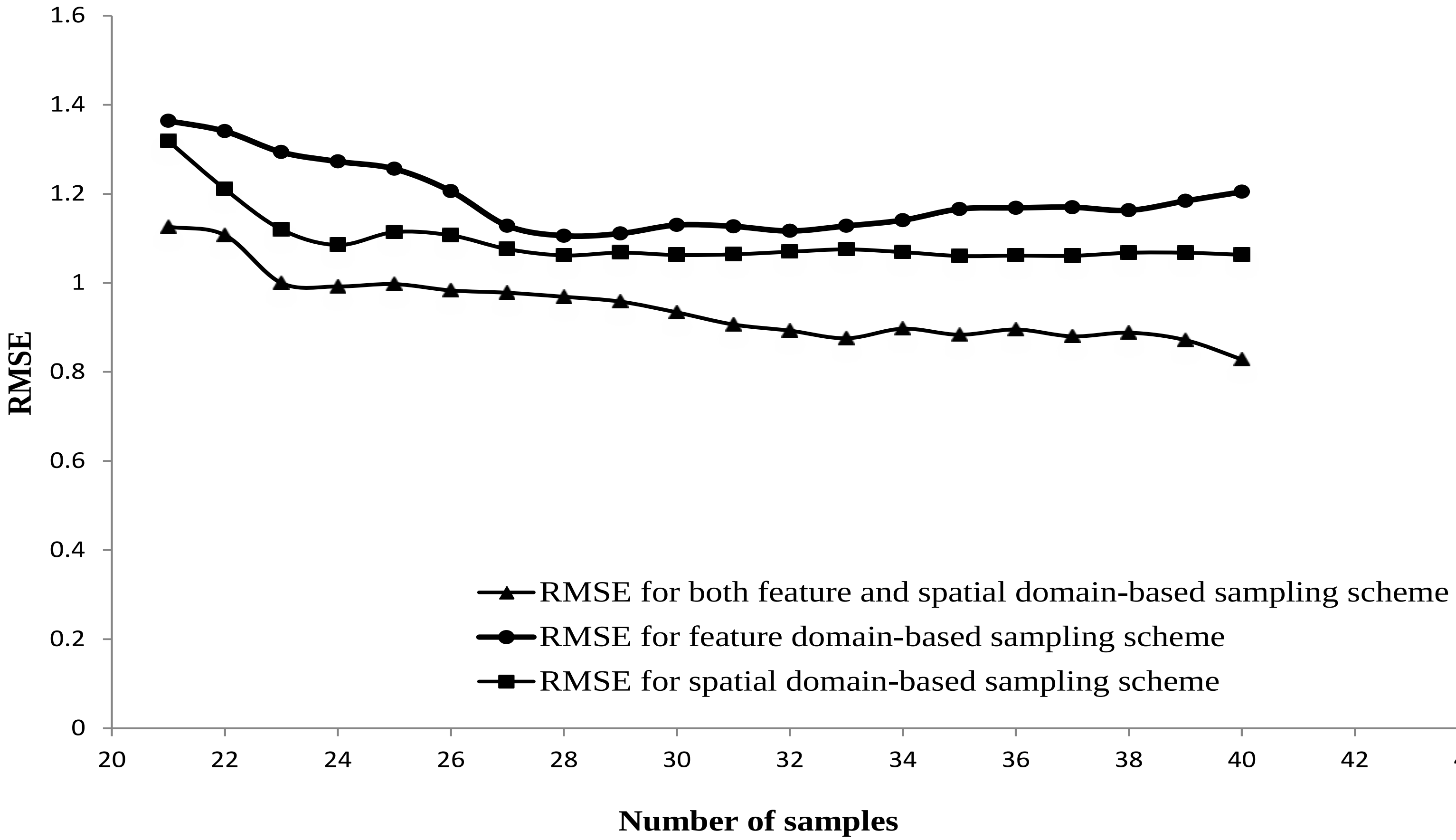
The first 10 points from each of the two lists were selected and merged in order to generate the third ordered list of 20 samples.

Each list was added to the existing samples (20) to form the pooled samples (total of 40) for digital soil mapping.

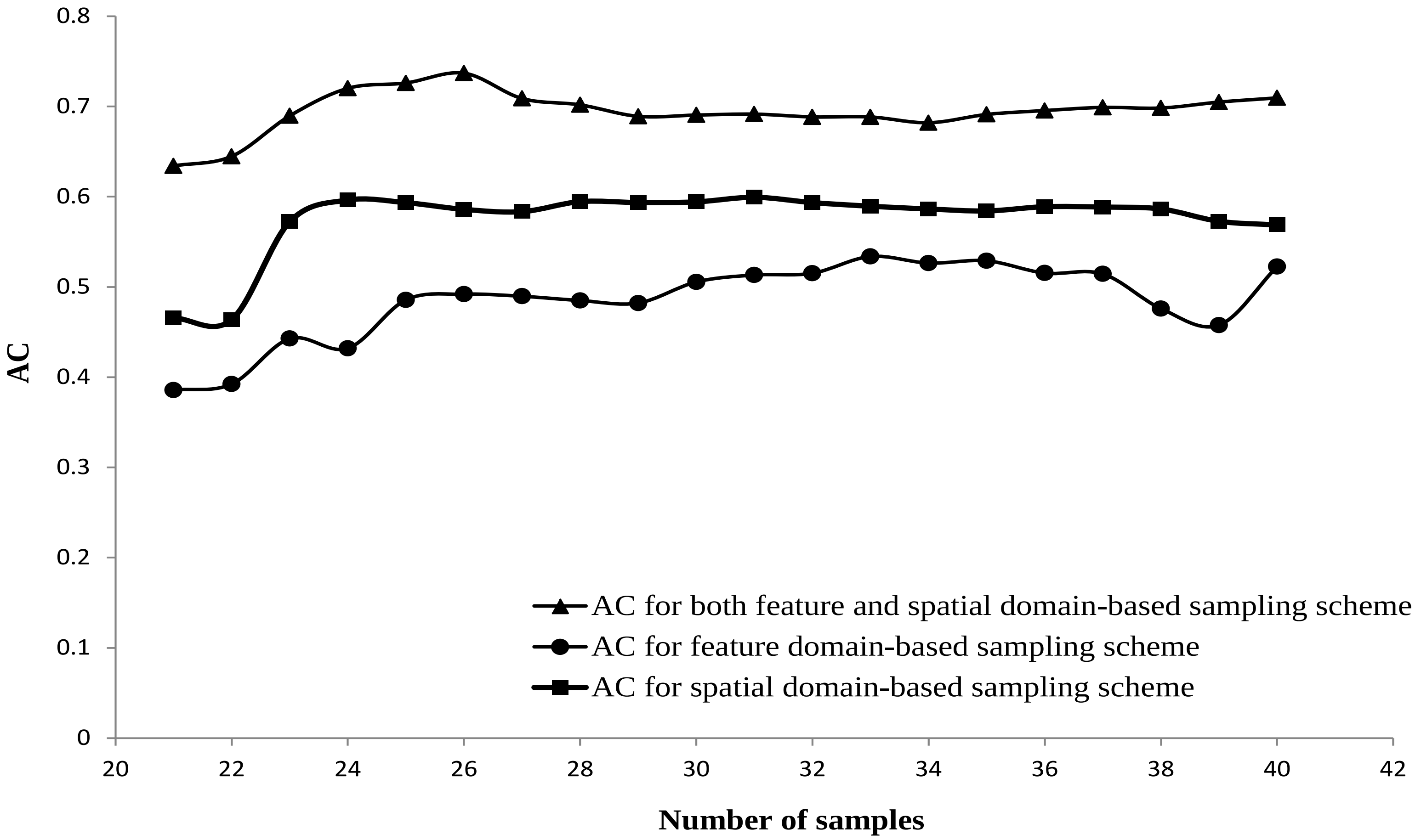
Two soil mapping methods, iPSM and ordinary kriging, were used to produce the prediction uncertainty maps of SOM

42 independent validation points were used to assess the accuracy of the predicted maps.

Results – RMSE



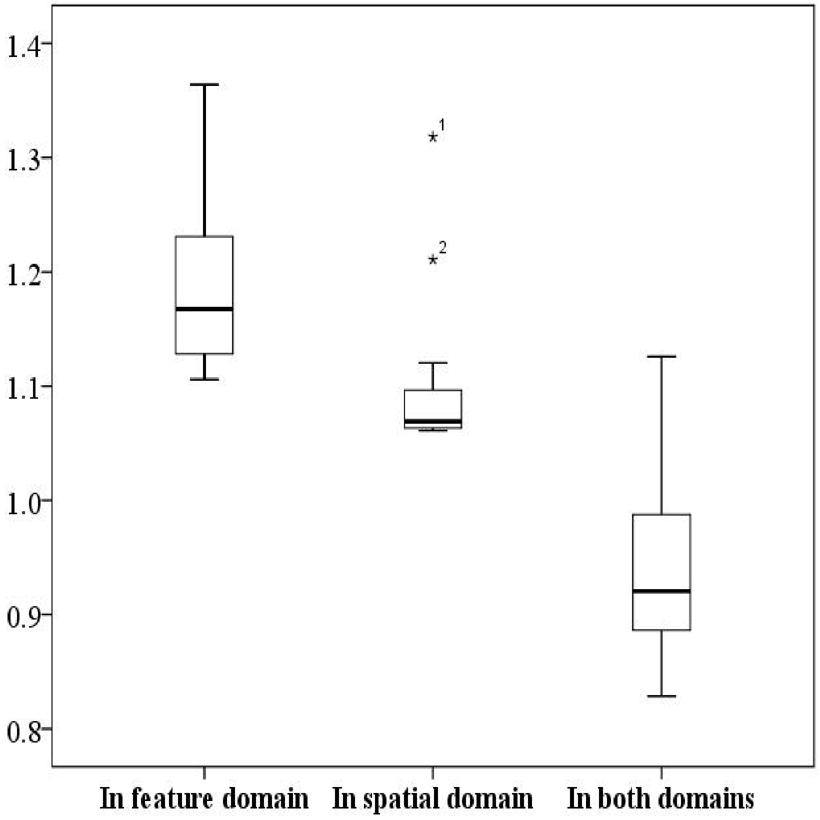
Results – Agreement of Coefficient (AC)

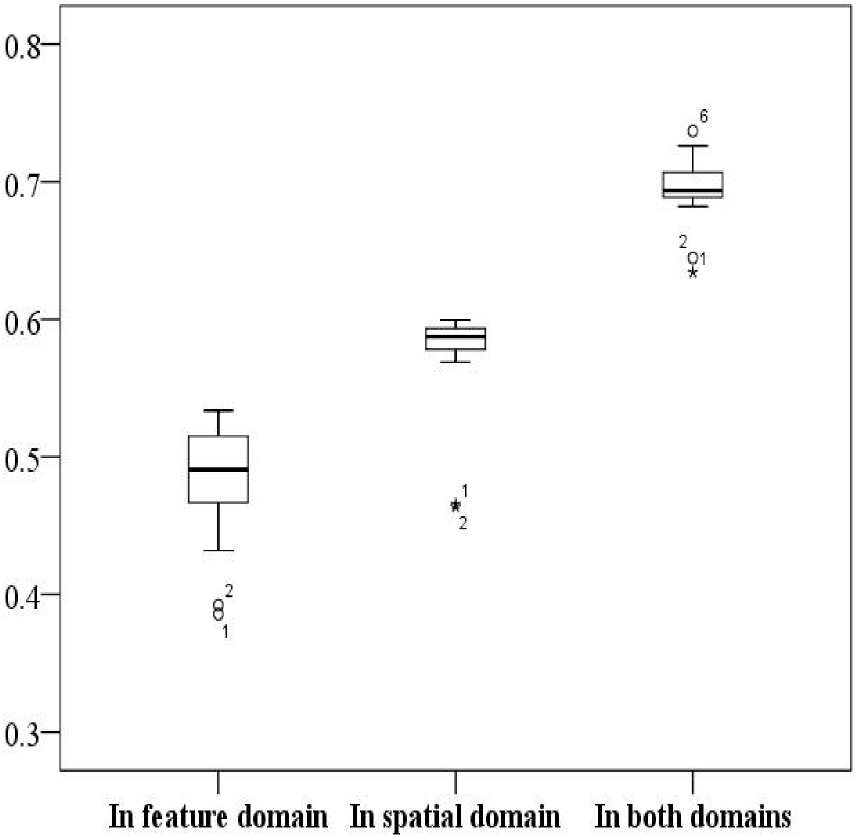


Results – Repetition

To further validate the performance of the sampling design the experiment was repeated 5 times with each having different set of starting samples (existing samples) by selecting randomly existing samples from the 50 original samples.

Results – Repetition





Box plot of the average of RMSE (left) and AC (right) between measured and kriging predicted SOM through repeating the experiment 5 times

Conclusion

Sampling based on both uncertainties is more "efficient" than based on either one uncertainty.

Should it be the way to sample for more in the future?

Thank you!

A-Xing Zhu
azhu@wisc.edu