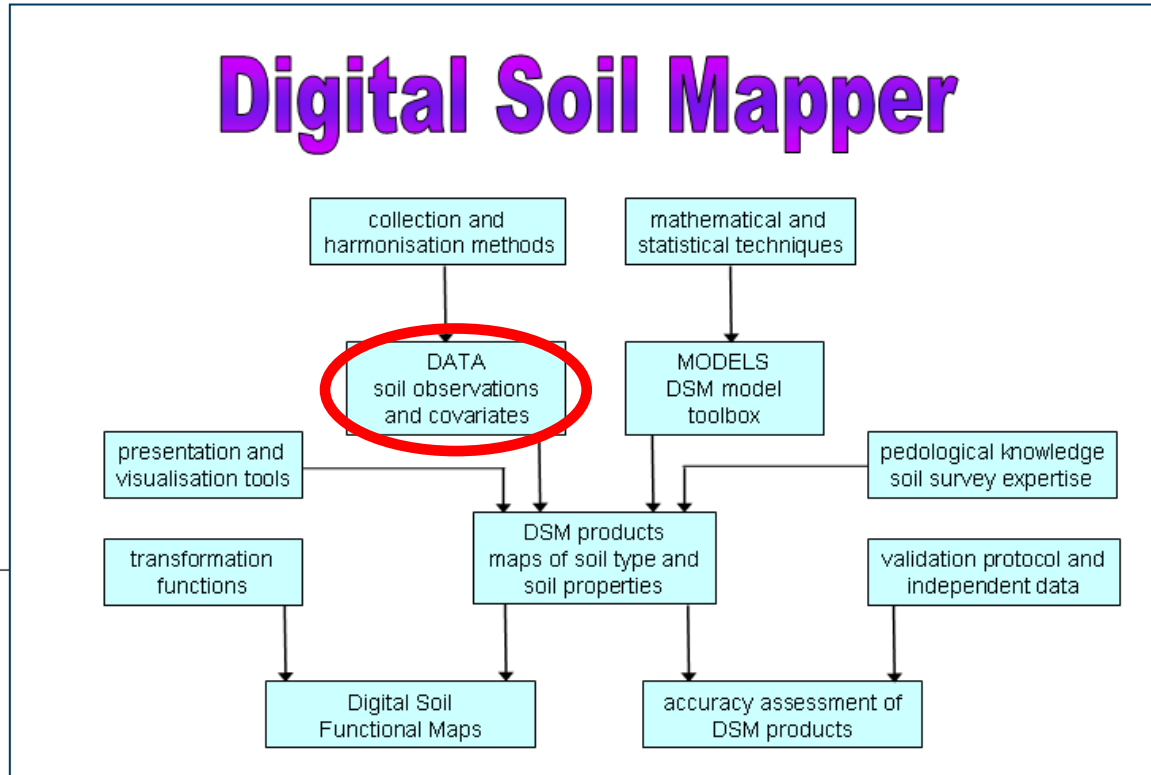
Digital soil mapping using data with different accuracy levels

ISRIC World Soil Information

Gerard Heuvelink, Dick Brus, Tom Hengl, Bas Kempen, Johan Leenaars and Maria Ruiperez-Gonzalez

Soil profile data are key to DSM

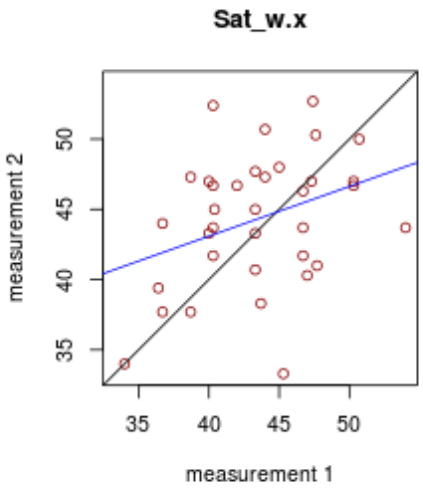
- They are used to calibrate soil prediction models that predict soil properties from covariates
- They are used to condition predictions to nearby observations using kriging

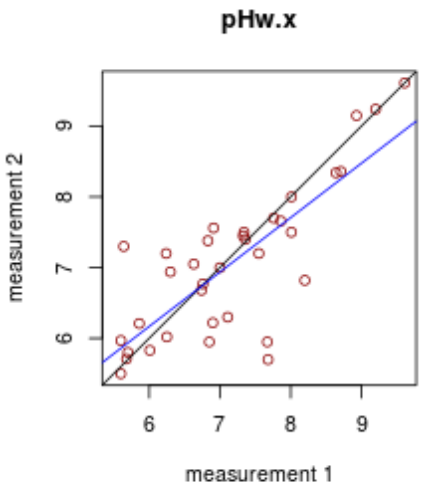


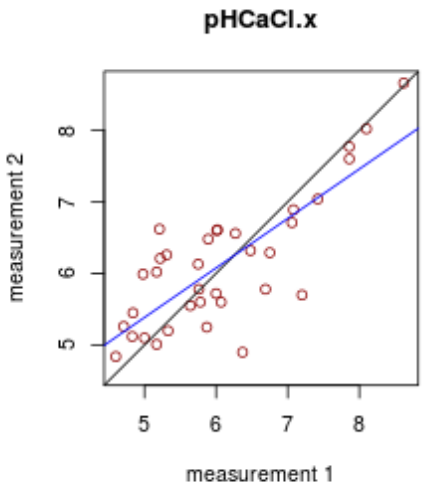
But soil profile data are not without error

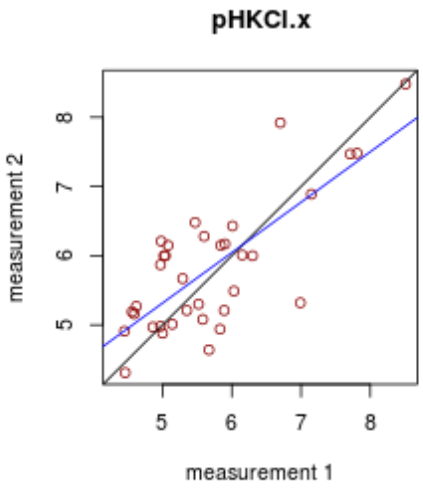
- Lab data can have substantial measurement errors
- There are quality differences between labs
- Field data ('educated guesses') tend to be less accurate than lab data
- Many soil 'observations' are measured indirectly (such as through soil spectroscopy, e.g. PLSR)
- Some of us make use of pseudo-observations
- In future the use of volunteered soil information (crowd-sourcing) will grow, but these data may not be very reliable

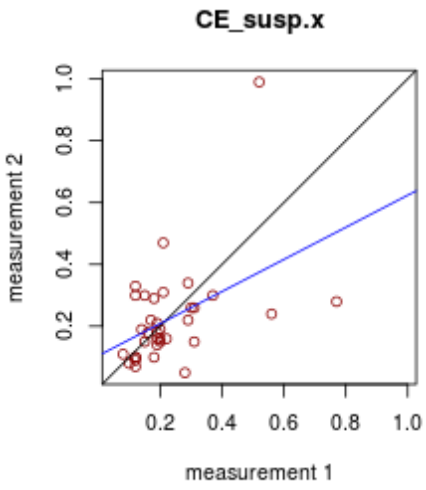
Quality of lab data: look and shudder

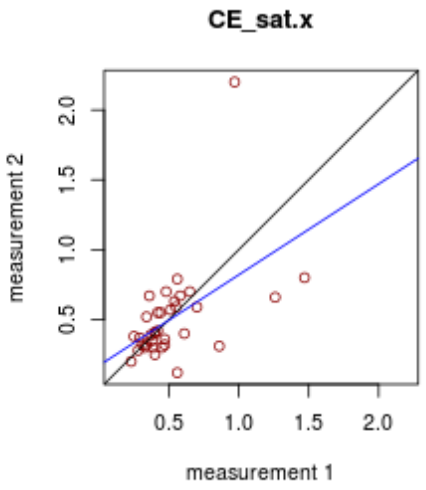












The geostatistical approach can account for measurement errors (KED case)

$$Z(s) = m(s) + \varepsilon(s) = \sum_{j=0}^{p} \beta_j \cdot x_j(s) + \varepsilon(s)$$

$$Y(s_i) = Z(s_i) + \delta(s_i)$$

measurement errors with mean vector μ and variance-covariance matrix V

$$\hat{Z}(s_0) = E[Z(s_0)|Y(s_i) = y(s_i), i = 1 \cdots n]$$

Result that goes back to Delhomme (1978)

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

$$Var(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = (\boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{X})^{-1}$$

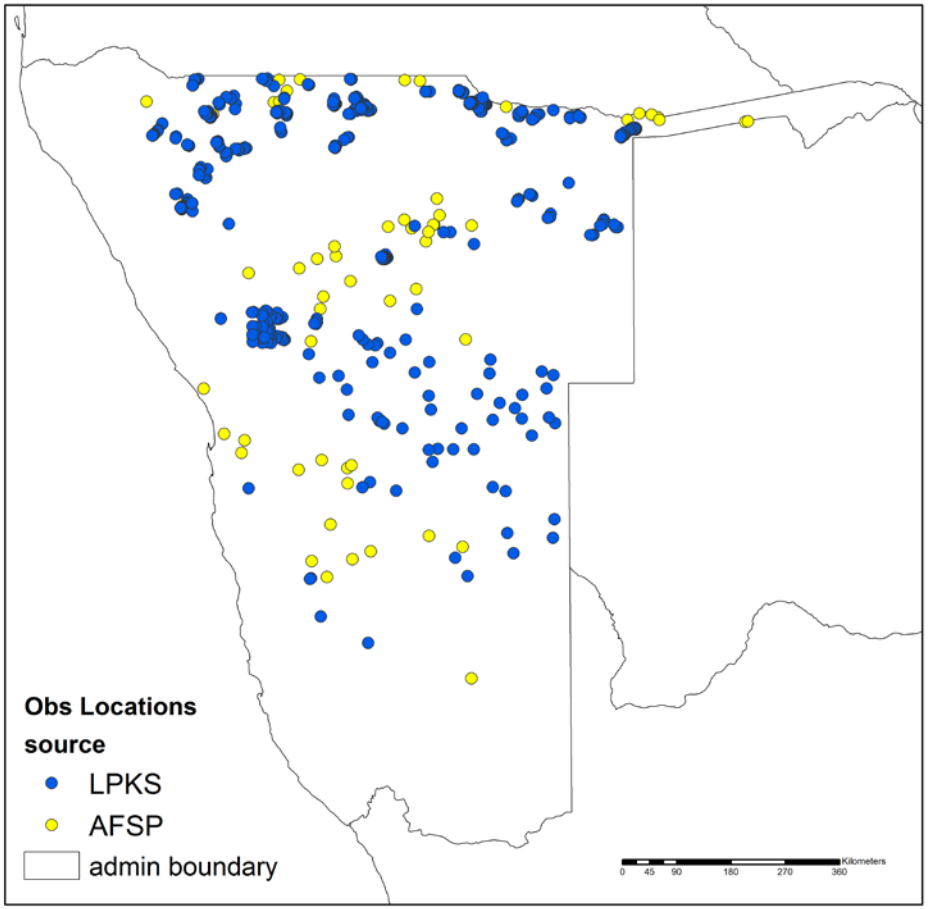
$$\hat{Z}(s_0) = (\boldsymbol{x}_0 + \boldsymbol{X}(\boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{c}_0))^T \cdot (\boldsymbol{C}+\boldsymbol{V})^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$$

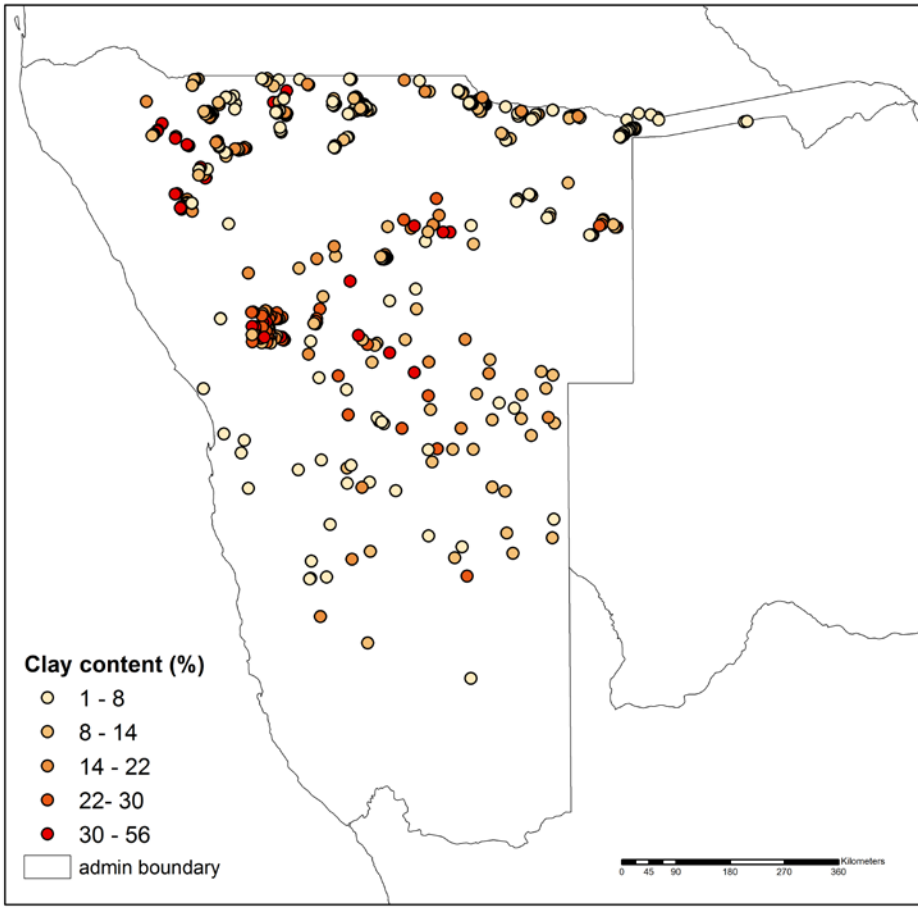
$$Var\left(\hat{Z}(s_0) - Z(s_0)\right) = c_{00} - \boldsymbol{c}_0^T \cdot (\boldsymbol{C}+\boldsymbol{V})^{-1} \cdot \boldsymbol{c}_0 + (\boldsymbol{x}_0 - \boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{c}_0)^T(\boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{X})^{-1} \cdot (\boldsymbol{x}_0 - \boldsymbol{X}^T(\boldsymbol{C}+\boldsymbol{V})^{-1}\boldsymbol{c}_0)$$

Example: mapping topsoil clay content for Namibia

LPKS = LandPKS database, field estimates of soil texture (by texture class)

AFSP = Africa Soil Profiles database (merge of numerous legacy soil datasets)

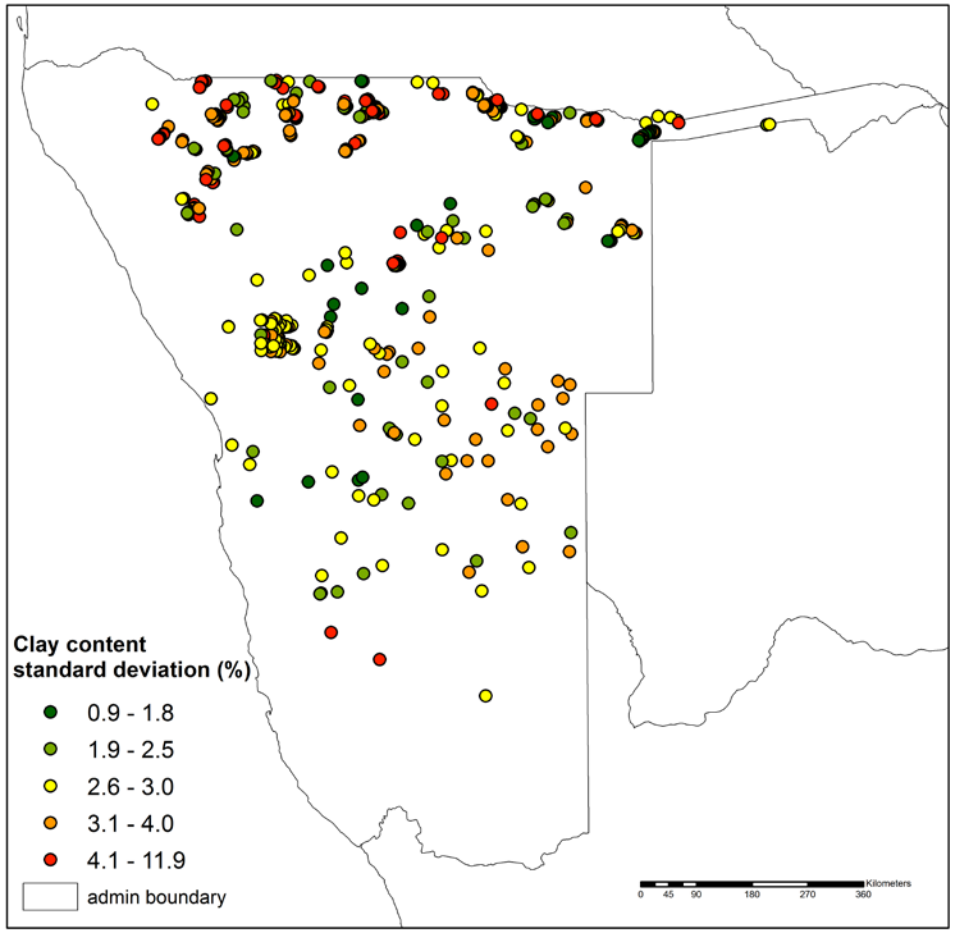




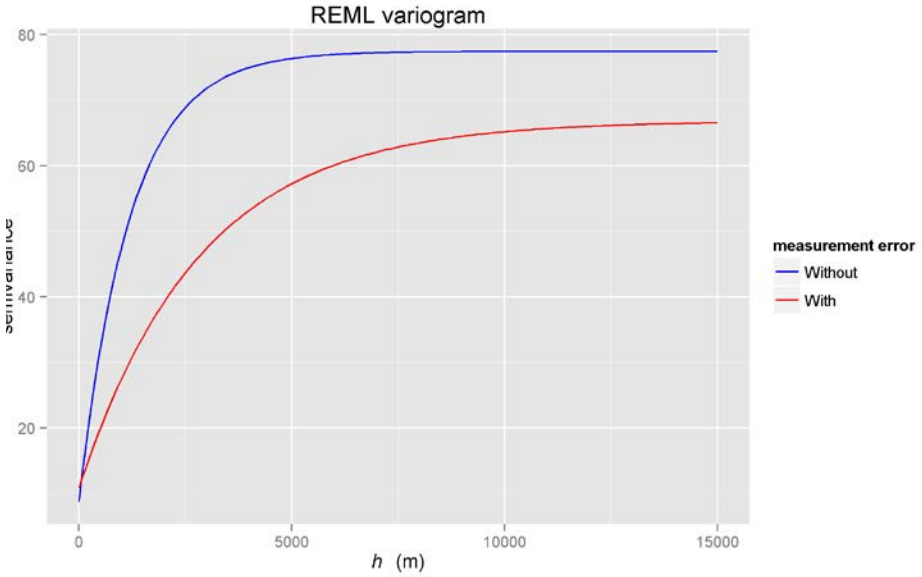
AFSP data four accuracy levels (depending on source credibility)

LPKS data accuracy based on variability within soil texture class (GSIF TT2tri function in R)

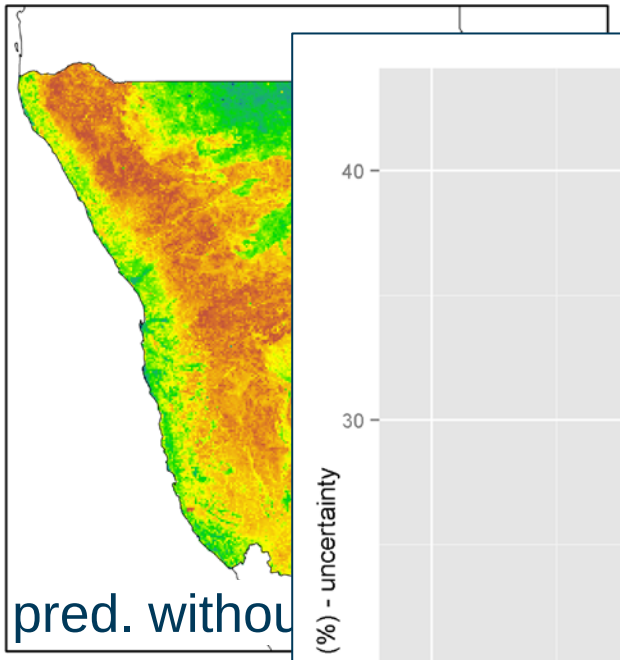
clay sd ranges from 0.9% to 11.9%



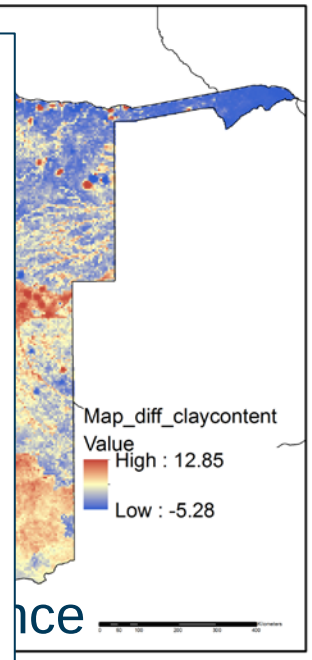
	Beta_GLS (without uncertainty)	Beta_GLS (including uncertainty)
Intercept	8.596	8.835
ASSDAC3	-0.063	-0.057
VBFMRG5	-0.0075	-0.0073
T03MSD3	0.227	0.208



Resulting maps: meaningful differences



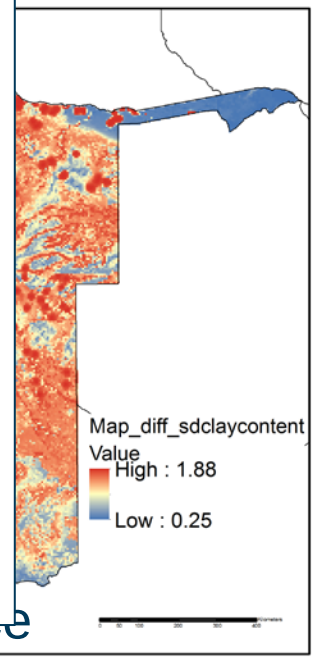
pred. without



nce

sd without u

sd with unc.



sd difference

This was just an illustrative example, possible extensions:

- Include (known) systematic differences between sub-datasets
- Include unknown systematic differences by representing these as random errors that are perfectly correlated within a sub-dataset
- Include serial correlation between errors (instrument drift, anchoring effect)
- Take different soil data accuracy levels into account for variogram estimation (including its uncertainty, using a Markov chain Monte Carlo approach)
- Estimate part of the measurement error parameters (i.e. elements of μ and V) from the data

Also important and challenging:

- How to include differences in soil data accuracies in machine-learning algorithms?
- One possibility is to assign weights, but how large should these weights be?

Published in final edited form as:
Stat Anal Data Min. 2013 December 1; 6(6): 496–505. doi:10.1002/sam.11196.

A Weighted Random Forests Approach to Improve Predictive Performance

Stacey J Winham[1,§], **Robert R Freimuth**[1], and **Joanna M Biernacka**[1,2,§]
[1]Department of Health Sciences Research, Mayo Clinic, 200 First Street Southwest, Rochester, MN 55905 USA

Concluding remarks

- Soil measurements are not error-free
- Measurement error can be taken into account in DSM, so why don't we do it?
- Perhaps it is because often we do not know how accurate the soil measurements are?
- But this is not true for data such as derived from soil spectroscopy, and why don't we routinely send replicates to the laboratory (without telling the lab)?
- We can do so much better. And we should. Is there anyone in this room who has not wasted valuable time on trying to fit models to 'poor' (rubbish) data?

Submission of pre-conference workshop proposals
1 November, 2016

Abstract submission deadline
1 February, 2017

Early-bird conference registration
1 April, 2017

Submission of special issue manuscripts
1 October, 2017

Evolution
Proximal Soil Sensing
Soil Monitoring

Please join us in Wageningen, The Netherlands!

26 June – 1 July, 2017

Come share and learn about the latest developments in the field and join the festive 25th anniversary of pedometrics.

Stay up to date at
www.pedometrics2017.org

SoilCares

ISRIC World Soil Information

International Union of Soil Sciences

WAGENINGEN UR

ΠΕΔΟ METRICS

PEDOMETRICS.ORG