Prediction of soil properties with NIR data and site descriptors using preprocessing and neural networks

Matt Aitkenhead
Malcolm Coull
Jean Robertson

The James Hutton Institute

Introduction to NSIS

- A component of the Scottish Soils Database
  - One of the most detailed and systematic collections of national soil data in Europe.
  - Soil Survey of Scotland produced a range of digitised and paper maps at a number of scales from full national coverage at 1:250000 scale to more local surveys at scales of 1:10560 or larger.
  - Comprehensive database was developed that currently contains chemical and physical information on over 13000 georeferenced soil profiles.
- The National Soils Inventory for Scotland (NSIS) is an objective sample of Scottish soils.
  - Soil and site conditions of 183 locations throughout Scotland were sampled using a 20km grid across the entire country (NSIS 2).
  - Samples taken at multiple depths from soil pits and analysed to determine their physical and chemical properties (approx. 800 datasets)

NSIS data

VIS-NIR spectra (pre-processed)	Temperature (12 monthly means)
Topography (8 parameters)	Rainfall (12 monthly means)
Land cover (10 parameters)	Geology (19 classes)
Soil (9 classes)	

inputs…

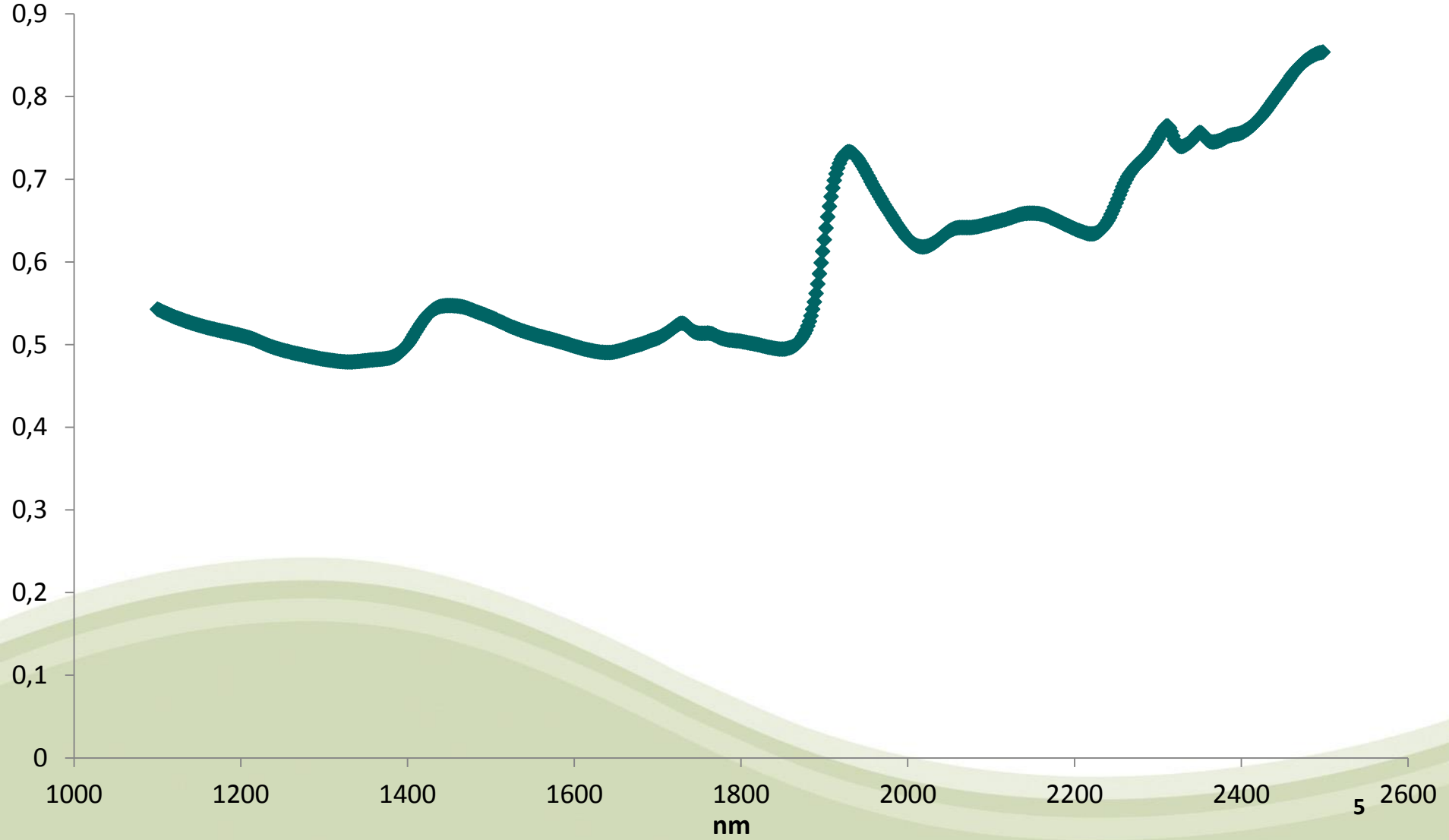
…and outputs

Ag (aqua-regia digestion, ppm)	Cd (aqua-regia digestion, ppm)	K (exchangeable, meq per 100g)	Mo (aqua-regia digestion, ppm)	pH (in H_2O)
Al (exchangeable, meq per 100g)	Co (aqua-regia digestion, ppm)	K (aqua-regia digestion, ppm)	H_2O loss (105°C)	Pt (aqua-regia digestion, ppm)
Al (aqua-regia digestion, ppm)	Cr (aqua-regia digestion, ppm)	LOI (loss on ignition, 450°C)	Na (exchangeable, meq per 100g)	S (aqua-regia digestion, ppm)
As (aqua-regia digestion, ppm)	Cu (aqua-regia digestion, ppm)	LOI (loss on ignition, 900°C)	Na (aqua-regia digestion, ppm)	Se (aqua-regia digestion, ppm)
B (aqua-regia digestion, ppm)	H (exchangeable, meq per 100g)	Mg (exchangeable, meq per 100g)	Ni (aqua-regia digestion, ppm)	Sr (aqua-regia digestion, ppm)
Ba (aqua-regia digestion, ppm)	Fe (exchangeable, meq per 100g)	Mg (aqua-regia digestion, ppm)	P (aqua-regia digestion, ppm)	Ti (aqua-regia digestion, ppm)
Ca (exchangeable, meq per 100g)	Fe (aqua-regia digestion, ppm)	Mn (EDTA extraction, ppm)	Pb (aqua-regia digestion, ppm)	P (total, derived from P_2O_5 ppm)
Ca (aqua-regia digestion, ppm)	Hg (aqua-regia digestion, ppm)	Mn (aqua-regia digestion, ppm)	pH (in $CaCl_2$)	Zn (aqua-regia digestion, ppm)

NIR data - introduction

- WinISI software used for on-board analysis of spectra

NIR data - example



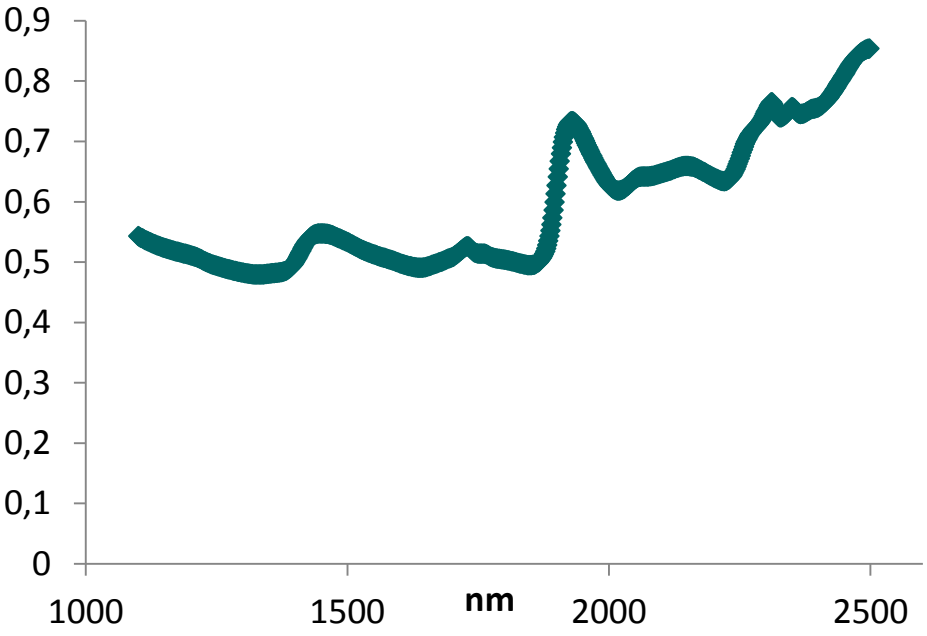
Experimental design

- Multiple steps, based on on-going NIR/FTIR soil work:
  - Moving window transform
  - Derivative transform
  - Normalisation
  - Input subsampling
  - Neural network layer size
- 3600 combinations explored:
  - Moving window/derivative transform first
  - Moving window size of 5, 10, 20, 50, 100
  - Derivative transform options of (1) no derivative, (2) 1st derivative, (3) 2nd derivative, (4) Savitsky-Golay 0-order, (5) S-V 1st order, (6) S-V 2nd order
  - Spectral normalisation over either entire range of values, or by min/max for each spectrum
  - Dataset subsampling rate of 1, 2, 5, 10, 20 or 50
  - NN hidden layer sizes of 5, 10, 20, 50 or 100

Moving window/smoothing

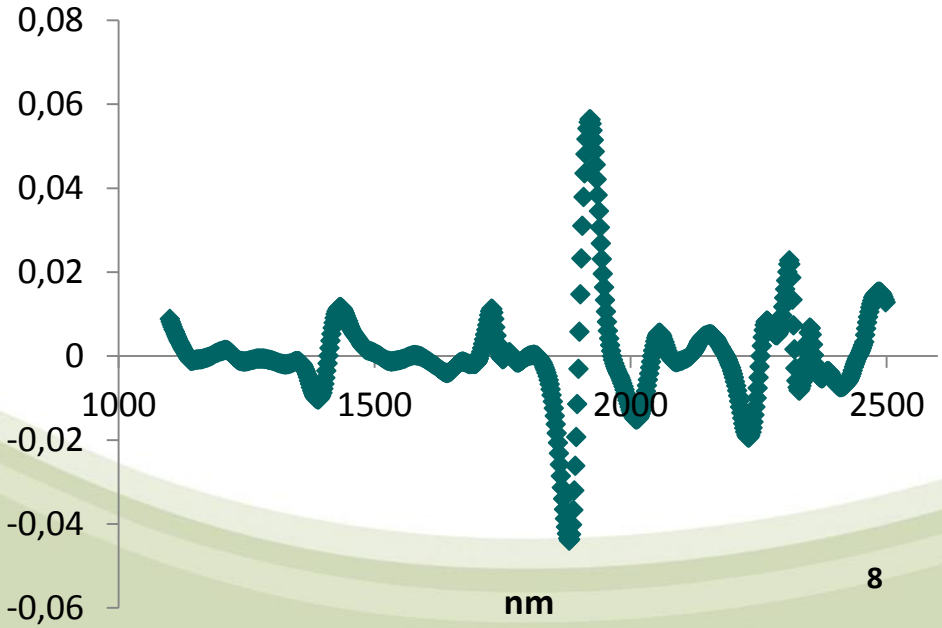
- Smoothing/derivation of the spectra prior to interpretation is common
  - Reduces noise and accentuates useful data
- Many different smoothing/derivative functions exist
- Using a ‘moving window’ subtraction makes peaks stand out from their surroundings
- Which should be chosen?
  - Moving window – what radius of window?
  - Smoothing/derivative – what function?
  - Both? In what order?

Moving window

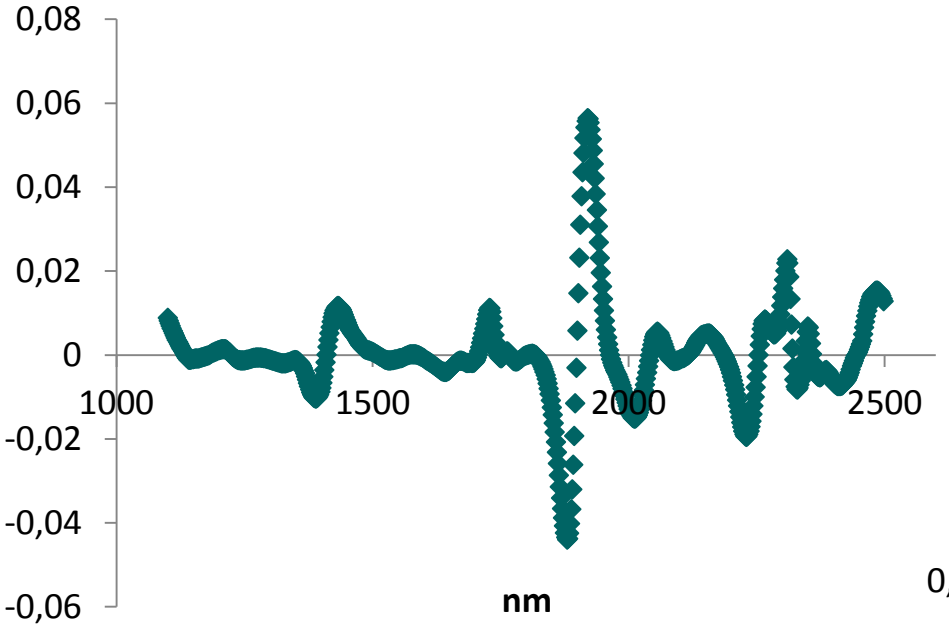


Before

After
(window radius 20)

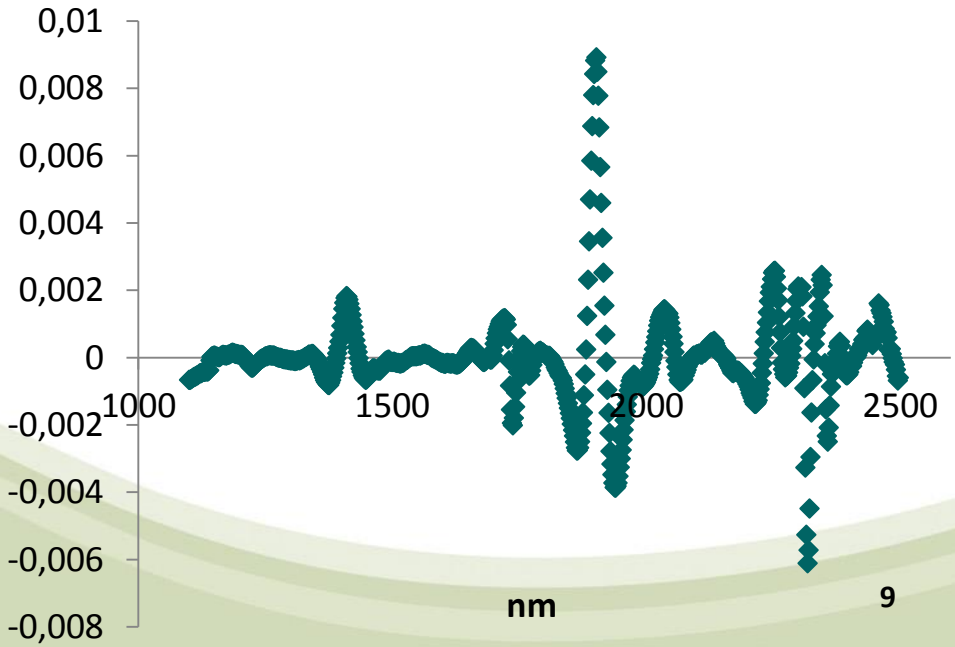


Smoothing/derivation

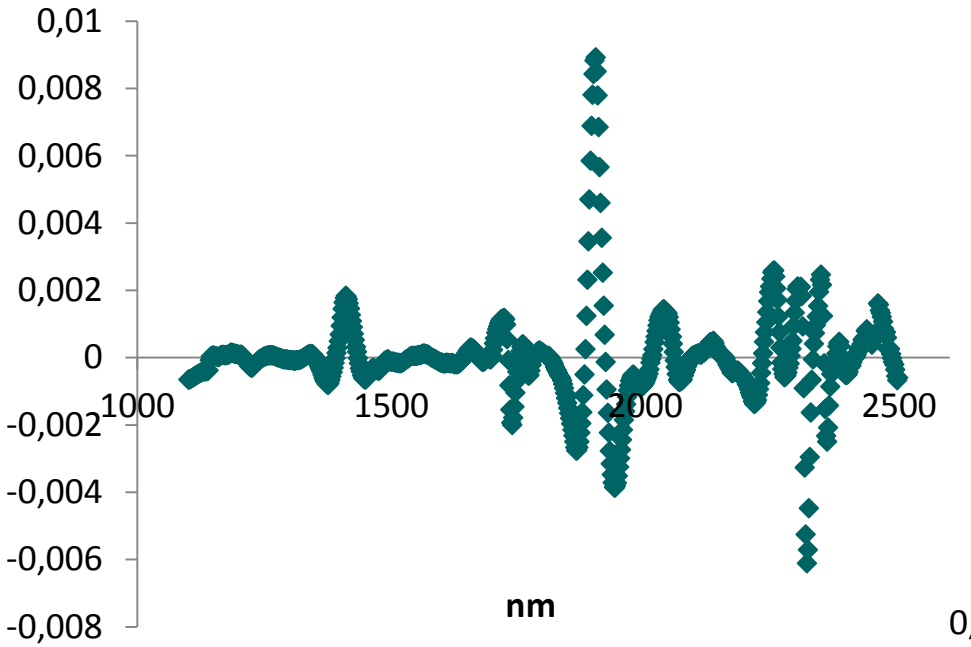


Before

After
(1st order derivative)

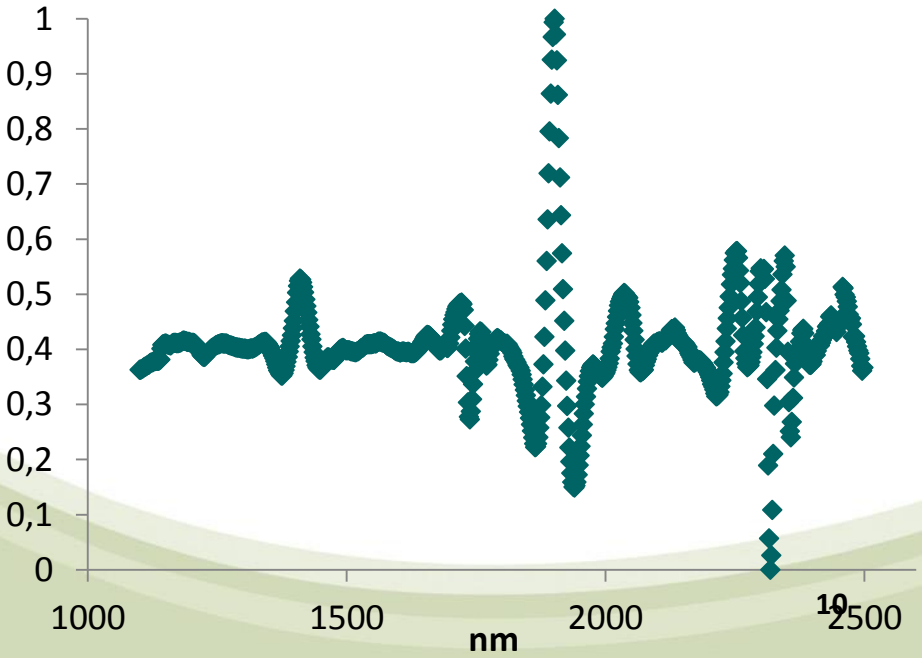


Normalisation

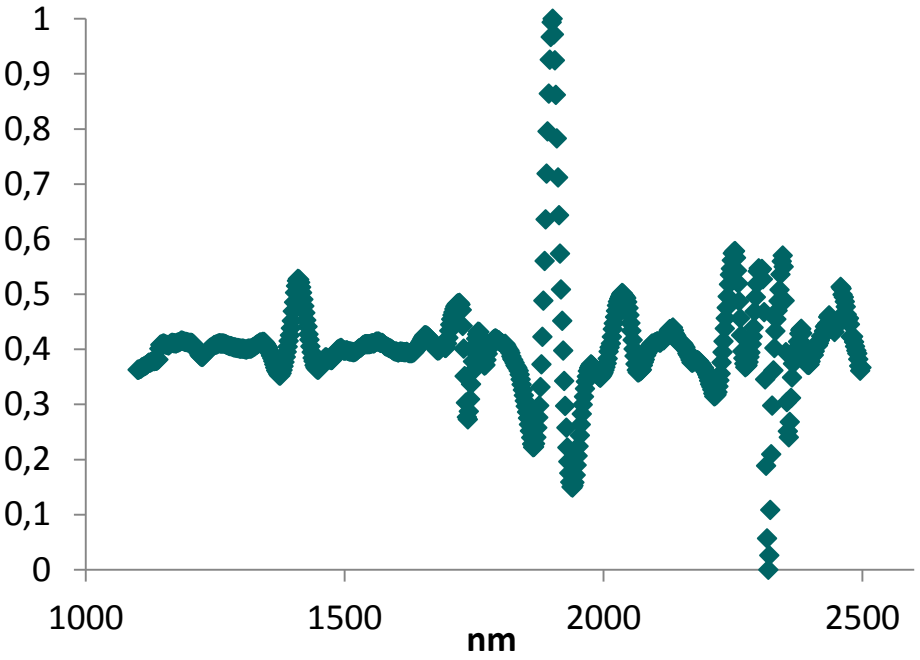


Before

After



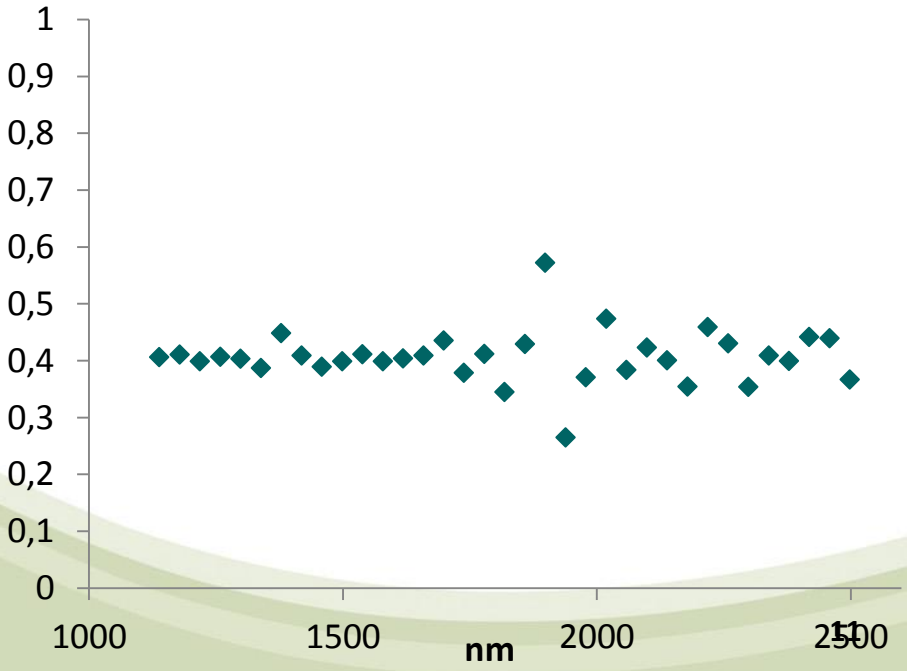
Sampling



The James Hutton Institute

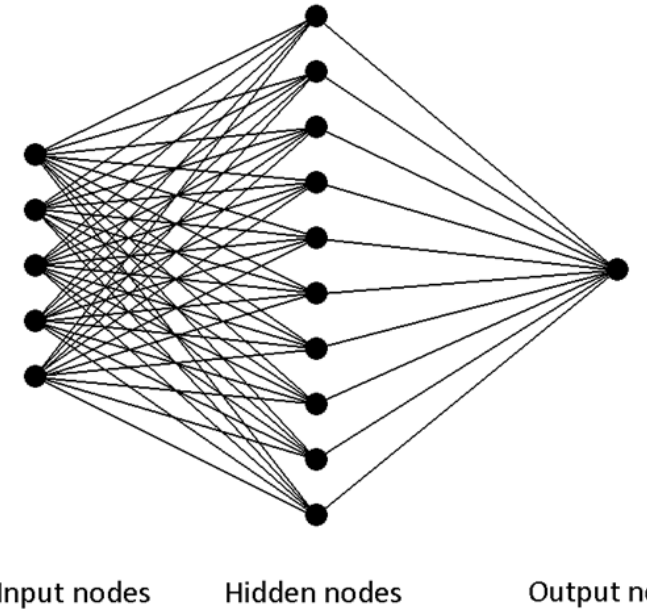
Before

After



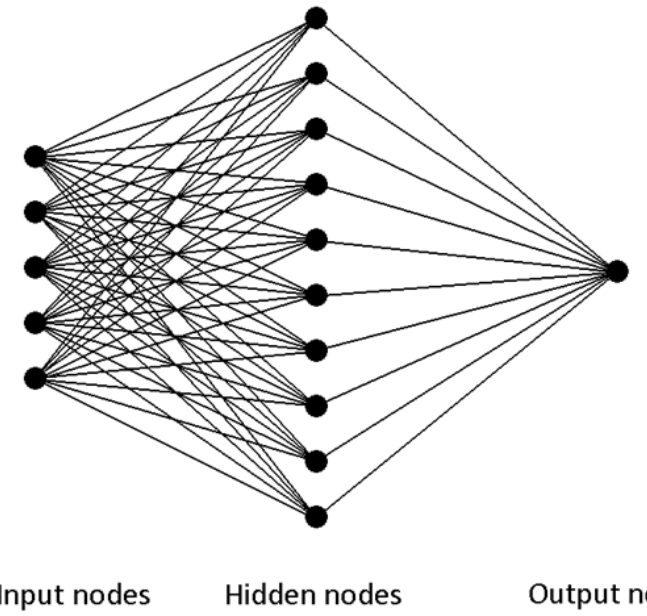
Neural network modelling

- Simplified model of biological learning
  - Useful for large, ‘messy’ datasets
  - Can handle large numbers of input and output parameters
- Backpropagation training method
  - Relatively old, simple NN approach
  - Based on error minimisation
  - Standard for data mining/modelling
  - Allows the ‘black box’ to be opened



Neural network design/training

- One input node for each value in the pre-processed spectrum (700)
- Additional nodes can be added if other input data is to be used
- Two hidden layers of 100 nodes each
- One output node for each of the output parameters (40)
- Dataset split into training/testing (75/25) at random
- Testing at every 1000 training steps to find optimal network

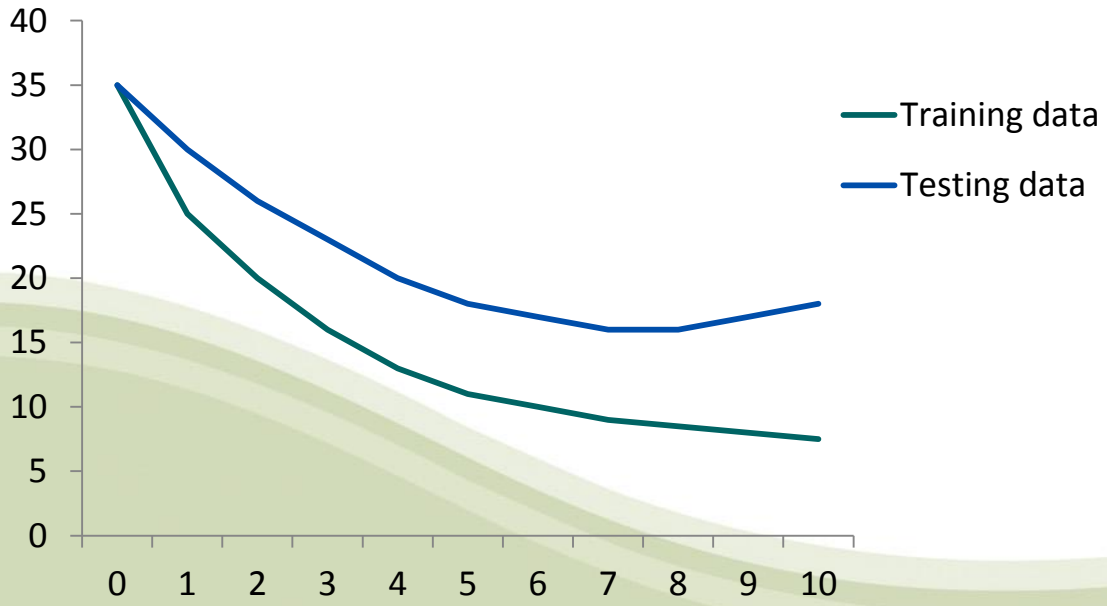


Statistical evaluation

- Statistics of predictive accuracy:
  - R-squared
  - RMSE
  - MAE
  - ME
- Weighting of network input/output relationships
  - Partial derivatives method (Olden & Jackson, 2002; Olden et al., 2004)
  - Looks at the relationships between every input/output parameter combination

Variation in results

- Neural network can underfit or overfit the data
  - Underfitting if not sufficiently trained
  - Overfitting if trained too well on the training data
  - Need to identify ‘stopping point’
  - Testing data (separate from training data) used for this

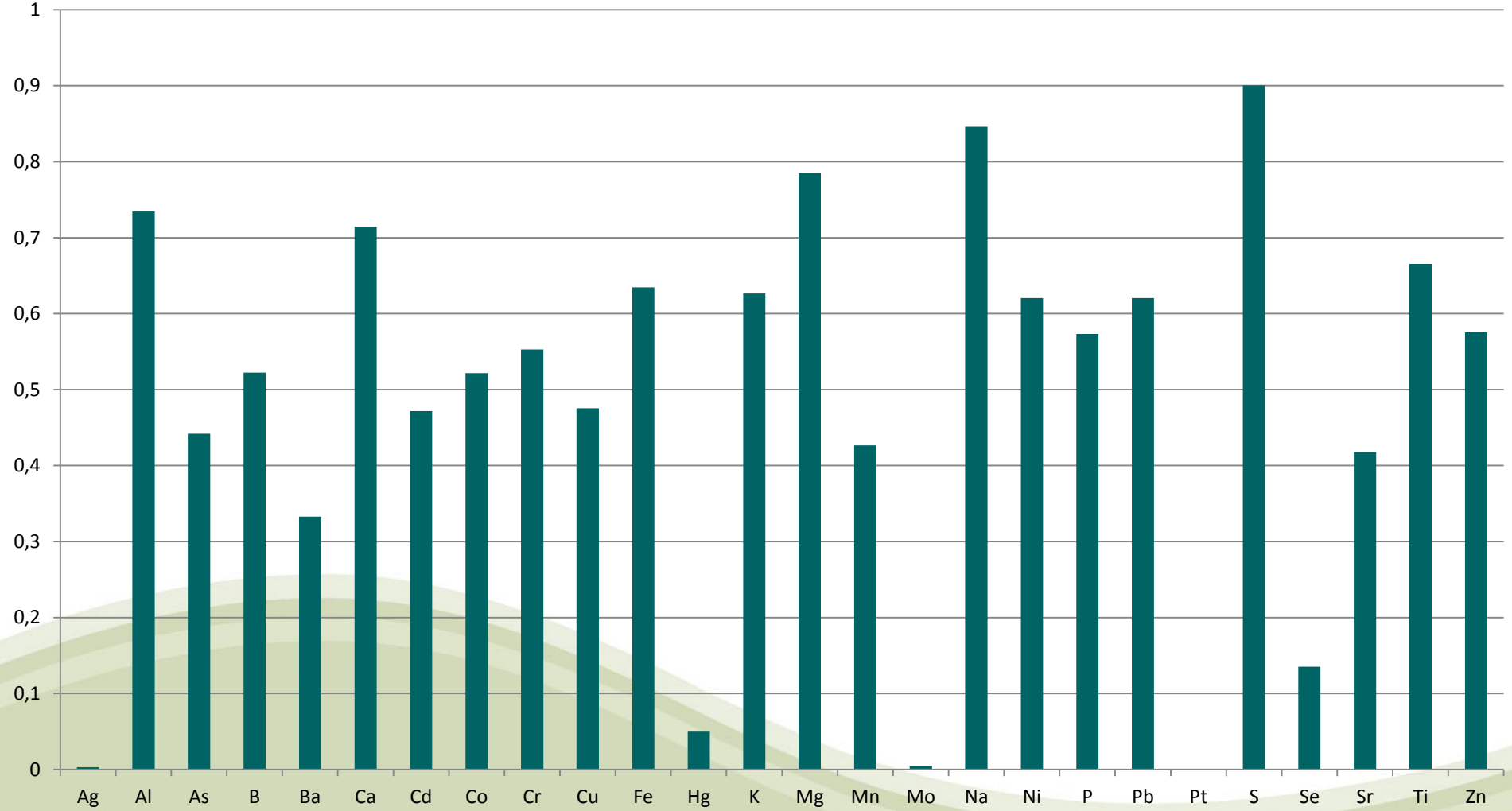


Best preprocessing algorithm

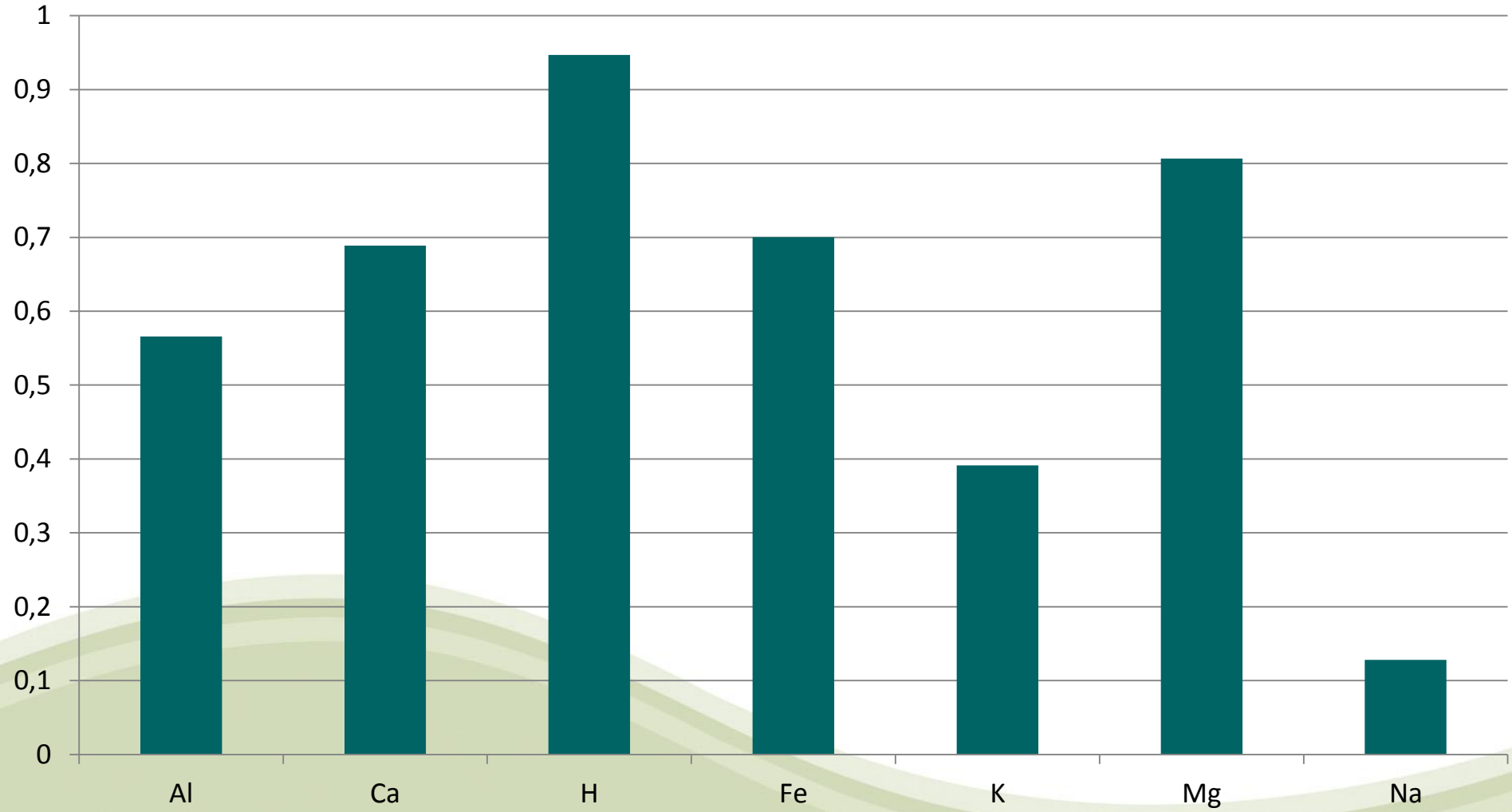
The James Hutton Institute

- Moving window first, with window radius 50
- Then 1st-order Savitsky-Golay smoothing
- Normalisation by min-max range for each spectrum
- Minimise data subsampling (no subsampling at all is best)
- Maximise NN hidden layer size (100 was largest used)
- Demonstrable variation in results between experimental combinations:
  - All statistical measures varied greatly between worst and best combinations
  - Trends seen in subsampling & NN hidden layer size effects

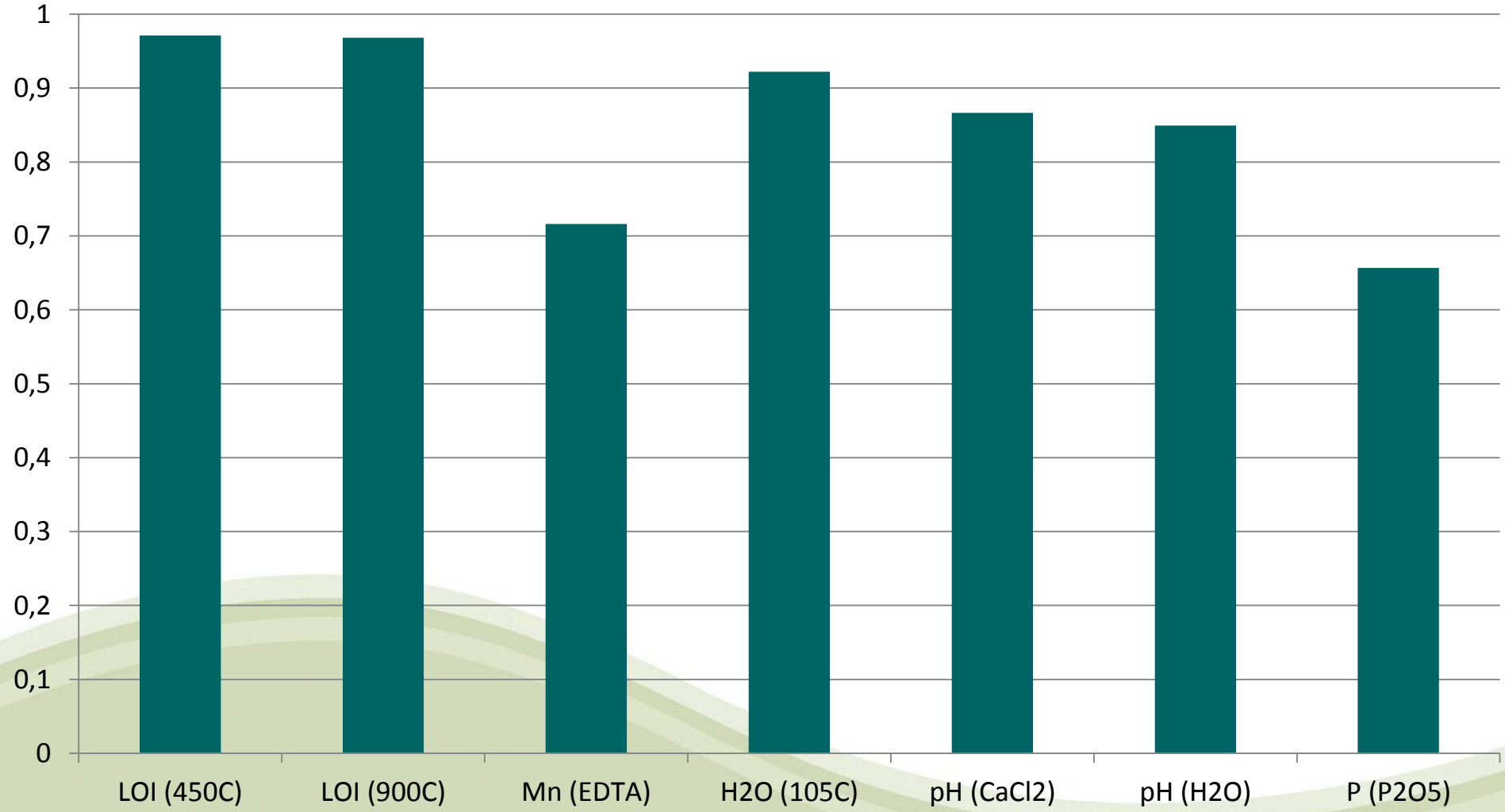
Best results (aqua regia r-squared)



Best results (exchangeable r-squared)



Best results (other r-squared)



Site characterisation

- Environmental factors influence the character of the soil
  - Topography
  - Vegetation
  - Climate
  - Geology
- Sample locations were recorded to within 10m accuracy (in most cases!)
- With sufficiently large dataset, can be used to develop an 'environment-specific' calibration of the model
- NN approach is sufficiently flexible to incorporate this information 'automatically'

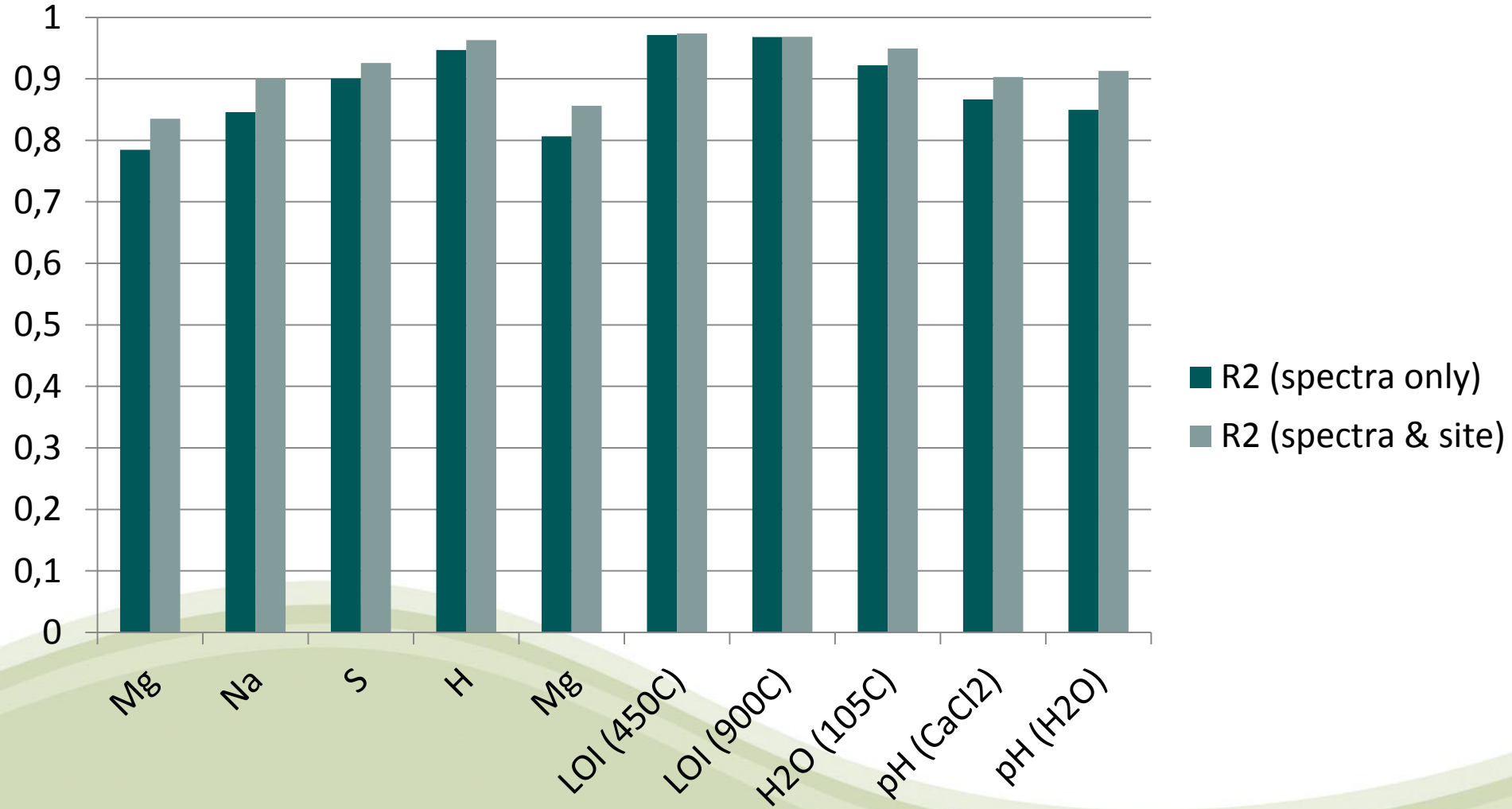
Inclusion of site character

- 8 extra input parameters for topography
  - Elevation, slope, curvature, curve-plan, curve-profile, aspect, aspect-east, aspect-north
- 20 extra input parameters for vegetation
  - 10 classes for each of 2 land cover maps (LCS88 & LCM2007)
  - Cropland, improved grassland, rough grassland, deciduous, coniferous, peat, heath, bare, water, montane
- 9 extra input parameters for soil
  - Alluvial, alpine, bare, brown earth, gley, peat, podzol, lithosol, regosol
- 24 extra input parameters for climate
  - Monthly means for temperature and rainfall
- 19 extra input parameters for geology
  - Derived from geological information produced during soil survey work (Lilly, Towers and others)

Modelling with all of the data

- 80 extra input nodes for 80 extra input parameters
- Identical training regime
- Identical NN architecture
- Sensitivity analysis to identify important input parameters (spectroscopy inputs included in this)
- Site characterisation derived from existing spatial datasets, all adjusted to 100m resolution

Changes in the results

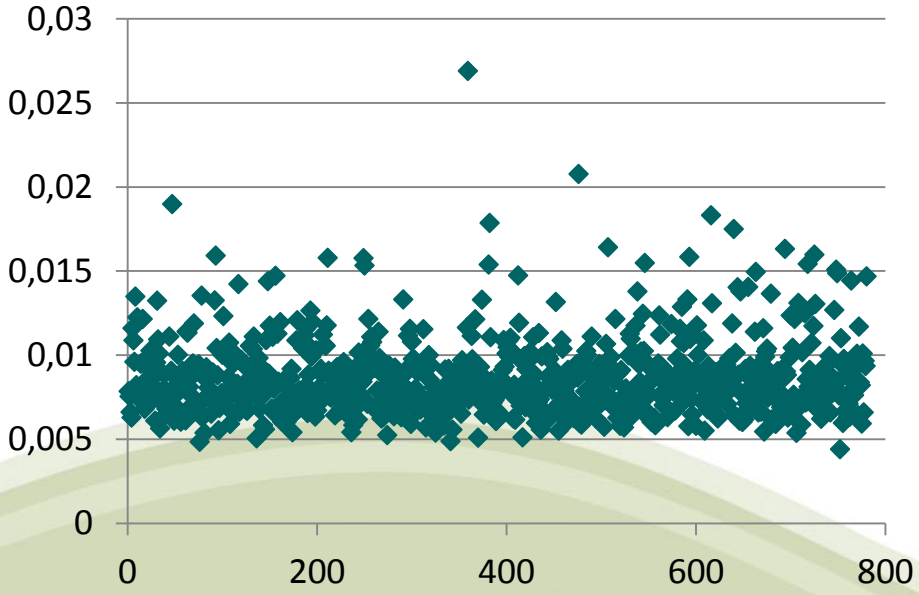


Specific parameter (LOI450)

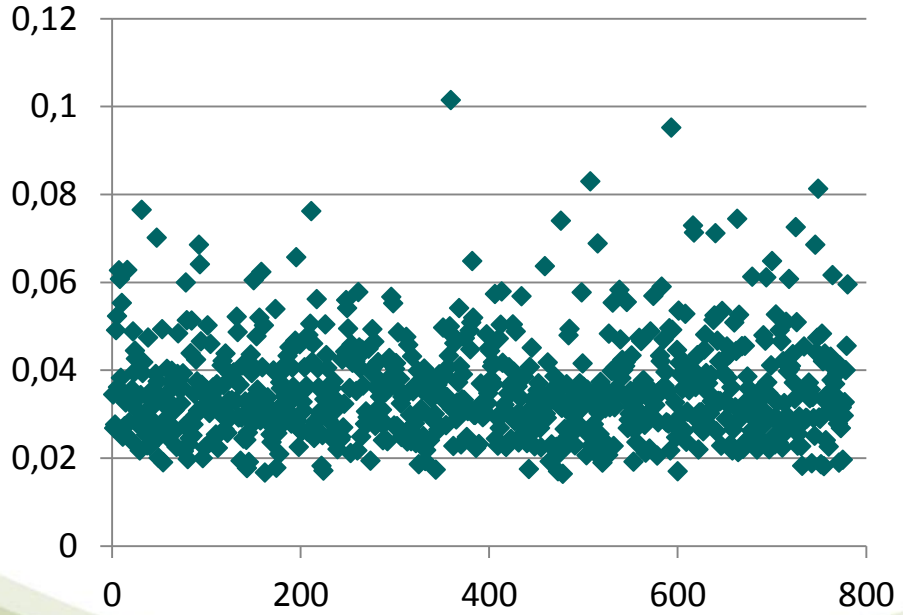
- Almost suspiciously good!
  - So I went back and checked
  - R-squared (all inputs) of 0.974
  - Accurate within 1% of LOI >90% of the time for LOI < 20%
  - Can still be out by up to 4% in this range...
  - Accuracy better at low and high LOI values, slightly worse in the middle range
  - Overall RMSE: 0.046
  - Overall MAE: 0.035
  - Overall ME: 0.001

Sensitivity analysis

- Several inputs in the spectra/environmental data have relatively high mean absolute or maximum weightings
  - No clear pattern or clustering of 'important' inputs'
  - Environmental inputs no more important than spectra



Mean weighting



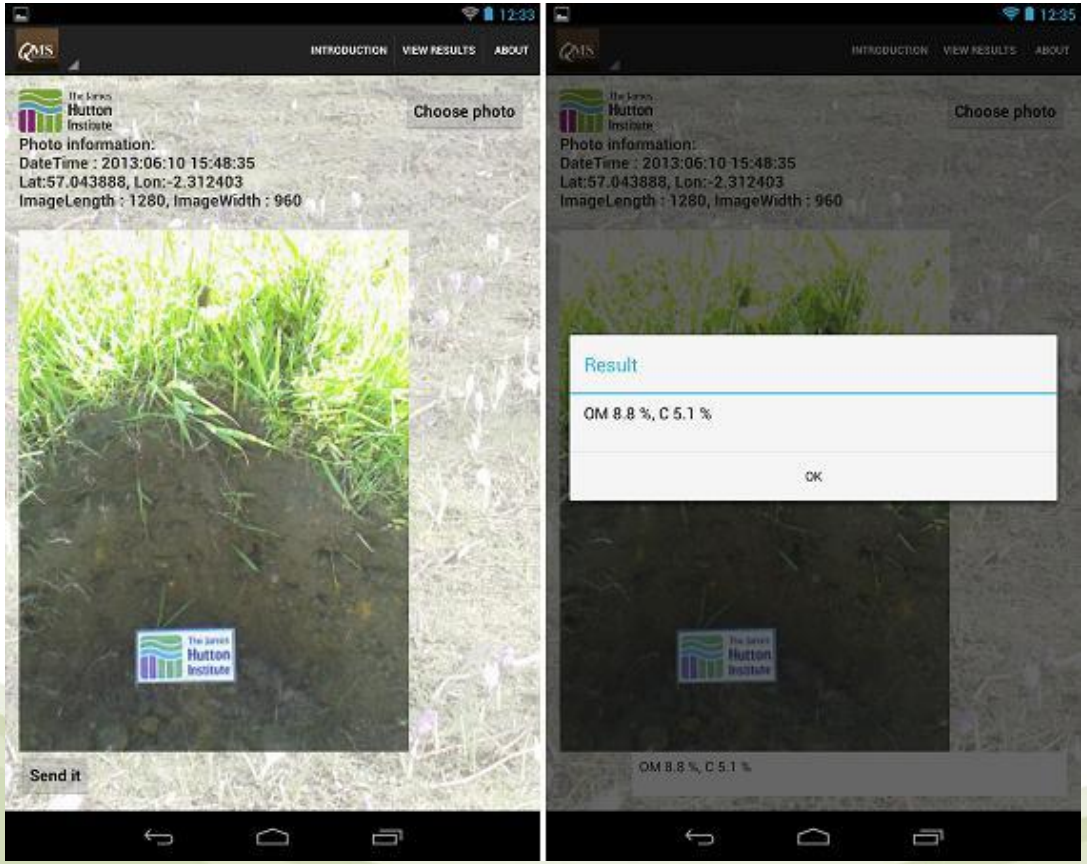
Maximum weighting

Ongoing and future work

- Redo the sensitivity analysis using other approaches
  - Current sensitivity analysis is noisy, tells us less than it could
- Comparison of prediction accuracies with standard approaches
  - Jean Robertson's analysis for matching these soil samples
  - Literature, for wider comparison
- Local calibration – automated real-time stratification based on site characteristics (real-time model training, testing)
- LUCAS data analysis (for the future!)
  - Similar approach as described here
  - Need to develop site descriptor data (topography, climate, vegetation, geology, soil type)

A potential side-route?

- SOCIT mobile phone app (iPhone/Android)
- Estimates soil OM and soil C using mobile phone imagery & site descriptors
- LUCAS spectroscopy could be used to produce RGB estimates
- A soil C estimation app for Europe?



Conclusions

- Some soil parameters can be predicted ‘well’ using NIR data
  - Depends on your definition of ‘well predicted’
  - Mg, Na, S, Ti, H, Fe, Mn, LOI, H2O, pH all above r2 of 0.75
  - C (0.94) , N (0.88) also found to be predicted well in ongoing study
  - P (0.72), K (0.48) not predicted so well
- Some important parameters not predicted so well (totals generally better than exchangeables)
- Preprocessing of the spectral data can improve the prediction accuracy if done appropriately
- Inclusion of site characteristics improves prediction accuracy
- Predictions can be made using a trained network in <5 seconds