AN EVALUATION OF METHODS FOR IMPUTATION OF MISSING TRACE ELEMENT DATA IN GROUNDWATERS

BRUCE DICKSON[1]

&

ANGELA GIBLIN[2]

[1]Dickson Research Pty Ltd

[2]CSIRO Honorary Research Fellow

Background to this study

- Uranium in saline waters can concentrate in salt-lakes through reduction and precipitation by *desulfovibrio spp* bacteria. This process has been observed to operate in salt lakes both in USA and Australia.

- In USA (San Joaquin valley, California) issues arising from concentration of U, Se and Mo in sediments and their effects on wildlife led to closure of salt disposal ponds

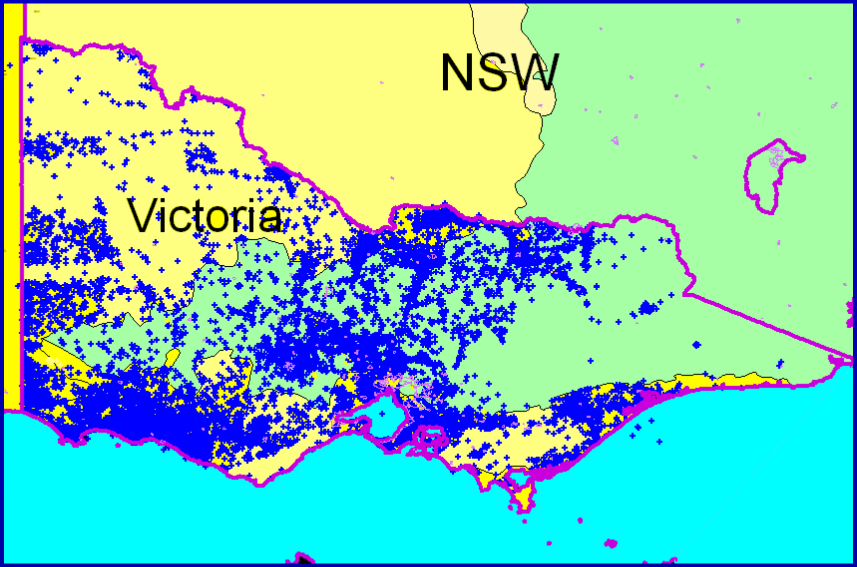
Factors in the formation of U concentrations in salt-lakes

- Sources of uranium (granites, marine shales)
  - That uranium must be “available” for leaching

- Elevated U in groundwaters
  - Cases where U accumulations have been found have coincided with levels of dissolved U being much higher than average

Groundwater data

- Vast amounts of groundwater data exists within state databases
  - e.g. for Vic there are over 75,000 analyses
- This could be a valuable resource for exploration but unfortunately most of this data covers major components only



Objective of Study

Can we use this groundwater resource in any way to extend the smaller database of samples with multi-element analyses?

Imputation

In statistics, imputation is the prediction of a missing value using a mathematical model in combination with available information.

Methods for Imputation

- Traditional
  - mean-substitution - MS
- Model-based
  - regression methods
  - expectation minimization - EM
  - multiple imputation
- Model-free
  - self-organizing maps - SOM

Imputation by Expectation Minimization

“With the EM algorithm, the parameters of a probability distribution are estimated from incomplete data by maximizing iteratively the likelihood of the available data”

- For Gaussian data, the probability distribution can be parameterized by matrices of means M and of covariance C
- Estimates of the M and C of an incomplete dataset depend on unknown missing values
- Conversely, estimates of missing values depend on the unknown statistics of data set (M & C)
- Assume process causing missing data is random

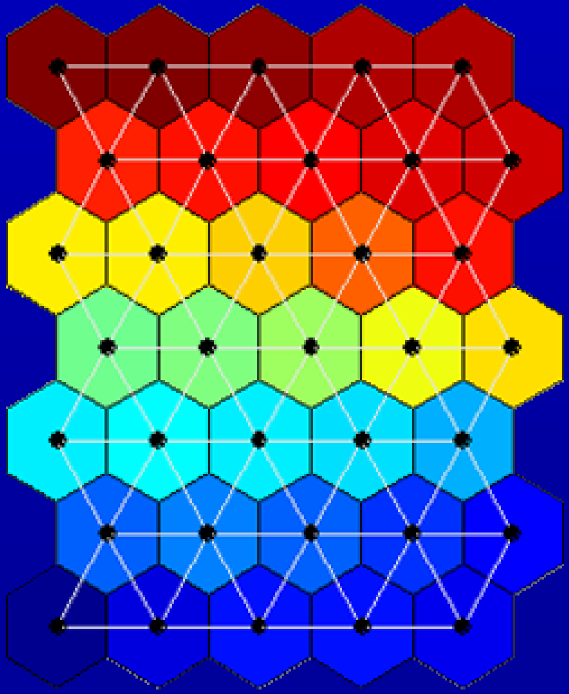
Expectation Minimization Algorithm

- EM algorithm starts with initial guess for M & C
- Cycles through alternating steps of
  - ➢ imputing missing values
  - ➢ re-estimating M & C
- Process stops when changes in M and C become smaller than a preset limit

Self-Organizing Maps

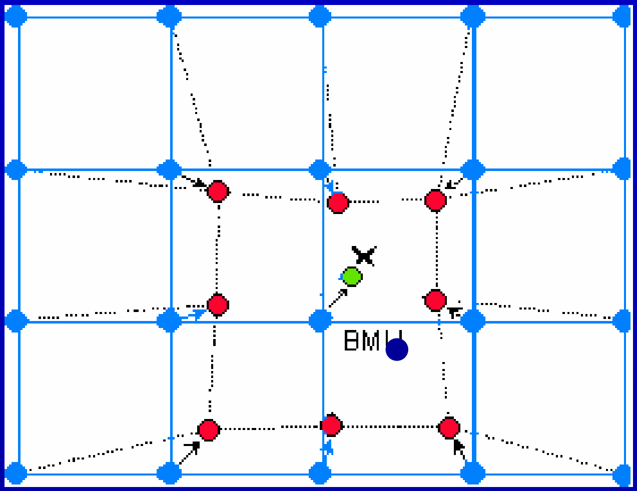
- A self-organizing map (SOM) is an unsupervised technique for visualisation of multivariate data.
- SOM summarizes the “essence” of a data set

- A SOM consists of a set of nodes set out in a pattern, each node comprising a vector of weights of the same dimension as the input data vectors.
- We refer to these as *code vectors*



Training a SOM

- The SOM is trained by presenting the data repeatedly and adjusting the weights to "learn" the structure of the data
- The weight adjustment is constrained by two processes-
  - Competitive learning
  - Cooperative learning



- Closeness is determined on basis of Euclidean distances between sample and code vectors

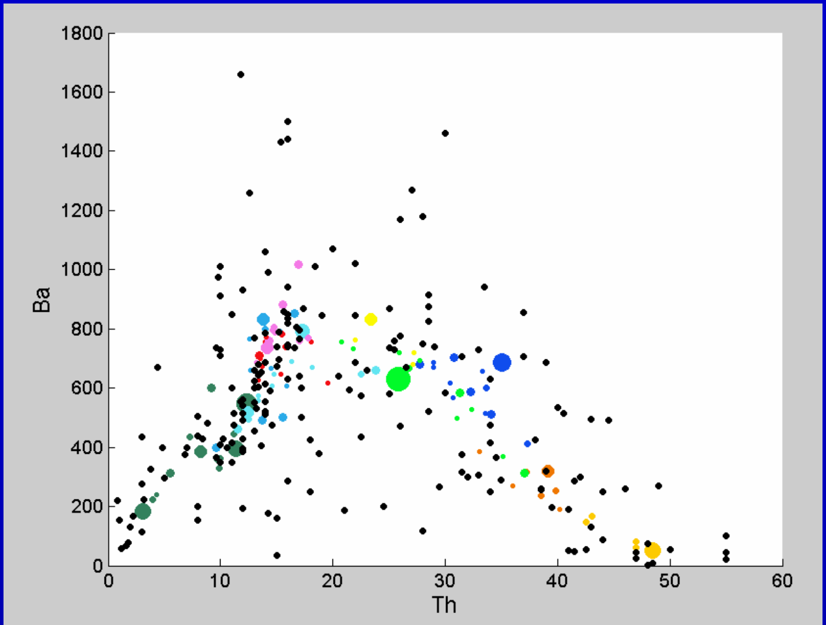
Imputation with SOM

- Due to its vector basis, a SOM can handle missing data by determining distances on the basis of what data is available
- Each sample vector has a “best-matching” code vector
- The value for a missing item in the sample vector is taken as the value for that item in the best-matching code vector

Performance of SOM with missing data

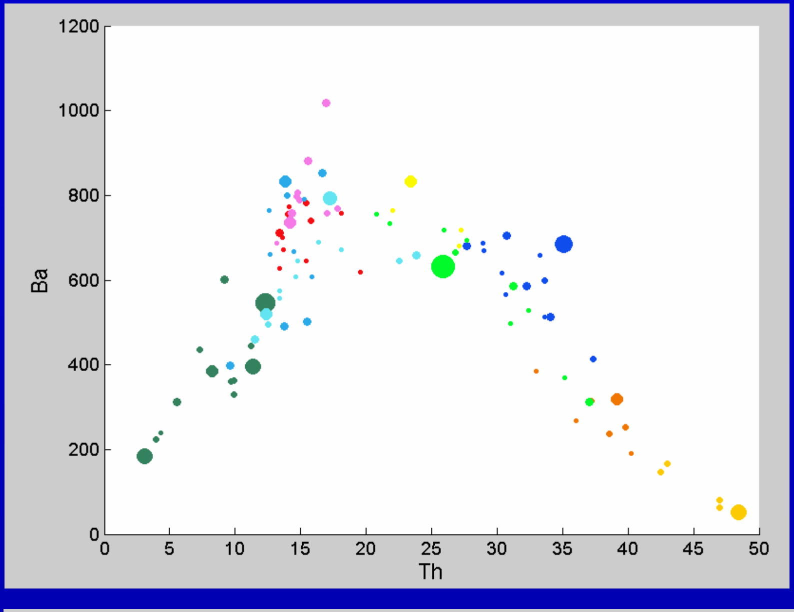
Data comprises analyses of 32 elements in 220 igneous rocks from NE Qld

SOM identifies Ba trend in fractioning igneous rocks, using Th as index

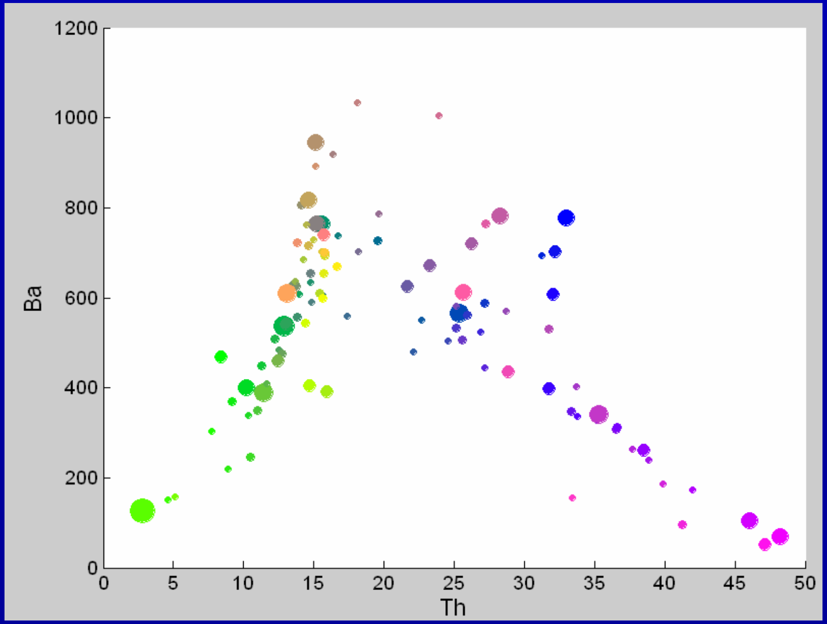


Performance of SOM with missing data

Without data we can see how SOM identifies Ba trend in fractioning igneous rocks, using Th as index

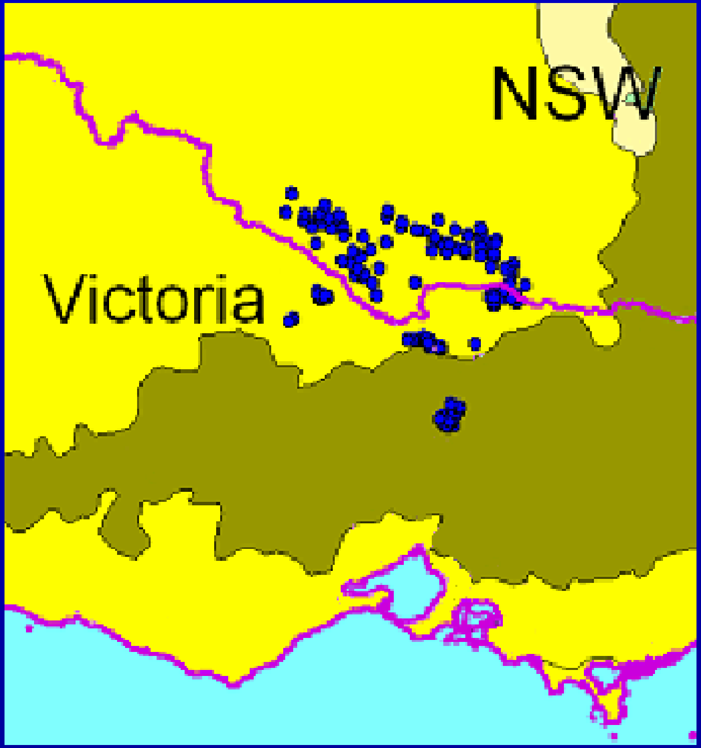


Same trend is still clearly seen although now 50% of the data has been set, at random, to missing

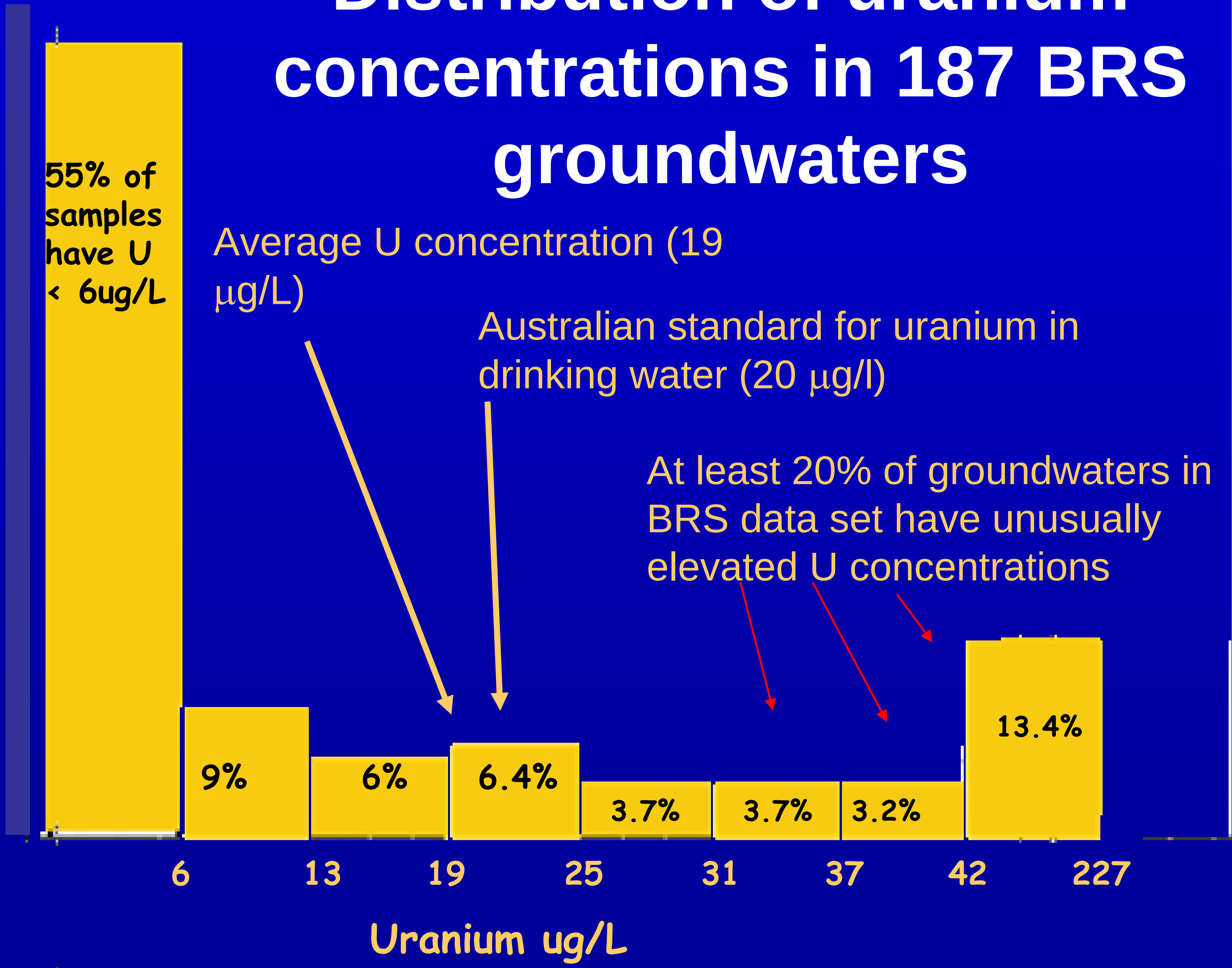


Data set for this study

- To test the methods we required a data set with comprehensive analyses
- Data sets published by Bureau of Rural Sciences, Canberra, for five areas in the eastern Murray Basin were selected and combined
- From the data we have selected 187 analyses of pH, major cations and anions and U and F
- Note that this data comes from areas of similar geology (the Parilla Sand unit)



Distribution of uranium concentrations in 187 BRS groundwaters



Average U concentration (19 μg/L)

Australian standard for uranium in drinking water (20 μg/l)

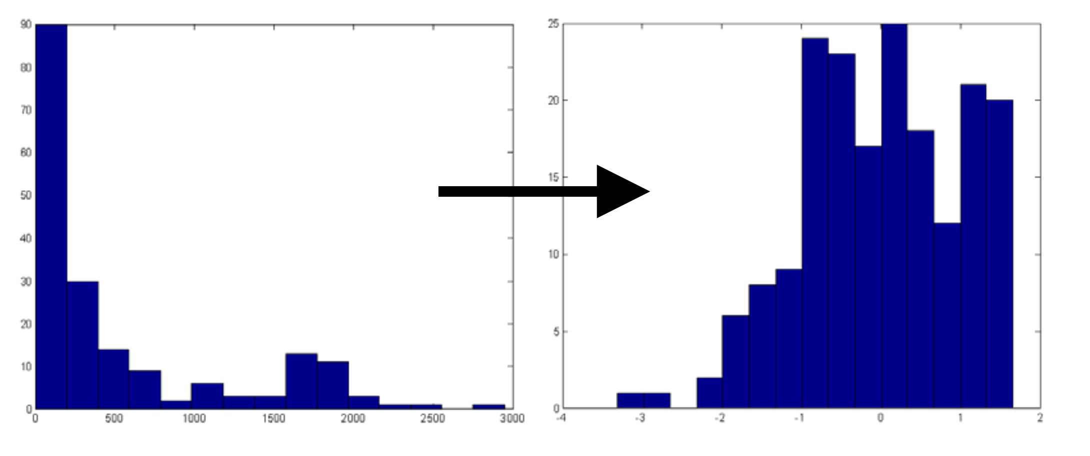
At least 20% of groundwaters in BRS data set have unusually elevated U concentrations

Why include F?

- Elevated U found in these waters is probably leached from granites
- Granites in this area of the Murray Basin are fractioned Siluro-Devonian S-type with elevated P, Rb and U
  - Some similar granites further west in Vic have secondary growths in vugs of uranium phosphates
- F in groundwaters is an indicator of granite

Pre-processing of the data set for SOM

- Data for all except pH were log-transformed then normalized to the range 0 – 1
- This ensures each variable is equally weighted in the analysis



Methodology

- Set a percentage of U values to missing
- Determine how well the imputation methods can determine replacement values for the missing values

Performance indicators

1. Plots of original vs imputed values
2. Estimation of mean and standard deviation for full data set after imputation
3. Root mean square error between original (O_i) and imputed (P_i) values (i = 1…N)

$$RMSE = ((\Sigma[P_i – O_i]^2)/N)^{½}$$

4. Mean average error

$$MAE = (\Sigma|P_i – O_i|)/N$$

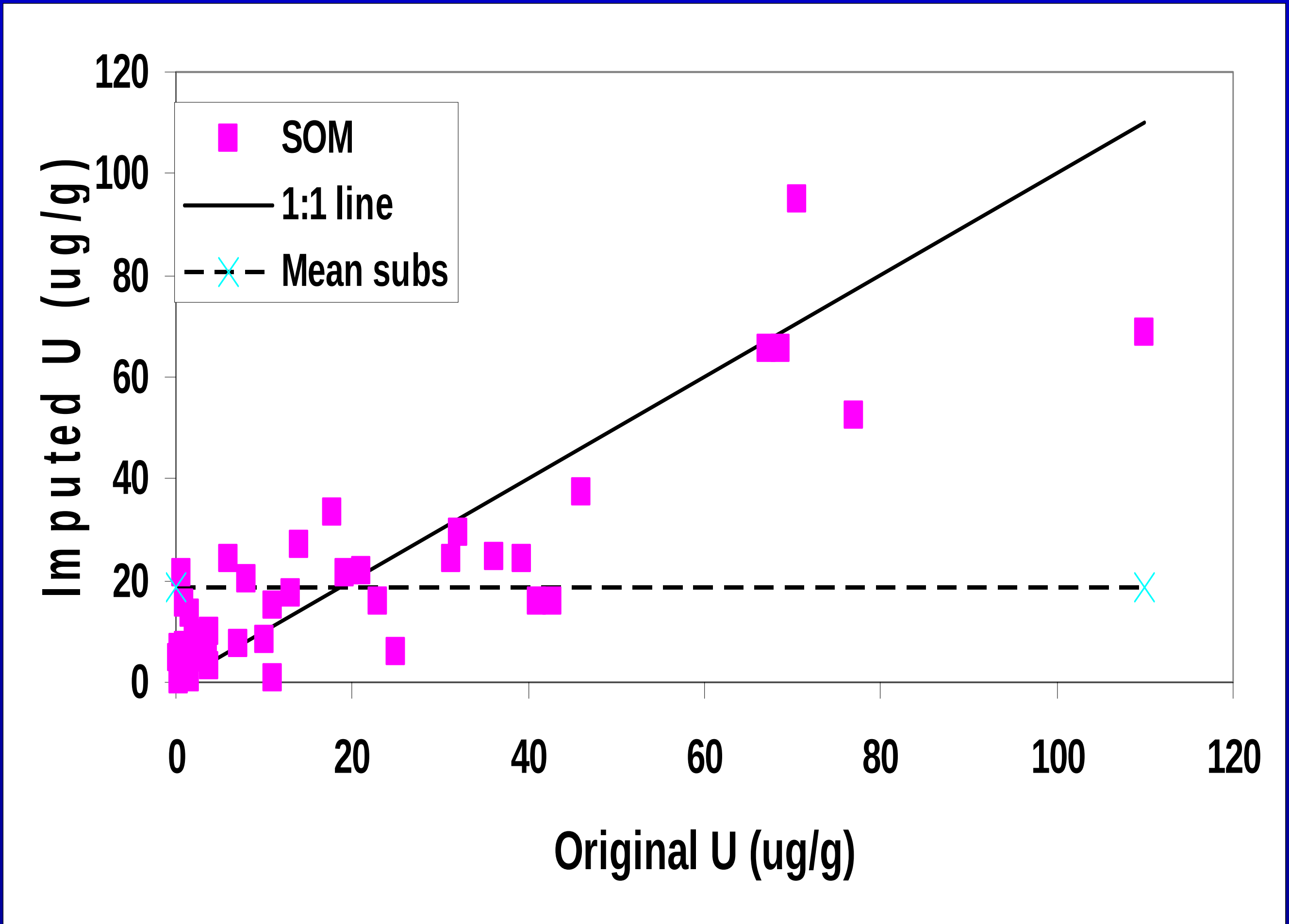
5. Index of agreement

$$Ia = 1 - (\Sigma[P_i – O_i]^2 / (\Sigma(|P_i – Õ| + |O_i – Õ|)^2)$$

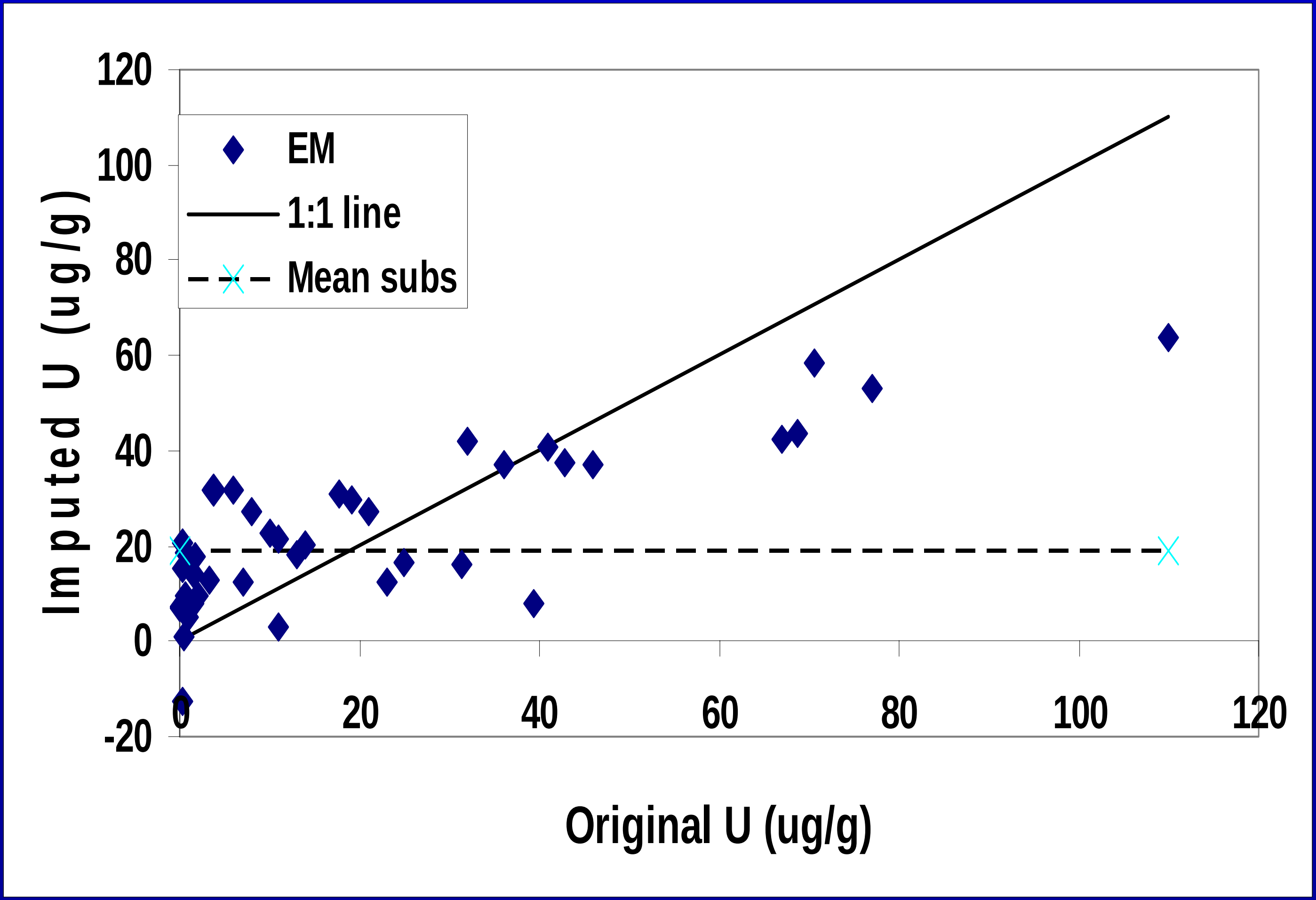
$$Õ = mean(O_i)$$

0 = no agreement; 1 = complete agreement

SOM results for 25% of U missing



EM results for 25% of U missing



Accuracy of imputation for 25% missing

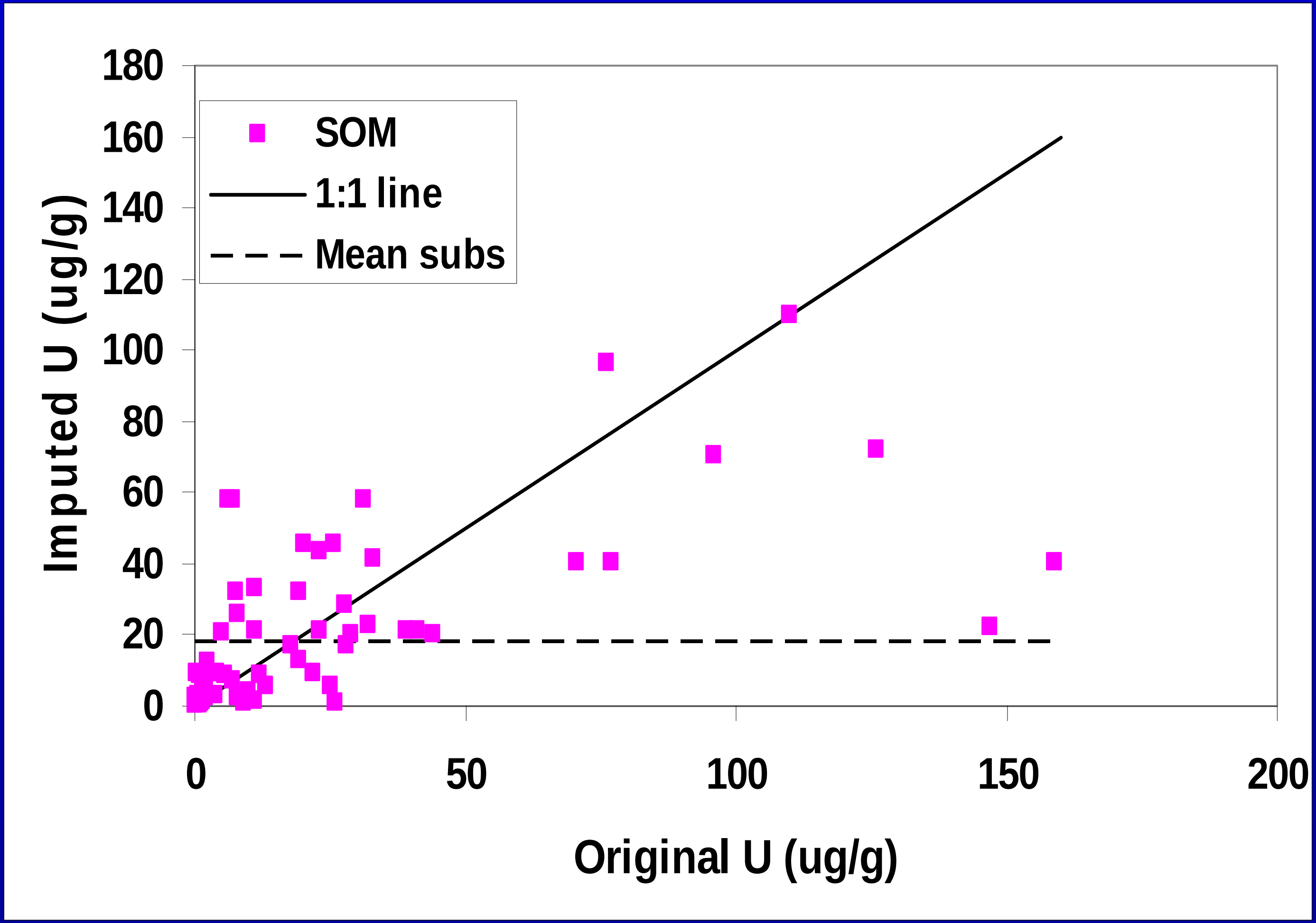
Case	mean	RMSE	MAE	Ia
All	19.2 ± 32.5			
25% miss	18.8 ± 34.2			
Mean sub	18.8 ± 30.1	25.8	19.4	0.11
EM	19.8 ± 31.1	16.7	13.7	0.92
SOM	19.1 ± 31.7	13.1	9.4	0.92

RMSE = root mean square error

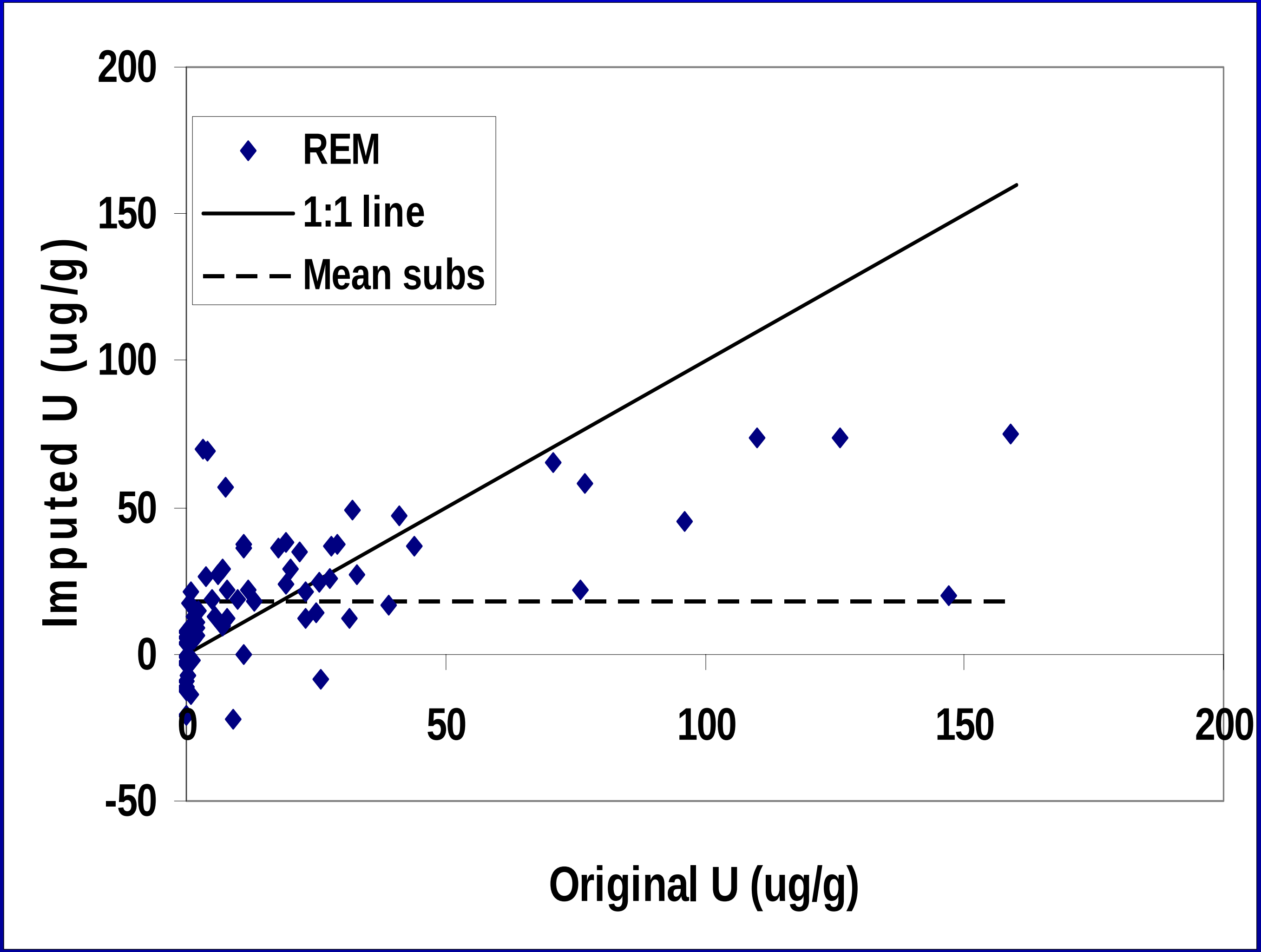
MAE = mean average error

Ia = Index of agreement

SOM result for 50% of U missing



EM results for 50% of U missing



Accuracy of imputation for 50% case

Case	mean	RMSE	MAE	Ia
All	19.2 ± 32.5			
50% miss	18.3 ± 31.6			
Mean sub	18.3 ± 24.4	33.9	21.5	0.08
EM	18.8 ± 28.4	21.2	17.2	0.89
SOM	18.6 ± 28.6	25.5	13.8	0.76

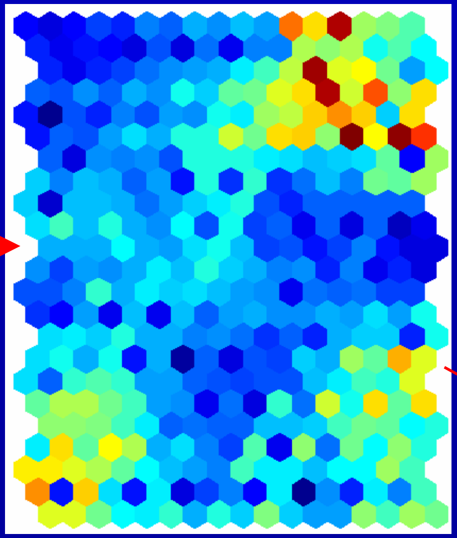
RMSE = root mean square error

MAE = mean average error

Ia = Index of agreement

Using SOM to predict U values

A precomputed SOM Map can be used to sort and impute values for new data



BRS data 187 samples with U

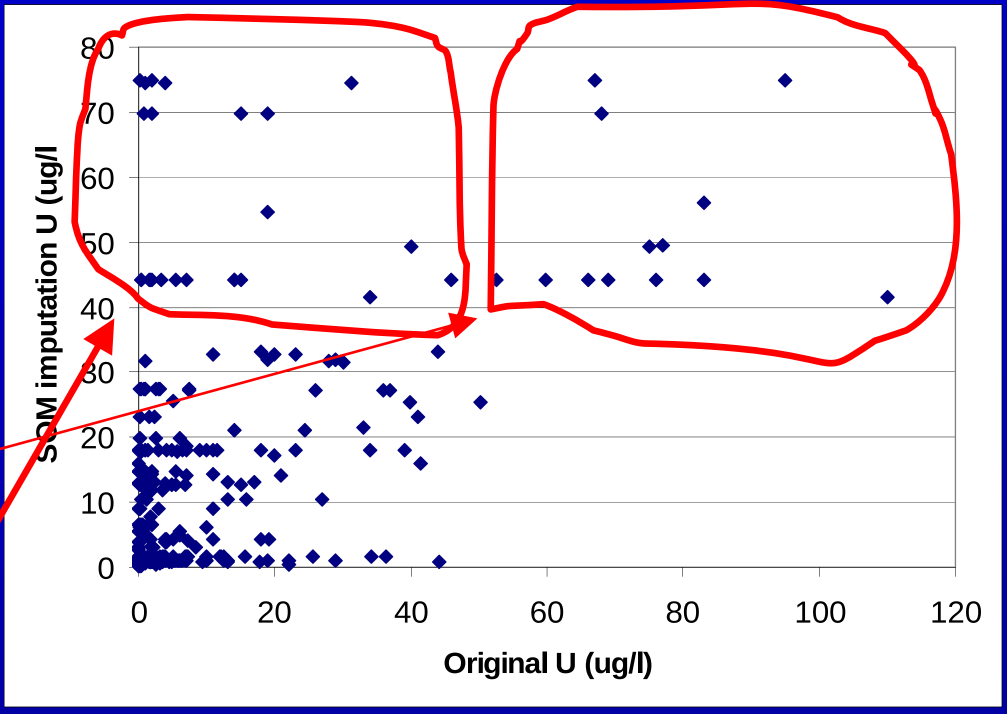
New data lacking U

Imputed values for U

SOM map

Predicting samples with elevated U values

- 361 extra samples from elsewhere in Murray Basin with pH, majors, F and U
- Assume U missing
- Figure shows 13 of 14 samples originally > 50 μg/l are > 40 μg/l (14 samples)
- Extra 23 samples also show > 40 μg/l

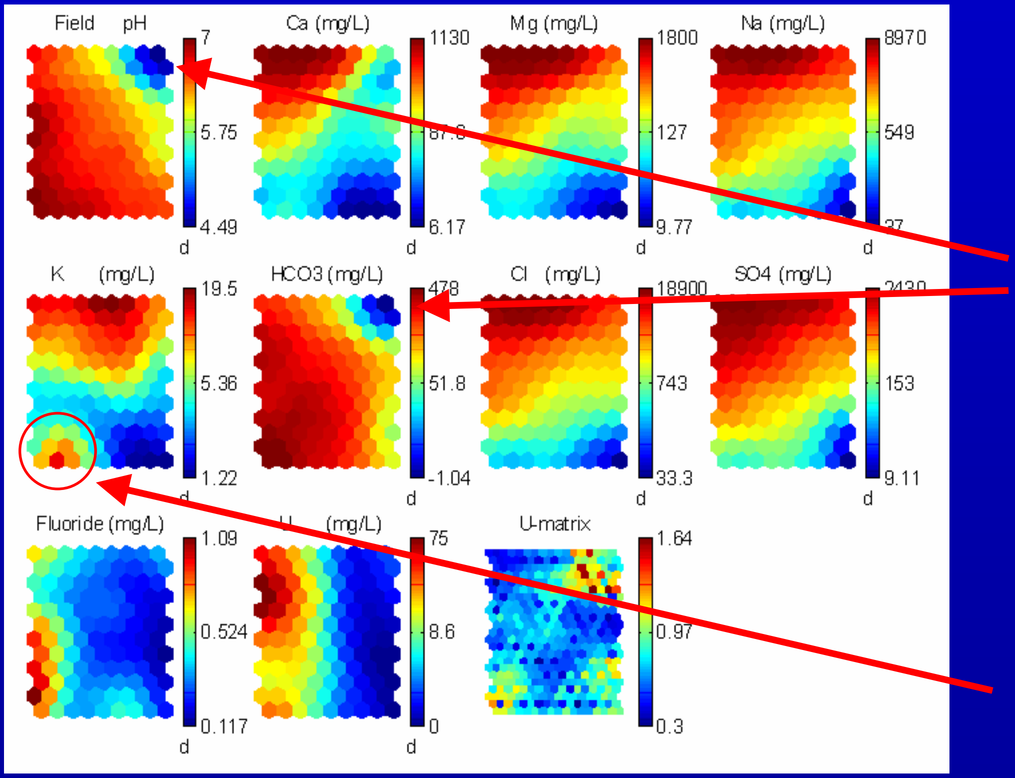


Method can be used to reduce number of samples for further investigation by 90%

Conclusions

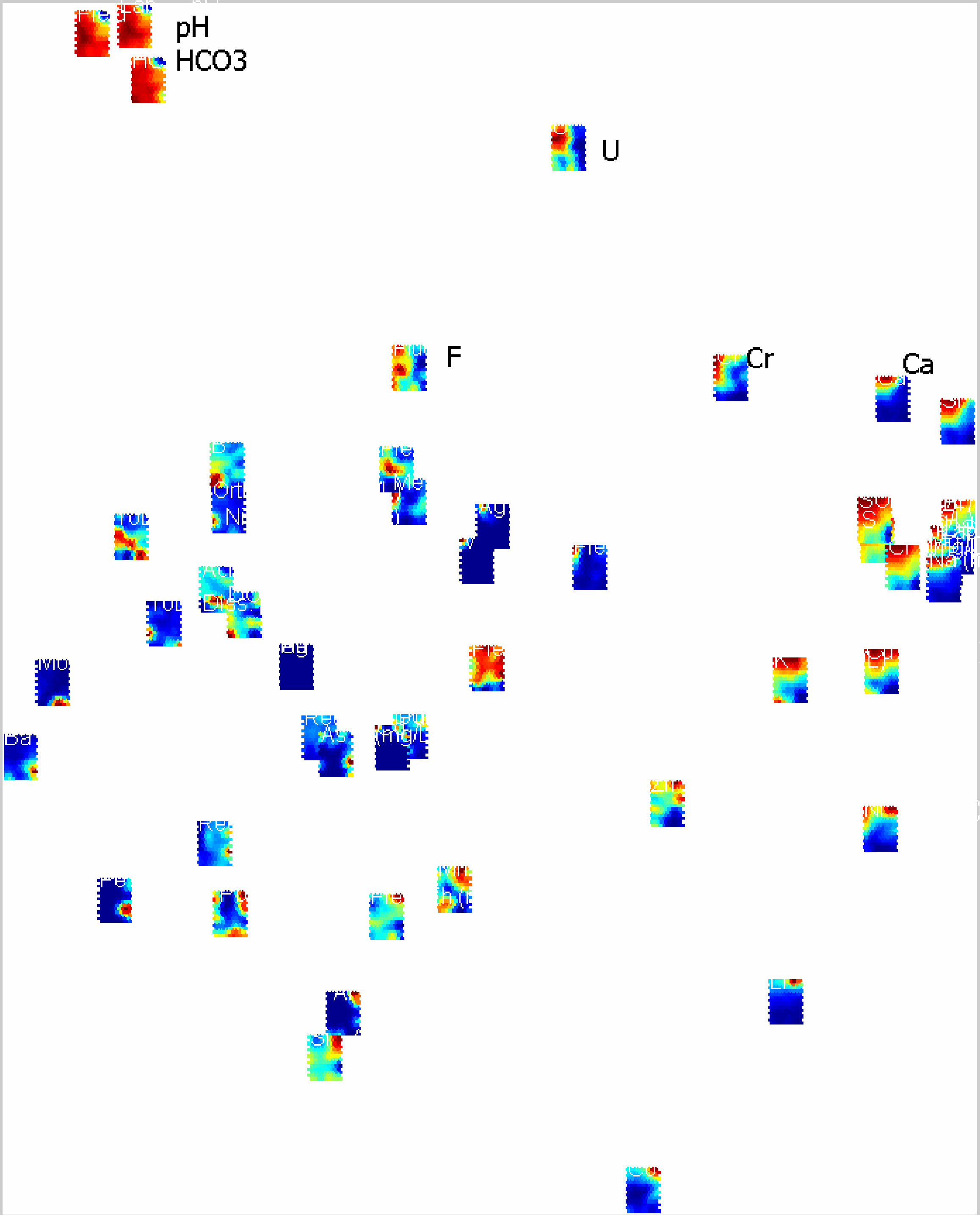
- U concentrations are a result of a complex interaction of groundwater, geology and mineralogy
- Major composition of water alone is poor indicator for U
- Thus, we conclude there is no substitute for measuring U directly
- But SOM can be used to select samples for further analysis

Could other elements stand in for U?
SOM component plots



These plots show how the variables distribute in SOM space. For example low pH waters (blue) are, of course associated with low HCO_3 waters.

Plot also shows high K occurs in two groups of water, one saline and the other low salinity and alkaline.



PCA analysis of component plots

Plot shows similarity of components (e.g. pH & HCO_3).

PCA analysis of SOM from complete data set (187 samples, 52 variables) show NO other trace element can act as a better proxy for U

Final Word

- The poor imputation results shown here for U should NOT be taken as a impugning imputation in general
- We undertook a difficult task
- SOM through its model-free approach is a useful addition to the array of imputation tools

Acknowledgments

- The SiroSOM program used in this study was written whilst Bruce was a member of CSIRO Exploration & Mining
- It uses the SOM Toolbox (*Vesanto et al, 2000, http://www.cis.hut.fi/projects/somtoolbox/*)
- The Expectation Minimization program is made available by T. Sceneider (*J. Climate, 14, 853-871, 2001*)
- All programs are written in MATLAB code