Jenny, PCA and Random Forests

By Budiman Minasny & Alex. McBratney
University of Sydney

The 2011 Hans Jenny Memorial Lecture in Soil Science was delivered by Prof. Garrison Sposito from UC Berkeley. He called his talk - The Genius of Soil. The video is available at http://youtu.be/y3q0mg54Li4. In the last part of the lecture, Gary drew attention to one of the little known paper by Hans Jenny in 1968 which was presented at the *Study Week on Organic Matter and Soil Fertility*, April 22-27, 1968, organised by the Pontificia Academia Scientiarum a scientific academy of the Vatican. This is probably the early Global Soil Carbon workshop.

The paper by Jenny et al. (1968) was the first chapter in this book, available at http://tinyurl.com/begpa2h (Jenny's appendix paper on "The image of soil in landscape art, Old and New" in the same book is better-known than this paper). In the study, Jenny collected 97 soil samples across a moisture transect in the Sierra Nevada, California where the variation in the factors of soil formation were to some degree controlled. The mean annual precipitation (MAP) is between 80 and 2000 mm, and mean annual temperature (MAT) between 10 and 16℃. The flora was restricted to pine tree and grass. The aspect is always southeast, with slopes varying from 0 to 30%. The parent materials are acidic and basic igneous rocks. Jenny used this data to quantitatively fit "an integrated clorpt model" where all factors were simultaneously modelled in the form of a multivariate linear regression:

$$s = a + k_1\,MAP + k_2\,MAT + k_3\,Parent\ Material + k_4\,slope + k_5\,Flora + k_6\,Latitude$$

In addition, Jenny also realised there would be correlation among the independent factors:

"When the independent variables X_1, X_2 , X_3, … are highly self-correlated (collinear) the slope coefficients *b* become unstable, even meaningless as to sign. Regressing for example N against Precipitation (*P*) (in.) and Temperature (*T*) (°F) gives

$$N = 0.350 + 0.0012\,P - 0.0055\,T$$

with R^2= 0.324. Introducing Leaching value (*Li*, in.), which is highly collinear with *P*, results in

$$N = 0.375 + 0.0037\,P - 0.0062\,T - 0.0029\,Li$$

The slope coefficient of *P* has tripled and that of *Li* is negative, which is absurd from the viewpoint of soil leaching. R^2 remains essentially unchanged as 0.327.

The handicap of self-correlation can be overcome by computing "*principal components*" (Some of the content of this paper is later used in the last chapter of Jenny's 1980 *The Soil Resource Book*, pp. 361-363).

Gary in his talk indicated that this was the first paper that used PCA in soil studies. Intrigued by his comment, we tried to find out whether there are predecessors. Jim Wallis one of the co-authors of Jenny's paper (who then worked at I.B.M. Watson Research Center, Yorktown Heights, N.Y.) wrote to Gary (email message from Jim Wallis to Gary Sposito, May 15, 2011):

"It is not possible for me to say definitely that my work/paper was the first use of principal-components regression in soil science, but the probability that it was is extremely high. What is certain is that very few people at the time seemed able to understand the methodology or provide references to similar work. It seemed that it would help me with my dissertation on accelerated soil erosion, and I used it in my dissertation - it was highly controversial at the time.

A sidelight on how I arrived at the methodology follows. There was a Professor Meredith in the Psychology Department at the time, he taught a graduate course in Factor Analysis which I unofficially audited, and it seemed to me that if one used principal-component regression to determine the number of factors at work in soil formation (eigenvalues >1) and rotated the matrix into the variable space by Varimax that you would have a quantitative measure of Jenny's CLORPT equation. I wrote a 120-variable computer program to do just that. Jenny was not on campus that year, but he came back in the spring of 1966, got excited by its possibilities for pedology, although I had little to do with the writing of our joint paper, beyond a few conversations and notes that did not get preserved. He demanded that I give a seminar to the Soils Department on the subject, so I did."

Jim Wallis introduced PCA to hydrology in a 1965 paper[1] (Wallis 1965), and also wrote a FORTRAN program called WALLY1. Jim Wallis is a well-known hydrologist who wrote the first paper on fractal in hydrology with Benoit Mandelbrot (1968), and he was the president of the Hydrology section at the AGU.

[1]The hydologists always seem to be a couple of years ahead of the soil scientists

Jenny, PCA and Random Forests

The earliest references to techniques in Principal Component Analysis (PCA) were Karl Pearson in 1901 and Hotelling (1933). However it was not until the 1960s with the availability of computers that the analysis became practical. Earlier papers on soil can be found on mostly on factor analysis (rotated principal components or principal factors that are not necessarily orthogonal). The thrust for factor analysis was largely from the social sciences (Psychometrics) rather than the physical ones. Rayner's 1966 paper was on principal coordinate analysis which involves finding the principal components corresponding to similarity matrices. This analysis was invented by John Gower at Rothamsted principally to help James Rayner with the soil similarity problem, however as it turns out it is virtually the same as multidimensional scaling which was invented in the 1950's by the psychologists. The first use of multivariate statistical methods for soil (data) that we know of is Cox and Martin (1937). See the Ordination article by Dick Webster in Pedometron 29.

Searching through the Web of Knowledge, we found several earlier papers that used PCA in soil studies. A paper by Gyllenberg (1964) from Finland and another one by Skyring and Quapling (1968) from Canada used PCA as a way to describe soil microbial diversity. Yamamoto and Anderson (1967) used PCA (instead of multiple linear regression) to find the degree of association between soil physical properties (soil aggregate stability, erodibility) and the soil-forming actors for wildland soils of Oahu, Hawaii. This bears the closest resemblance to Jenny's 1968 paper. Their study was also inspired by Jim Wallis' paper in hydrology (Wallis, 1965). There was also a PhD dissertation by John Berglund in 1969 from State University College of Forestry at Syracuse University, where PCA was used to develop and interpret prediction equations to estimate forest productivity from its soil properties. Dick Webster and his student Ignatius Wong (1969) used PCA to analyse soil data collected along a transect. The main use here was for ordination - many soil properties were combined into a first principal component so that soil property variation could be plotted as a graph along a transect.

While Jenny may not be the first to use PCA in soil studies, the 1968 paper lays the fundamentals of what is now called digital soil mapping. It should be a good reminder for us on how to mindfully choose the best covariates and model. We need to remember that Jenny's linear model is used to explain the factors that control the distribution of soil properties, not specifically as a spatial prediction function.

Jenny (1980) wrote:

"The computer's verdict of tangible linkages of soil properties to the state factors pertains to today's environment. Either the pedologically effective climate has been stable for a long time, or past climates are highly correlated with modern ones, or the chosen soil properties have readjusted themselves to today's precipitation."

Nowadays (notwithstanding its simplicity) PCA is still extensively used in soil science and pedometrics, for drastically reducing the number of variables in soil spectral data, finding patterns (clusters) in the data, reducing dimensions of microbial diversity data, or satellite images, etc. According to Scopus, since 2010, there has been an average of 450 papers per year on the application of PCA to soil data.

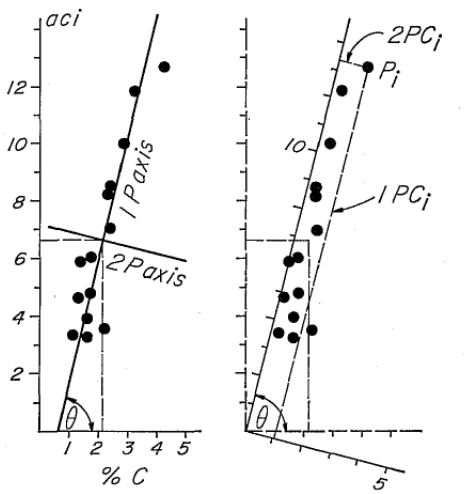

Figure 1. Illustration of converting original variables (aci, C) to first and second principle components (from Jenny et al., 1968).

Pygmy Forest to Random Forests

Research in digital soil mapping now has moved from carefully controlled environmental factors to "real-world" soil data, either collected from stratified random sampling or using legacy soil data. Models of the Jenny et al. (1968) type are still being developed (Gray et al., 2012), while others prefer to use data-mining techniques. Data-mining models are usually treated as a black-box as they are complex and cannot be easily or explicitly written out. However, the results can be expressed as significant predictors or variables of importance and usually interpreted as 'knowledge discovery' from databases which are then sometimes justified a *posteriori* by principles of soil genesis. As opposed to a process-based model, where

the process needs to be specified first, the data-mining approach is said to "learn" the process through the data. As an example, the Random Forests technique has been used a lot in digital soil mapping as it is freely available and it has been claimed that "Random forests does not overfit. You can run as many trees as you want" (From the Random Forests Manual by Breiman and Cutler). In addition, the author also made claims that it is: "The most accurate current prediction", "a complex predictor can yield a wealth of 'interpretable' scientific information about the prediction mechanism and the data."

An example of the use of Random Forests is given in Figure 2, which shows the prediction for some surface soil carbon data in the Hunter Valley, NSW, Australia, where the fit is excellent on the training data (using 100 trees), R^2 = 0.94. The variable of importance indicated that in addition to indices calculated from Landsat bands, terrain attributes of MrVBF (Multi-resolution Valley Bottom Flatness) and TWI (topographic wetness index) play important roles. The map confirmed this and it is in accordance with our pedological knowledge, where carbon concentration is expected to be larger in areas with higher moisture and areas of deposition (knowledge discovery).

But wait a minute, Figure 3, shows the fit on an internal validation (out of bag estimates) and an independent validation data, where there is no fit at all. The soil carbon data has very little correlation with any of the terrain attributes and is very weakly correlated with some Landsat imagery. It is obvious that Random Forests can easily overfit the data. Overfitting implies the model describes the noise in the data (perfect fit on the training data), while has poor predictive capability in the validation data. (The data and R code are available to download from here, and you can experiment yourself with the notion that RF can fit anything). It is quite interesting that scientists take the statement "Random Forests does not overfit" as the truth, and repeatedly quote this in many papers without any question.

A recent news article mentioned the latest breakthrough in technology: "With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart." Perhaps we should tell you that we need to explore Deep Learning tools for pedometrics. But we think now we need to remind ourselves that explicit linear models should be at least considered as a starting point for exploratory data analysis before trying the fancy tools. There is no magic algorithm that can fit everything — yet not overfit.

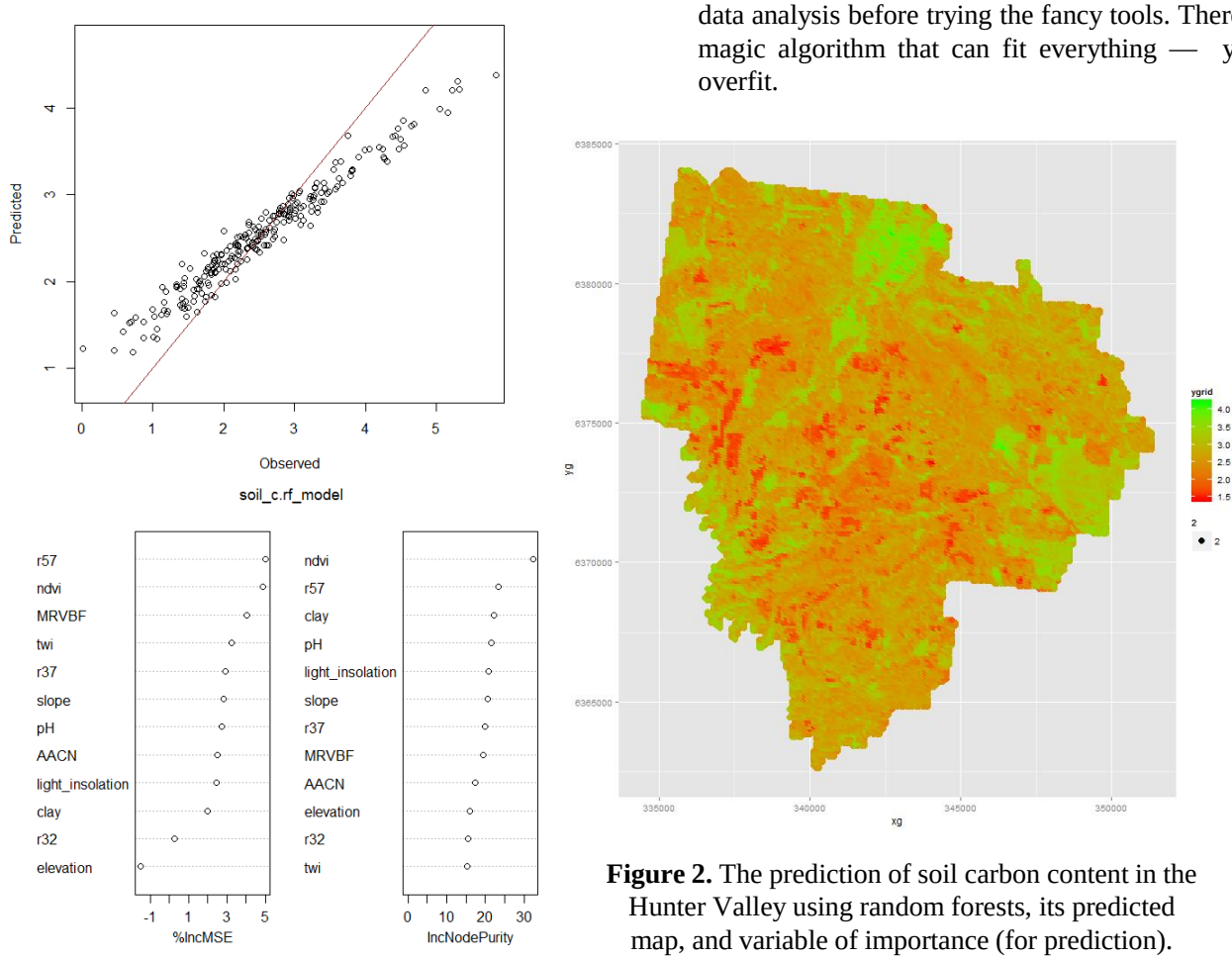

Figure 2. The prediction of soil carbon content in the Hunter Valley using random forests, its predicted map, and variable of importance (for prediction).

Summary

We've come a long way from Jenny's pedological study in the Pygmy Forest to using Random Forests for making soil predictions. Technology has advanced, powerful computers that can handle complex algorithms and there is now widespread availability of high-resolution covariates. We still stick to the same principle that while we need to make use of all of the new technologies, common sense and parsimony must prevail over fancy tools.

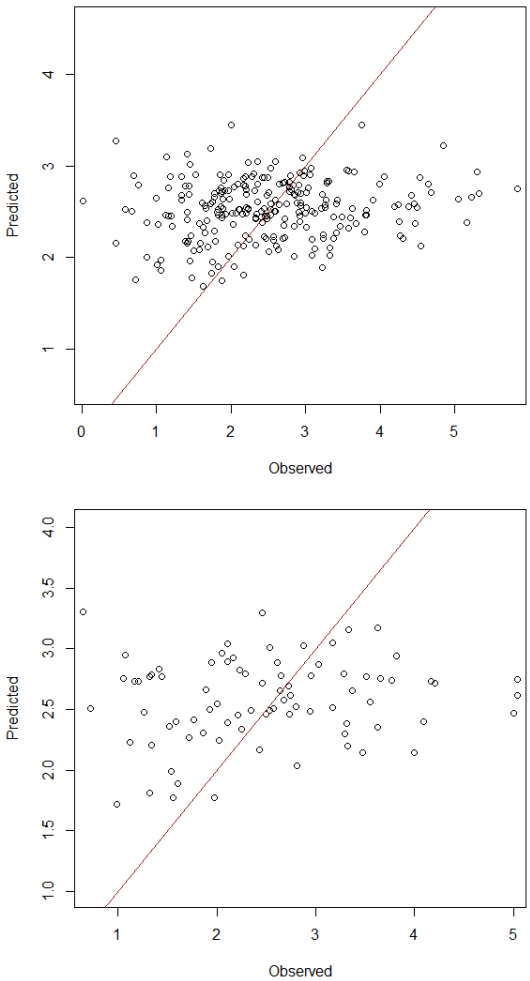

Figure 3. The out of bag fit vs. observed carbon content fitted using random forests (up) and the fit for an independent validation dataset (down).

References

Berglund, J.V., 1969. The Use of Modal Soil Taxonomic Units for the Prediction of Sugar Maple Site Productivity In *Southern New York. Dissertation Abstracts* 29B, 2696-2696.

Cox, G.M., Martin, W.M., 1937. Use of a discriminant function for differentiating soils with different azotobacter populations. *Journal paper No. J451 of the Iowa Experimental Station*, pp. 323–332.

Gray, J., Bishop, T., Smith, P., Robinson, N., Brough, D., 2012. A pragmatic quantitative model for soil organic carbon distribution in eastern Australia. In *Digital Soil Assessments*, CRC Press, pp. 115-120.

Gyllenberg, H.G., 1964. An approach to numerical description of microbial isolates of soil bacterial populations. *Ann Acad Sci Fennice*, Ser. A Biol 81, 1-23.

Jenny, H., 1980. *The Soil Resource*. Springer-Verlag.

Jenny, H., Salem, A.E., Wallis, J.R., 1968. Interplay of soil organic matter and soil fertility with state factors and soil properties. *Study Week on Organic Matter and Soil Fertility*, Pontif. Acad. Sci. Scripta varia, 32, 5-36.

Mandelbrot, B.B., Wallis, J.R., 1968. Noah, Joseph, and operational hydrology. *Water Resources Research* 4 , 909-918.

Rayner, J.H. 1966. Classification of soils by numerical methods. *Journal of Soil Science*, 17, 79-92.

Skyring, G.W., Quapling, C., 1968. Soil bacteria: principal component analysis of descriptions of named cultures. *Canadian Journal of Microbiology* 15, 141-158.

Wallis, J.R., 1965. Multivariate statistical methods in hydrology - A comparison using data of known functional relationship. *Water Resources Research* 1, 447-461.

Webster, R., 2010. An early history of ordination in soil science Ordination. *Pedometron* 29, 20-24.

Postscript by Alex.

The availability of principal components and more general multivariate methods for looking at soil took off fairly quickly after the sixties. When I did my first serious pedometrics work, which was in the long hot summer of 1976, with Dick Webster at Yarnton, software for doing PCA, discriminant analysis, principal coordinates etc. was readily available in programs such as Genstat, BMDP, SPSS and SAS. They were the powerful forerunners of S and then R. In my alma mater at Aberdeen another mentor the soil physical chemist Michael Court very much favoured principal factor analysis over principal components analysis. Largely with Dick Webster's help I learned the mechanics of the multivariate methods – and they continue to serve well. They should be in any pedometrician's toolbox.