

Provision of soil information for biophysical modelling

José Sergei Padarian Campusano

A thesis submitted in fulfilment of the requirements for the degree of

Master of Philosophy

2014

Faculty of Agriculture and Environment

The University of Sydney

New South Wales

Australia



Certificate of Originality

This thesis is submitted to the University of Sydney in fulfilment of the requirements
for the degree of Master of Philosophy.

The work presented in this thesis is, to the best of my knowledge and belief, original
except as acknowledge in the text. I hereby declare that I have not submitted this
material, either in full or in part, for a degree at this or any other institution.

Signature: José Padarian

Date: October 1, 2014

Thesis summary

This thesis is concerned with the generation of a framework for addressing soil data needs, specifically for biophysical modelling. The soil system is an important ecosystem actor, supporting most of the worlds' food production and being the major terrestrial carbon stocks, thereby information about it is crucial for management and policy making. To provide this information, it is important to deliver information of the highest possible quality; thus the need to define guidelines to standardise, not only the methodologies, but the minimum requirements that information must meet.

In this project, providing soil data is addressed in two ways. The first scenario investigates the use of soil information to predict other soil properties, using pedotransfer function (PTFs).

Chapter 1 addresses a usual problem of soil data non-uniformity. Two major soil textural classifications are used in the world, the International and the USDA/FAO systems. The difference between these two systems is the limit between the silt and sand particle sizes: $20\mu m$ for the International and $50\mu m$ for the USDA/FAO. A conversion between both systems is proposed through the use of PTFs generated using symbolic regressions (genetic programming technique). There are previous works on this topic, but the decision of generating new PTFs lays on the availability of new soil textural data, with measurements for both classification systems. The generated PTFs outperform the previous ones, reducing the prediction error by 15%-24%.

Chapter 2 extends the method used in Chapter 1, using the fuzzy k-means with extragrades (FKM_{ex}) algorithm to assess the uncertainty of the predictions. It is stressed that quantifying uncertainty levels for any model (including PTFs) is essential to evaluate risk involved in using the predictions for a decision-making process. The chapter begins with a summary of the main soil properties used by biophysical models in Australia. After identifying eight common soil properties, several PTFs related to soil water content were generated, using symbolic regressions. The incompatibility between field and laboratory measurements is also addressed, proposing PTFs to correct the water content measured in laboratory conditions. Besides the fact of obtaining the error magnitude of predictions, an important concept is integrated with the uncertainty estimation method: end-users are capable of identifying when their samples are too dissimilar compared with the datasets used to generate the PTFs.

The FKM_{ex} algorithm penalises those samples scaling the error magnitude up to two times, depending how far from the original dataset they are.

In the second scenario, it is assumed that the end-user does not have extra information about the soil properties at a specific location. In this case, the use of existing soil maps is a traditional solution, thus in Chapter 3 a framework for generating maps at national/continental scale, using digital soil mapping (DSM) techniques, is proposed.

Chapter 3 presents the spatial distribution of available water content (AWC) using environmental covariates to make predictions over Australia's wheatbelt. The aim of this chapter is to reconcile model parsimony (number of covariates), accuracy (numerical performance) and realism of the visual representations (maps). To achieve this, several combination of covariates were used, varying the complexity of the model inputs. Spatial predictions were made using three modelling techniques: symbolic regression, Cubist, and support vector machines. The concept of model averaging was also explored, trying to obtain an ensemble model that combines the best of all the individual models. After a numerical and visual evaluation of maps generated with all the combinations of covariates, modelling techniques and ensemble methods, the ensemble model using all the available covariates showed the highest accuracy levels, but it was incapable of realistically representing the spatial structure of AWC. From this, it is stressed the need to consider the knowledge about the modelled process and not only focus on the numerical performance in order to obtain a flexible and stable model, but to also produce a realistic visual representation of it. The uncertainty concept is reinforced in this chapter, delivering a map of uncertainty levels along with the final map of AWC predictions.

Finally, Chapter 4 presents a synthesis of the previous chapters and main findings of the project. There are always new opportunities for further work in how to provide information due to the evolving nature of end-users, data availability and analytic methodologies.

Chapters of this thesis that have been submitted and/or published in scientific journals

Chapter 1

Padarian, J., Minasny, B., and McBratney, A. 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. *Soil Research* 50 (6): 443–446

Chapter 3

Padarian, J., Minasny, B., McBratney, A. B., Dalgliesh, N. 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. *Geoderma Regional* (in press).

Contents

General introduction	1
1 Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system	5
1.1 Introduction	9
1.2 Data sets	10
1.3 Genetic programming	10
1.4 Particle-size conversion	13
1.5 Conclusions	14
1.6 References	16
2 Provision of soil water retention information for biophysical modelling: an example for Australia	19
2.1 Introduction	20
2.2 Soil data requirements of biophysical models	20
2.2.1 Australian biophysical models	22
2.3 Prediction of soil water retention properties	22
2.3.1 Data sets	25
2.3.2 PTF development	26
2.3.3 Uncertainty estimation	27
2.4 Results	30
2.4.1 Drained upper limit	30
2.4.2 Water content at -10 kPa	31
2.4.3 Relation between DUL and -10	31
2.4.4 Crop lower limit	33

2.4.5	Water content at -1500 kPa	33
2.4.6	Relationship between CLL and -1500	34
2.4.7	Uncertainty estimation	35
2.4.8	External validation	35
2.5	Making predictions with new data	36
2.6	Conclusions	38
2.7	References	39
	Appendix A - Models description	46
	Appendix B - PTF use diagram	49
	Appendix - PTFs cluster information	50
3	Predicting and mapping the soil available water capacity of Australian wheatbelt	61
3.1	Introduction	65
3.2	Materials and methods	66
3.2.1	Data sets and study area	66
3.2.2	Digital soil mapping model	67
3.2.3	Prediction and mapping	70
3.3	Results	72
3.3.1	Covariates selection	72
3.3.2	Ensemble Model	73
3.3.3	Visual evaluation	74
3.3.4	Validation	77
3.3.5	AWC map	78
3.4	Conclusions	79
3.5	References	82
4	General discussion, conclusions and future research	87
4.1	General discussion	87
4.2	Overall research conclusions	89
4.3	Future work	90
4.4	References	92

List of Figures

1.1	Example of an initial population of four randomly created individuals representing GP models: (a) $x + 1$, (b) $x^2 + 1$, (c) 2 and (d) x . This representations should be read from left to right and bottom to top.	12
1.2	Surface plot of 2-50 μm fraction prediction at different clay ($< 2\mu m$) and sand (20-2000 μm) contents. Note that 20-200 μm represents the fine-sand fraction.	14
1.3	Surface plot of residuals of Eq. 1.4, as a function of clay and sand content, using IGBP-DIS data set.	15
2.1	Example of an initial population of two randomly created individuals representing GP models: (a) $x + 2$ and (b) 2. This representations should be read from left to right and bottom to top.	28
2.2	Prediction comparison between: a-c $\hat{\theta}_{DUL}$ and $\hat{\theta}_{-10}$ using field measurements dataset (Table 2.3). Predictions were made using Eq. 2.9 and Eq. 2.11; d $\hat{\theta}_{DUL}$ and predictions made with PTF proposed by Twarakavi <i>et al.</i> , (2009).	32
2.3	Prediction comparison between $\hat{\theta}_{CLL}$ and $\hat{\theta}_{-1500}$ using field measurements dataset (Table 2.3). Predictions were made using Eq. 2.13 and Eq. 2.15.	34
2.4	PICP and MPI behaviour with different number of cluster for Eq. 2.9. Dotted circles highlight the optimum number of clusters.	35
2.5	Relative positioning of observations in relation of class centroids. Convex hull represents the limit to consider an observation as an extragrade.	37
B.1	PTF use diagram	49
B.1	Location of soil profiles from APSRU database. Greyed area represents the bioregion subset where predictions were made.	66

B.2	Boxplot of model cross-validated RMSE (20 iterations) trained with different combinations of covariates. The prediction corresponds to DUL at 0-5 cm depth. bio: bioregion; topo: slope, TWI and MRVBF; weathering: weathering index; gamma: gamma-ray spectrometry; landsat: Landsat 7 bands; climate: air temperature, rainfall, evapotranspiration, Prescott Index. Letter on the right margin represent mean groups after an ANOVA/Tukey analysis.	73
B.3	Subarea showing artefact caused by ^{232}Th data on DUL map. (a) Map with artefact, and (b) map without artefact.	75
B.4	Standard deviation of DUL predictions, between 0 and 5 cm depth, of SVM, Cubist, and GP models, with different combinations of covariates. (a) COV 1, (b) COV 2, and (c) COV 3.	77
B.5	Water content to 1 meter depth (mm m^{-1}) based on ensemble model, using COV 3. Red colour represents negative values. (a) Drainage upper limit, and (b) Crop lower limit.	78
B.6	Diagram of estimation of AWC based on DUL and CLL values with their respective uncertainty levels.	79
B.7	Prediction interval width (mm m^{-1}) based on SVM predictions, using COV 3, to 1 meter depth.	80
B.8	Available water content (mm m^{-1}) based on SVM predictions, using COV 3, to 1 meter depth.	81

List of Tables

1.1	Data sets used in this work	10
1.2	Statistics of data sets by particle fractions	11
1.3	External validation statistics of prediction quality	15
2.1	Some biophysical models commonly used in Australia and related soil properties	23
2.2	Soil properties commonly used in reviewed models and predictors mentioned in literature	24
2.3	Statistics of soil samples used for PTF generation of field measurements.	25
2.4	Statistics of soil samples used for PTF generation of laboratory measurements.	26
2.5	Statistics of soil samples used for PTFs' external validation.	27
2.6	External validation statistics of prediction quality	36
2.7	Membership (m) in clusters C1, C2 and extragrade (*), prediction intervals (PI), prediction limits (PL) and DUL prediction for example observations, using Eq. 2.9.	38
C.1	Calibration cluster information for use with Eq. 2.5.	50
C.2	Calibration cluster information for use with Eq. 2.6.	51
C.3	Calibration cluster information for use with Eq. 2.7.	52
C.4	Calibration cluster information for use with Eq. 2.8.	53
C.5	Calibration cluster information for use with Eq. 2.9.	54
C.6	Calibration cluster information for use with Eq. 2.10.	55
C.7	Calibration cluster information for use with Eq. 2.11.	56
C.8	Calibration cluster information for use with Eq. 2.13.	57
C.9	Calibration cluster information for use with Eq. 2.14.	58

C.10	Calibration cluster information for use with Eq. 2.15.	59
C.11	Calibration cluster information for use with Eq. 2.16.	60
3.1	Statistics of soil samples used for model generation.	67
3.2	RMSE values of DUL validation between 0 and 5 cm depth. Mean of 50 iterations by model and covariate combinations. RMSE range between brackets. Units in $m\ m^{-1}$).	74
3.3	R^2 values for validation of DUL and CLL in depth, using SVM and COV 3.	78

General introduction

Undoubtedly, we are immersed in an era where data generation is much faster than it used to be. Whole genome sequencing and astronomical data are some examples where data flow in the last decade experienced an important change. Is it possible to achieve a similar data volume in all the fields of science?

In soil science, most data collection must be conducted in the field, making difficult to reduce the costs to produce more data. Two important techniques emerge to try to overcome this issue: the use of pedotransfer functions (PTFs) and digital soil mapping (DSM) with the use of spatial environmental data.

PTFs are models to estimate soil properties using other available or more easily measured soil properties Bouma (1989). The use of PTFs is extensive, including filling gaps in soil databases (Wösten *et al.*, 2001), and soil mapping (Noble *et al.*, 2002; Scheinost *et al.*, 1997). They have been included in computer software like Rosetta (Schaap *et al.*, 2001), and the inclusion of these kinds of models into expert systems has also been discussed (McBratney *et al.*, 2002).

DSM modelling estimates a soil property using diverse information, including other soil properties at the same location, but also information related to soil forming factors. These predictors are also known as *scorpan* factors (McBratney *et al.*, 2003). Digital soil maps are meant to be continuous representations of the planet surface (more-or-less continuous depending on the map scale), hence data to generate them should try to capture the intrinsic heterogeneity. So far, the use of spatial environmental data is the most adequate alternative to represent this spatial (and temporal) heterogeneity, specially at national/continental scales. Environmental data is generally derived from sensors mounted on satellites, making it possible to detect a vast range of signals, from the radio-waves to gamma-rays (McBratney *et al.*, 2003). They have been widely used to represent scorpan factors and to model the spatial distribution of soil properties (Mulder *et al.*, 2011; Singh and Dwivedi, 1986).

Interdisciplinary data requirements

The demand to increase soil data generation rate comes from soil scientist but also from other disciplines where soil is an important factor. Need of soil data for soil carbon assessment (Zhang *et al.*, 2014), in ecology (Wigley *et al.*, 2013), and climate

studies (Khodayar *et al.*, 2013; Khodayar and Schädler, 2013) are some examples of areas where soil information is of critical importance. Soil scientists should not only meet this demand of data, but also promote the use of soil information in new research fields.

In response to the increasing demand of soil information, various research groups and governmental organisations have decided to supply data in a more organised manner. GlobalSoilMap (GSM, <http://globalsoilmap.net/>) is an initiative to provide accurate, up-to-date and spatially referenced soil information demanded by many stakeholders, including policymakers, the climate change community, farmers, other land users, and scientists in the form of a digital soil map of the world. Another example is the Terrestrial Ecosystem Research Network (TERN, <http://tern.org.au/>) project in Australia. TERN aims to connect ecosystem scientists and enables them to collect, contribute, store, share and integrate data across disciplines. One of the areas covered by TERN is soil, whose aims parallel those of the GSM specifications to generate maps of soil properties but with national coverage only.

Information quality

Soil information required by the scientific community is generated by different research groups or individuals, using different methods and datasets. How to ensure uniformity in the quality of the delivered models?

A key factor on how to properly deliver the new soil models is uncertainty assessment. In every modelling exercise, it is recommended to estimate the uncertainty associated with the predictions. It is important to understand how the errors propagate through the model, but especially because it is a way of evaluating the risk involved in using the predictions for a decision-making process (Goovaerts, 2001). In the soil science literature, the Monte Carlo method (Minasny and McBratney, 2002) has been most frequently used, and more recently empirical methods using the fuzzy k-means algorithm to generate prediction intervals has been suggested (Tranter *et al.*, (2010) for PTFs and Malone *et al.*, (2011) for DSM).

The aim of this thesis is to derive a framework for addressing soil data needs, using as example drained upper limit (DUL) and crop lower limit (CLL) in Australia, subdivided in two specific objectives:

1. Propose a workflow to generate pedotransfer functions (PTFs)
2. Obtain a continuous spatial prediction of available water content over Australia's "wheatbelt", using digital soil mapping techniques

References

- Bouma, J. 1989. Using soil survey data for quantitative land evaluation. *Advances in soil sciences*. 9: 177–213.
- Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma* 103 (1): 3–26.
- Khodayar, S. and Schädler, G. 2013. The impact of soil moisture variability on seasonal convective precipitation simulations. Part II: sensitivity to land-surface models and prescribed soil type distributions. *Meteorologische Zeitschrift* 22 (4): 507–526.
- Khodayar, S., Kalthoff, N., and Schädler, G. 2013. The impact of soil moisture variability on seasonal convective precipitation simulations. Part I: validation, feedbacks, and realistic initialisation. *Meteorologische Zeitschrift* 22 (4): 489–505.
- Malone, B., McBratney, A., and Minasny, B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160 (3): 614–626.
- McBratney, A., Mendonça Santos, M. d. L., and Minasny, B. 2003. On digital soil mapping. *Geoderma* 117 (1): 3–52.
- McBratney, A. B., Minasny, B., Cattle, S. R., and Vervoort, R. 2002. From pedotransfer functions to soil inference systems. *Geoderma* 109 (1–2): 41–73.
- Minasny, B. and McBratney, A. 2002. Uncertainty analysis for pedotransfer functions. *European Journal of Soil Science* 53 (3): 417–429.
- Mulder, V., De Bruin, S, Schaepman, M., and Mayr, T. 2011. The use of remote sensing in soil and terrain mapping—A review. *Geoderma* 162 (1): 1–19.
- Noble, A., Middleton, C, Nelson, P., and Rogers, L. 2002. Risk mapping of soil acidification under Stylosanthes in northern Australian rangelands. *Soil Research* 40 (2): 257–267.

- Schaap, M. G., Leij, F. J., and van Genuchten, M. T. 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of hydrology* 251 (3): 163–176.
- Scheinost, A., Sinowski, W, and Auerswald, K. 1997. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. *Geoderma* 78 (3): 129–143.
- Singh, A. and Dwivedi, R. 1986. The utility of LANDSAT imagery as an integral part of the data base for small-scale soil mapping. *International Journal of Remote Sensing* 7 (9): 1099–1108.
- Tranter, G., Minasny, B., and McBratney, A. 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades. *Soil Sci. Soc. Am. J.* 74 (6): 1967–1975.
- Wigley, B. J., Coetsee, C., Hartshorn, A. S., and Bond, W. J. 2013. What do ecologists miss by not digging deep enough? Insights and methodological guidelines for assessing soil fertility status in ecological studies. *Acta Oecologica* 51: 17–27.
- Wösten, J., Pachepsky, Y., and Rawls, W. 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology* 251 (3–4): 123–150.
- Zhang, X., Sahajpal, R., Manowitz, D. H., Zhao, K., LeDuc, S. D., Xu, M., Xiong, W., Zhang, A., Izaurrealde, R. C., Thomson, A. M., *et al.*, 2014. Multi-scale geospatial agroecosystem modeling: A case study on the influence of soil data resolution on carbon budget estimates. *Science of The Total Environment* 479: 138–150.

Chapter 1

Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system

Summary

The difference between the International (adopted by Australia) and the USDA/FAO particle-size classification system is the limit between silt and sand fractions (20 and 50 μm respectively). In order to work with pedotransfer functions generated under the USDA/FAO system with Australian soil survey data, a conversion should be attempted. The aim of this work is to improve prior models using larger data sets and a genetic programming technique, in the form of a symbolic regression. 2-50 μm fraction was predicted using a USDA data set which included both particle-size classification systems. The presented model reduced the RMSE (%) in 14.96 - 23.62% (word-based data set and Australian data set respectively), compared with the previous model.

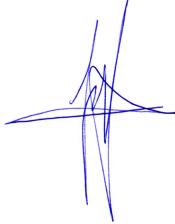
Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B., and McBratney, A. 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. Soil Research 50 (6): 443–446

Contributors	Statement of contribution
José Padarian Signature:  Date: October 1, 2014	Data analysis; Writing
Budiman Minasny	Data analysis; Writing
Alex McBratney	Data analysis; Writing

1.1 Introduction

Two major soil textural classifications are used in the world, the International and the USDA/FAO systems. The difference between these two systems is the limit between the silt and sand particle size: $20\mu m$ for the International and $50\mu m$ for the USDA/FAO. This could be considered a problem when a pedotransfer function (PTF) generated in one system is used with data of the other system, thus a conversion between both systems is necessary. Several attempts to achieve this has been made (Rousseva, 1997; Buchan, 1989; Shirazi *et al.*, 1988; Marshall, 1947). Minasny *et al.*, (1999) predicted the fraction P_{20-50} to convert from 2-20 to 2- $50\mu m$ fraction with the model:

$$\begin{aligned}\hat{P}_{20-50}(\%) = & 48.4593 - 0.2225P_{20-2000} - 0.0029(P_{20-2000})^2 \\ & - 0.6952P_{<2} + 0.0018(P_{<2})^2 \quad (R^2 = 0.76)\end{aligned}\quad (1.1)$$

where $P_{<2}$ and $P_{20-2000}$ correspond to clay and sand (International) fractions respectively.

In order to achieve better prediction, Minasny and McBratney (2001) used a larger data set than that used for Model 1.1 (Eq. 1.1), and generated a model using a multiple linear regression. The model was:

$$\begin{aligned}\hat{P}_{2-50}(\%) = & -18.3914 + 2.0971P_{2-20} + 0.6726P_{20-2000} - 0.0142(P_{2-20})^2 \\ & - 0.0049(P_{20-2000})^2 \quad (R^2 = 0.823)\end{aligned}\quad (1.2)$$

$$\text{If } \hat{P}_{2-50} < 0 \text{ then } \hat{P}_{2-50} = 0.8289P_{2-20} + 0.0198P_{20-2000}$$

This model was reported to produce unreasonable estimates at high clay and low sand contents. It is also a two-part model that produces an unnatural “break”. The aim of this work is to improve Model 1.2 (Eq. 1.2) with a new tool based on genetic programming.

1.2 Data sets

Three data sets were used in this work. How they were used and their size is shown in Table 1.1.

Table 1.1: Data sets used in this work

Data set	Reference	N° of records	Use
USDA/NRCS	Soil Survey Staff (1995)	104,864	Calibration
Australian (CSIRO)	-	758	Validation
IGBP-DIS	Tempel <i>et al.</i> , (1996)	55,282	Validation

The USDA/NRCS data set correspond to the National Soil Characterization database. The samples had data on soil texture measurements at < 2, 2-20, 20-50, 50-100, 100-250, 250-500, 500-1000 and 1000-2000 μm fractions. The Australian data set contains data from soil profile observations collected by CSIRO from various soil projects in Australia that had measurements of < 2, 2-20, 2-50, 20-200, and 200-2000 μm . The IGBP-DIS data set contains global data of soil properties that can be used for the development of pedotransfer functions with particle measurement at: < 2, 2-20, 20-50, 50-100, 100-250, 250-500, 500-1000 and 1000-2000 μm .

The USDA/NRCS and IGBP-DIS data sets were standardised to: < 2, 2-20, 2-50, 20-200, and 200-2000 μm . Particles < 200 μm were estimated from a log-linear interpolation between < 100 and < 250 μm .

All the outliers (outside the 2*inter-quartile range) and abnormal observations were removed. In Table 1.2 statistics of particle fractions are presented.

1.3 Genetic programming

Genetic programming (GP) is a machine-learning method for evolving computer programs, following the concepts of natural selection and genetics, to solve problems. GP is generally used to infer the underlying structure of a natural or experimental process in order to model it numerically. GP applications to soil science are varied. They range from determining soil characteristics (Parasuraman *et al.*, 2007b; Makkeasorn *et al.*, 2006), to water and nutrients management in agriculture (Sharma

Table 1.2: Statistics of data sets by particle fractions

Data set	Fraction	Mean	St. Dev.	Min.	Median	Max
USDA/NRCS	< $2\mu m$	23.14	16.35	0.00	20.60	97.90
	2- $20\mu m$	21.19	12.76	0.00	20.30	93.80
	20- $2000\mu m$	55.64	23.42	0.00	55.70	100.00
CSIRO ¹	< $2\mu m$	31.21	17.34	3.20	27.00	77.70
	2- $20\mu m$	18.01	8.56	0.60	22.00	58.90
	20- $2000\mu m$	50.77	18.87	4.60	53.00	96.20
IGBP-DIS	< $2\mu m$	23.07	16.30	0.00	20.50	95.00
	2- $20\mu m$	20.98	12.60	0.00	20.10	93.80
	20- $2000\mu m$	55.96	22.79	0.30	56.30	100.00

*All statistics in percentage of mass basis

¹ National soil database

and Jana, 2009; Ines *et al.*, 2006), to development of PTFs (Parasuraman *et al.*, 2007a; Johari *et al.*, 2006).

In a recent work, Selle and Muttill (2011) test the structure of a hydrological model using GP and give a good description of how the GP process works.

GP works with a number of solution sets, known collectively as a “population”, rather than a single solution at any one time; thus the possibility of getting trapped in a “local optimum” is avoided. GP differs from the traditional genetic algorithms in that it typically operates on “parse trees” instead of bit strings. A parse tree is built up from a “terminal set” (the input variables in the problem and randomly generated constants, i.e. empirical model coefficients) and a “function set” (the basic operators used to form the GP model). The function set is user-defined and cannot only include algebraic operators, such as $\{+, -, *, \%\}$ but can also take the form of logical rules ($\{IF, OR, AND\}$) or more complex operators ($\{\sin, \cos, \exp\}$). An example of an initial population of parse trees can be found in Fig. 1.1.

Once the initial population of random parse trees is generated, GP calculates their fitness using the user-defined “fitness function”, e.g. absolute error, and subsequently selects the better parse trees for reproduction and variation to form a new population. This process of selection, reproduction and variation iterates until a user-defined “stopping criterion” is satisfied. The solutions in each iteration are collectively known

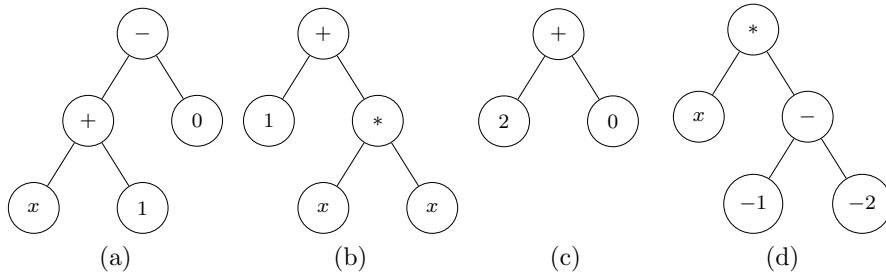


Fig. 1.1: Example of an initial population of four randomly created individuals representing GP models: (a) $x + 1$, (b) $x^2 + 1$, (c) 2 and (d) x . This representations should be read from left to right and bottom to top.

as a “generation”. As the population evolves from one generation to another, new solutions replace the older ones and are supposed to perform better. The solutions in a population associated with the best-fit individuals will, on average, be reproduced more often than the less-fit solutions. This is known as the Darwinian principle of “natural selection”.

During each successive generation a proportion of the existing population is “selected” to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions are typically more likely to be selected. The next step is to generate a second generation population of solutions from those selected, through the two variation operators —crossover and mutation. Crossover is the random swapping of sub-trees between the selected “parent” parse trees to generate the new “children”. The crossover tends to enable the evolutionary process to move toward promising regions of the solution space. In contrast to crossover, in mutation, a single parent parse tree is selected and random changes are made to it. The mutation operator is introduced to prevent premature convergence to local optima. A high crossover rate is usually used so that useful sub-trees from the previous generations are transmitted to the new generation. In contrast, the mutation rate is usually kept low since a high mutation rate can cause a big loss of useful sub-trees evolved in previous generations. This process of selection, reproduction and variation continues until a new population of solutions of appropriate size is generated. From generation to generation, the best solution evolved in previous generations is usually preserved, a process called “elitism”.

In this work we used a specific method called symbolic regression, which uses GP to fit a function to a specific data set, going from simple functions like those in Fig. 1.1 to a complex function like the solution proposed (Eq. 1.4).

For further reading about genetic programming, see Koza *et al.*, (1999) and Koza (1994) and Koza (1992).

1.4 Particle-size conversion

In a routine soil survey in Australia, particle-size could be measured at clay, silt and sand fractions (< 2 , $2-20$, $20-2000\mu m$) or with an extra intermediate fraction of fine sand ($20-200\mu m$). A symbolic regression was attempted (using the program *Formulize v0.96b*) for both cases, using $F = \{+, -, *, \%\}$ as the function set for the genetic programming routine, generating a model:

$$P_{2-50} = F(P_{frac}) + \epsilon$$

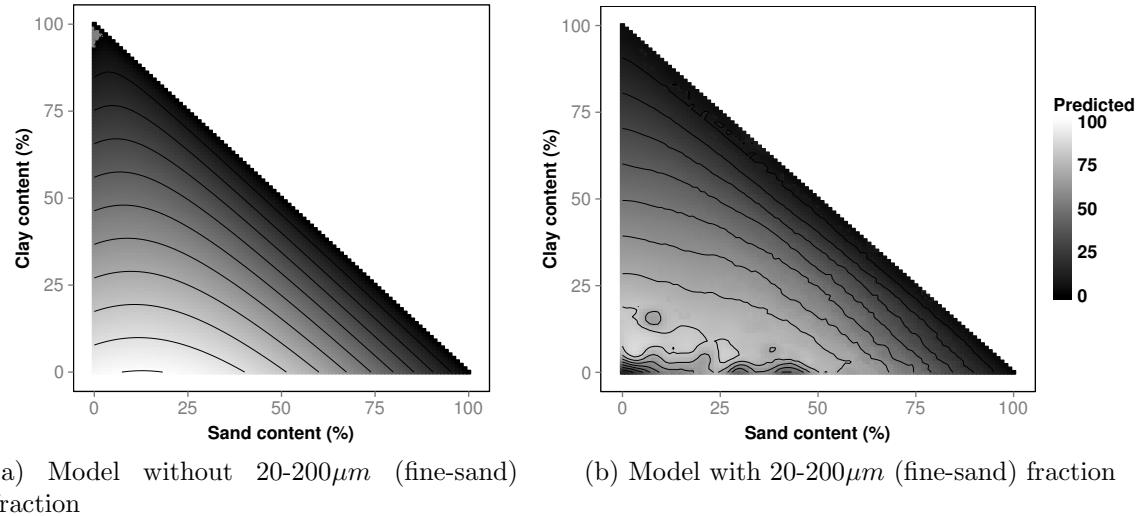
with P_{frac} as the available particles fractions of the Australian classification system, expressed in percentage, and ϵ as the error of prediction. Data was randomly split in two groups (50% for training and 50% for internal validation) and, minimising the absolute error as error metric, we obtained an approximate conversion as:

$$\hat{P}_{2-50}(\%) = 2.26P_{2-20} + \frac{5.55P_{2-20} + 1.513(P_{2-20})^2}{0.9966 - 1.236P_{2-20} - 1.349P_{20-2000}} \quad (1.3)$$

for survey data without the $20-200\mu m$ fraction, presenting an R^2 of 0.82 and a root mean squared error (RMSE), which measures the average error of the prediction, of 8.54% (internal validation). A surface plot of its predictions as a function of clay ($< 2\mu m$) and sand ($20-2000\mu m$) is shown in Fig. 1.2a. For survey data with measured $20-200\mu m$ (fine-sand) fraction a different solution was generated:

$$\begin{aligned} \hat{P}_{2-50}(\%) = & 1.561 + 0.9664P_{2-20} + 0.0003932P_{<2}P_{2-20}P_{20-200} \\ & + 0.0003634P_{2-20}(P_{20-200})^2 \end{aligned} \quad (1.4)$$

with an R^2 of 0.91 and a RMSE of 5.91% (internal validation). A surface plot of its predictions as a function of clay ($< 2\mu m$) and sand ($20-2000\mu m$) is shown in Fig. 1.2b.



(a) Model without $20-200\mu m$ (fine-sand) fraction
 (b) Model with $20-200\mu m$ (fine-sand) fraction

Fig. 1.2: Surface plot of $2-50\mu m$ fraction prediction at different clay ($< 2\mu m$) and sand ($20-2000\mu m$) contents. Note that $20-200\mu m$ represents the fine-sand fraction.

The surface plot of Eq. 1.3 (Fig. 1.2a) shows decreasing predictions of the $2-50\mu m$ fraction as the content of clay ($< 2\mu m$) or sand ($20-2000\mu m$) increases, with a slightly higher responsiveness to changes in sand content. The model including the $20-200\mu m$ fraction (Eq. 1.4; Fig. 1.2b) shows the same trend, but presenting some instability at high silt contents, also evident in the surface plot of the residuals Fig. 1.2b.

Table 1.3 presents the RMSE and R^2 between predicted and measured values in the external validation sets and a comparison with the previous model (Eq. 1.2).

Comparing with the model of Minasny and McBratney (Eq. 1.2), this work has a better performance when the $20-200\mu m$ fraction data is available. The model presents some limitations (higher absolute error) at low clay and high sand contents as shown in Fig. 1.3.

1.5 Conclusions

The use of a larger data set in conjunction with genetic programming techniques reduced RMSE (%) by 14.96% (from 8.69 to 7.39) in the IGBP-DIS data set and

Table 1.3: External validation statistics of prediction quality

Data set	Model	R^2	RMSE (%)
CSIRO	Minasny and McBratney (Eq. 1.2)	0.52	10.67
	without 20-200 μm (Eq. 1.3)	0.48	11.19
	with 20-200 μm (Eq. 1.4)	0.72	8.15
IGBP-DIS	Minasny and McBratney (Eq. 1.2)	0.81	8.69
	without 20-200 μm (Eq. 1.3)	0.81	8.66
	with 20-200 μm (Eq. 1.4)	0.86	7.39

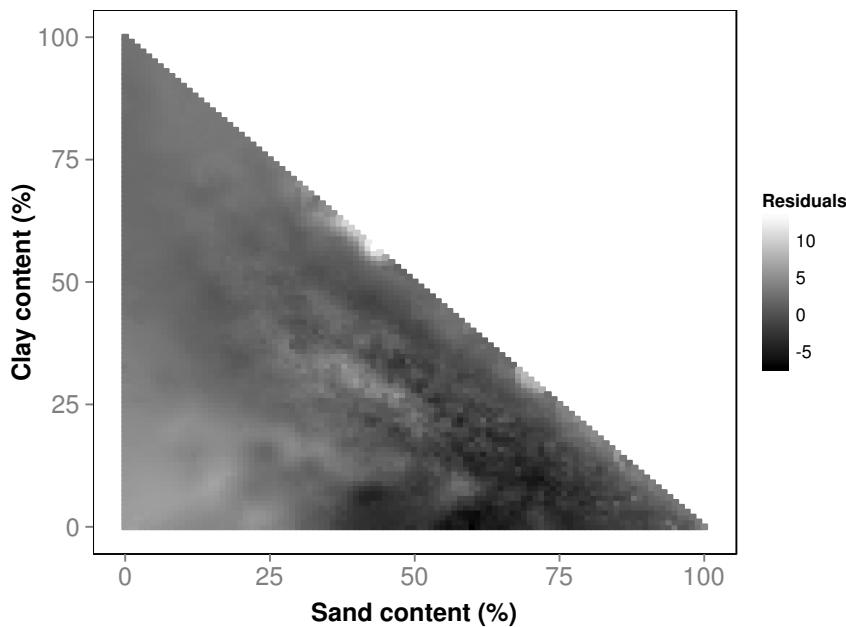


Fig. 1.3: Surface plot of residuals of Eq. 1.4, as a function of clay and sand content, using IGBP-DIS data set.

23.62% (from 10.67 to 8.15) in Australian data set, compared with the previous model of Minasny and McBratney (2001).

1.6 References

- Buchan, G. 1989. Applicability of the simple lognormal model to particle-size distribution in soils. *Soil Science* 147 (3): 155–161.
- Ines, A. V., Honda, K., Gupta, A. D., Droogers, P., and Clemente, R. S. 2006. Combining remote sensing-simulation modeling and genetic algorithm optimization to explore water management options in irrigated agriculture. *Agricultural Water Management* 83 (3): 221–232.
- Johari, A., Habibagahi, G., and Ghahramani, A. 2006. Prediction of Soil-Water Characteristic Curve Using Genetic Programming. *Journal of Geotechnical and Geoenvironmental Engineering* 132 (5): 661–665.
- Koza, J. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. *The MIT Press*.
- Koza, J. 1994. Genetic Programming II: Automatic Discovery of Reusable Subprograms. *The MIT Press*.
- Koza, J., Bennett, H., Andre, D., and Keane, M. 1999. Genetic Programming III: Darwinian Invention and Problem Solving. *Morgan Kaufmann Publishers*.
- Makkeasorn, A., Chang, N., Beaman, M., Wyatt, C., and Slater, C. 2006. Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. *Water Resources Research* 42 (9) W09401: W09401.
- Marshall, T. 1947. Mechanical composition of soil in relation to field descriptions of texture. *Council for Scientific and Industrial Research: Bulletin No. 224*. Melbourne, Australia.
- Minasny, B. and McBratney, A. 2001. The australian soil texture boomerang: a comparison of the australian and USDA/FAO soil particle-size classification systems. *Aust. J. Soil Res.* 39 (6): 1443–1451.
- Minasny, B., McBratney, A., and Bristow, K. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma* 93 (3–4): 225–253.
- Padarian, J., Minasny, B., and McBratney, A. 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. *Soil Research* 50 (6): 443–446.

- Parasuraman, K., Elshorbagy, A., and Si, B. C. 2007a. Estimating Saturated Hydraulic Conductivity Using Genetic Programming. *Soil Sci. Soc. Am. J.* 71 (6): 1676–1684.
- Parasuraman, K., Elshorbagy, A., and Carey, S. K. 2007b. Modelling the dynamics of the evapotranspiration process using genetic programming. *Hydrological Sciences Journal* 52 (3): 563–578.
- Rousseva, S. 1997. Data transformations between soil texture schemes. *European Journal of Soil Science* 48 (4): 749–758.
- Selle, B. and Muttgil, N. 2011. Testing the structure of a hydrological model using Genetic Programming. *Journal of Hydrology* 397 (1-2): 1–9.
- Sharma, D. K. and Jana, R. 2009. Fuzzy goal programming based genetic algorithm approach to nutrient management for rice crop planning. *International Journal of Production Economics* 121 (1): 224–232.
- Shirazi, M., Boersma, L., and Hart, J. 1988. A unifying quantitative analysis of soil texture: improvement of precision and extension of scale. *Soil Science Society of America Journal* 52 (1): 181–190.
- Soil Survey Staff. 1995. Soil Characterization and Profile Description Data. *Soil Survey Laboratory, Natural Resources Conservation Service, USDA, Lincoln, NE, USA.*
- Tempel, P., Batjes, N., and van Engelen, V. 1996. IGBP-DIS soil data set for pedotransfer function development. *Int. Soil Reference and Inf. Centre, Wageningen, Netherlands.*

Chapter 2

Provision of soil water retention information for biophysical modelling: an example for Australia

Summary

Soil is an important actor in ecosystem processes and data that represent soil processes in system models are not always available due to the intrinsic complexity and variability of soil over space. A frequently used method to overcome this problem is the use of pedotransfer functions (PTFs). We suggest the use of domain-specific PTFs with defined uncertainty levels to avoid erroneous predictions or extrapolation. PTFs with detailed uncertainty assessment are not always available, most of the time providing a single measurement (i.e.: standard error, variance), hence there is a necessity to generate new ones, with more detailed uncertainty assessment, and to identify if a PTF prediction is valid for a given soil domain. We selected Australia as example to generate a set of pedotransfer functions which predict soil water retention properties required by commonly-used biophysical models. PTFs were generated using symbolic regression and the fuzzy k-means with extragrades algorithm was used to estimate the uncertainty of prediction and identify when an observation is within the PTF data domain.

2.1 Introduction

Soil is a intrinsically complex system and an important component in ecosystem processes. There are many dynamic soil properties and trying to measure all of them would be a challenging task. This data collection activity is usually the most expensive and time consuming step in the ecological modelling process. A frequently used method to overcome this soil data availability problem is the use of pedotransfer functions (PTFs), term coined by Bouma (1989) as “translating (soil) data we have into what we need”, to estimate soil properties using other available or more easily measured soil properties.

Natural systems, including soil, vary in time and space (Frank and Slatkin, 1990) and a PTF should be able to consider this uncertainty. While many PTFs have been generated (see reviews by McBratney *et al.*, (2002) and Wösten *et al.*, (2001)), it has not been general practice to provide uncertainty levels for them. Taking in account this intrinsic uncertainty, it is recommended that a given PTF should not be extrapolated beyond the geomorphic region or soil type from which it was developed (McBratney *et al.*, 2002), since they may lose their validity (Minasny *et al.*, 1999). Hence the importance of having a *domain-specific* set of PTFs to prevent their misuse and avoid erroneous predictions and extrapolation.

The aim of this work is to propose a workflow to address the two issues: *a)* generation of PTFs with the corresponding uncertainty estimation, presented as upper and lower prediction limits, and *b)* identification of observations outside the data domain of the generated PTFs.

2.2 Soil data requirements of biophysical models

Soil intrinsic complexity and its interactions with the gaseous and liquid phases of the ecosystem, and the biota, makes it a subject of study in different disciplines. Biophysical models try to represent these interactions and we grouped them in the following domains.

Crop growth: Crop-growth models try to represent the soil-plant-atmosphere system taking into account this soil and water interaction, which depends on particle size,

hydraulic characteristics, and morphological and chemical properties (Rawls *et al.*, 1991). Soil water content is critical for agricultural production, thus, the amount of water stored in soil and its availability in time is an important factor in the decision-making process.

Models developed in this area try to estimate the amount of soil water stored in the soil and the nutrients extracted from soil solution into the plant. These nutrients come from diverse processes, including rock weathering and organic matter decomposition. The cycling of nutrients is governed by mass balances between dissolution and precipitation happening in the aqueous phase (Garrels and Mackenzie, 1967).

Watershed erosion: Topography is one of the factor involved in soil formation and also determines the path for surface runoff. Depending on the erodibility, infiltration rate and water retention capacity of the soil, a precipitation event could start the erosion process, leading to the transport of soil material to lower areas of the landscape or river basins. This soil loss has an tremendous impact in agricultural productivity, economy, and environment and models try to simulate different scenarios, including, for example, management practices and soil types.

Ecology: One of the areas covered by this discipline is associated with microbial activity and the transformations they could generate within the soil. This activity governs processes like biodegradation (pesticides), and carbon, nitrogen and phosphorous cycles and depends on the presence of adequate environmental conditions for microbial colonies growth (Han *et al.*, 2007; Buchmann, 2000).

Other important issues covered by this discipline is ecotoxicity (heavy metals) and the capacity of the soil to immobilise these compounds. Pampura *et al.*, (2007) highlight the importance of heavy metals (cadmium and lead) availability in the soil solution compared with the total metal concentration. They propose that this availability is ruled by soil properties like pH and organic matter content. Kuo and Baker (1980) include the clay content of soil as an important factor in the detoxification.

Climate: Is important to remember the constant interaction between soil and the atmosphere. Soil is an important water stock and, for that reason, influences

processes like evapotranspiration and the subsequent precipitation. Soil moisture information has been shown to enhance the prediction of precipitation and atmospheric circulations, modifying the distribution and intensity of precipitation (Walker and Houser, 2001). This relation is implicit in models like the one proposed by Rodriguez-Iturbe *et al.*, (1991), explaining the influence in local water recycling, or studies of mutual interaction (Entekhabi *et al.*, 1996).

2.2.1 Australian biophysical models

We reviewed 17 biophysical models commonly used in Australia and identified the soil properties required to successfully use them, either as inputs or internal values (Table 2.1). These models were generated and calibrated within Australia, putting into practice the previously mentioned precaution with model transportability. The description and purpose of the models reviewed can be found in Appendix A.

For the following PTF generation, and fulfilling the principle of *effort* proposed by McBratney *et al.*, (2002), which states that the cost and the effort to obtain the information on the predictor should be much less than that to obtain information on the predicted, eight soil properties were selected also taking into account its occurrence in the different models reviewed in this work (Table 2.2).

In the forthcoming sections we will demonstrate how to fill soil data hiatus using a genetic programming method to generate PTFs, and the fuzzy k-means algorithm to estimate the uncertainty of the generated PTFs. Undoubtedly, water is an important component of the ecosystem and most of the models reviewed take in account its presence as a regulator of chemical, biological and physical processes. That is the reason why, as an example, we focus in the properties which describe the water holding capacity of soils.

2.3 Prediction of soil water retention properties

Soil properties predicted in this work are drained upper limit DUL and CLL, both corresponding to field measurements. Soil water holding capacity (i.e.: the difference between DUL and CLL) is the main source of water for vegetation development and it is related to the potential amount of water a soil could make available for the atmosphere

Table 2.1: Some biophysical models commonly used in Australia and related soil properties

	Bulk density	Clay content	Silt content	Sand content	K_{sat}	Organic carbon	CLL or θ_{-1500}	DUL or θ_{-10}	SAT	AirD	pH	$K_{erosion}$	Curve number	Depth to BR	Salinity	CEC	Nitrogen content	Phosphorus content
PERFECT	✓																	
Hydrology																		
Mineralisation																		
Demitrification																		
Erosion																		
SWAT																		
Infiltration	✓	✓	✓															
Percolation	✓	✓	✓															
Nutrient cycle	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$K_{erosion}$																		
SedNet																		
$K_{erosion}$																		
FullCAM																		
Carbon stock	✓	✓																
APSIM																		
Water balance																		
Nutrient cycle																		
Mk3.5																		
Soil moisture																		
Soil temperature	✓																	
CENTURY	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DNDC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CLASS																		
CGM																		
PGM																		
SA																		
U3M-1D																		
DSSAT																		
OZCOT																		
GDAY																		
BIOME4																		
LPX																		
BC2C																		

K_{sat} : saturated hydraulic conductivity; DUL: drained upper limit; AirD: water content after air drying

CLL: crop lower limit; $K_{erosion}$: soil erodibility factor; CEC: cation exchange capacity; BR: bedrock

WRC: water retention curve; SAT: water content at saturation; $\theta_{-10/-1500}$: water content at $-10/-1500$ kPa

through evapotranspiration (Dunne and Willmott, 1996).

DUL, a practicable field measure of soil field capacity, represents the volumetric water content an initially saturated soil holds after draining for 2-3 days (Veihmeyer and Hendrickson, 1949). One of the issues related to this “steady state” is that it

Table 2.2: Soil properties commonly used in reviewed models and predictors mentioned in literature

Property	Predictors	Reference
BD	clay, silt, sand, OC, depth	Tranter <i>et al.</i> , (2007)
$\theta_{-10}, \theta_{-1500}$	clay, silt, sand, BD, OC	Rab <i>et al.</i> , (2011), Selle <i>et al.</i> , (2011), and Rawls <i>et al.</i> , (1982)
DUL, CLL	PSD, BD	Nemes <i>et al.</i> , (2011) and Romano <i>et al.</i> , (2011)
OC	clay, silt, colour	Viscarra Rossel <i>et al.</i> , (2006), Zinn <i>et al.</i> , (2005), and Schimel <i>et al.</i> , (1994)
K_{sat}	clay, silt, sand, BD	Minasny and McBratney (2000)
K_{erosion}	clay, sand, silt, OC	Torri <i>et al.</i> , (1997) and Williams (1995)

DUL: Drained upper limit; CLL: Crop lower limit; BD: Bulk density.

OC: Organic carbon; PSD: Particle size distribution

$\theta_{-10/-1500}$: Water content at $-10/-1500$ kPa.

K_{erosion} : soil erodibility factor; K_{sat} : Saturated hydraulic conductivity.

varies from soil to soil, strongly depending on soil properties like texture and structure (i.e. soil pore system). In practice, the previously mentioned 2-3 day drained period is only applicable to soils with uniform structure and texture, and that period may be extended, for example, to 3-6 months in a clayey soil (accessory publication of Dalglish *et al.*, 2009). Nachabe (1998) defined it as an expression which depends on water retention and hydraulic conductivity, field capacity is assumed when the water flux is equal to 0.05 mm day^{-1} . Meanwhile Twarakavi *et al.*, (2009) estimated field capacity based on Richards' equation and developed an analytical equation to predict field capacity from soil hydraulic parameters. Field capacity is assumed when the drainage flux is equal to 0.1 mm day^{-1} .

On the other hand, CLL corresponds to the volumetric soil water remaining in the soil after a healthy crop, with uninterrupted root development, has reached maturity under soil water-limited conditions (Hochman *et al.*, 2001). It depends on the ability

of the crop to extract water, but in practice is assumed to be the minimum of a group of crops. This concept is usually referred as permanent wilting point.

2.3.1 Data sets

The data set used correspond to a CSIRO Ecosystem Sciences (APSRU) compilation of 806 soil profiles that includes field measurements of DUL and CLL for the most commonly grown crops of Australia (Dalgliesh *et al.*, 2012). Procedures for determination of these properties are described in the accessory publication of the article by Dalgliesh *et al.*, (2009), “Procedures for determination of soil properties and states relevant to crop simulation and farmer crop management decision making”. The method is a modification of the techniques described by Ratliff *et al.*, (1983). Briefly, an area covering about 16 m^2 of soil was wetted using a trickle system. The water content and drainage were monitored using a neutron moisture meter at the access tube at the centre of the site down to a depth of 180 cm. Once the soil was judged to be thoroughly wet, it was allowed to drain until moisture monitoring indicated minimal change in profile water status. Samples for gravimetric moisture content and bulk density were taken. For CLL, crops were grown in the field, and a rain-exclusion tent of 9 m^2 was installed. At crop maturity, soil moisture were determined at different depths.

The soil properties used to generate PTFs and their statistics are presented in Table 2.3.

Table 2.3: Statistics of soil samples used for PTF generation of field measurements.

	Mean	S.D.	Min.	Median	Max.
Clay (%)	35.20	16.76	0.80	35.40	80.20
Sand (%)	54.20	20.90	9.00	51.00	97.00
BD (Mg m^{-3})	1.45	0.18	0.73	1.45	2.09
OC (%)	0.47	0.47	0.01	0.30	7.26
DUL (%)	30.20	11.54	3.00	32.00	56.00
CLL (%)	16.90	8.61	0.40	18.00	53.00

The soil orders according to the Australian Soil Classification System in this database correspond to Calcarosol (4.22%), Chromosol (4.96%), Dermosol (2.23%),

Ferrosol (0.99%), Kandosol (2.23%), Podosol (0.12%), Sodosol (5.21%), Tenosol (0.87%), Vertosol (22.08%), and 57.07% of unclassified soils. Based on the location of the unclassified soils and the dominant soil order map of Australia (ASRIS), they correspond to Dermosol (10%), Ferrosol (0.87%), Hydrosol (1.09%), Kandosol (32.61%), Kurosol (6.3%), Organosol (27.17%), Podosol (1.52%), Rudosol (0.87%), Sodosol (11.09%), Tenosol (0.43%), and Vertosol (8.04%).

We also used an Australian soil hydraulic properties database, compiled by Minasny *et al.*, (1999), from laboratory measurements of soil hydraulic properties throughout Australia. It includes 1403 soil samples collected using undisturbed soil cores, and measured in the laboratory for water retention at -10 and -1500 kPa using the pressure plate apparatus. These laboratory measurements of soil water content are usually assumed as equivalencies of DUL/field capacity (-10 kPa is the standard in Australia) and CLL/permanent wilting point respectively (White, 2009), thus the interest in comparing them in this work. The associated statistics are shown in Table 2.4.

Table 2.4: Statistics of soil samples used for PTF generation of laboratory measurements.

	Mean	S.D.	Min.	Median	Max.
Clay (%)	31.60	17.84	1.00	29.00	76.00
Sand (%)	50.50	22.11	6.49	50.40	97.90
BD (Mg m^{-3})	1.44	0.22	0.56	1.47	2.18
θ_{-10} (%)	33.20	9.60	8.00	33.00	70.00
θ_{-1500} (%)	18.40	9.07	1.80	18.00	48.10

External validation of the PTFs was performed using a database compiled by Gardner *et al.*, (1984). It contains properties of 628 horizons of soils located in Brisbane and Darling Downs area (Table 2.5) where DUL and CLL were measured in the field.

2.3.2 PTF development

We used symbolic regressions to model soil properties related to water retention (using the software *Formulize v0.98.1b*). Symbolic regression uses genetic programming (GP) to fit a function to a specific data set. It is a machine-learning method for evolving computer programs, following the concepts of natural selection and genetics, to solve

Table 2.5: Statistics of soil samples used for PTFs' external validation.

	Mean	S.D.	Min.	Median	Max.
Clay (%)	47.90	17.01	7.84	51.60	83.50
Sand (%)	38.70	18.66	7.37	34.00	86.30
BD (Mg m^{-3})	1.42	0.16	0.90	1.45	1.74
DUL (%)	27.50	10.17	7.00	27.50	58.00
CLL (%)	19.30	7.76	3.00	19.00	40.00

problems. GP is generally used to infer the underlying structure of a natural or experimental process in order to model it numerically. GP applications to soil science are varied. They range from determining soil characteristics (Parasuraman *et al.*, 2007b; Makkeasorn *et al.*, 2006), to water and nutrients management in agriculture (Sharma and Jana, 2009; Ines *et al.*, 2006), to development of PTFs (Parasuraman *et al.*, 2007a; Johari *et al.*, 2006).

In genetic programming, possible solutions (individuals) are typically represented as “parse trees” (Fig. 2.1), with nodes corresponding to basic algebraic operators such as $\{+, -, *, \%\}$, logical rules like $\{\text{IF}, \text{OR}, \text{AND}\}$ or more complex operators like $\{\sin, \cos, \exp\}$, the input variables of the function, or numerical constants. An initial random population of this individuals is generated and their fitness is assessed using a user-defined “fitness function”, e.g. absolute error, and subsequently the best individuals are selected to be the basis of the next generation. The “fittest” individuals are subjected to a mutation and crossover processes (random change of a random node and exchange of “branches” between individuals respectively) to introduce variation into the population as it evolves.

For further reading about genetic programming, see Koza *et al.*, (1999) and Koza (1994) and Koza (1992).

2.3.3 Uncertainty estimation

To assess the uncertainty of our predictions, we used a modification of the method by Shrestha and Solomatine (2006). The classic k-means clustering algorithm assumes each observation belongs to only one cluster. This approach seems inappropriate because, in a real world context, most ecosystem processes are continuous. Fuzzy

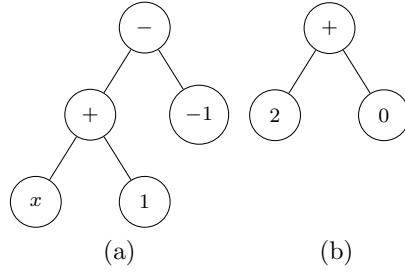


Fig. 2.1: Example of an initial population of two randomly created individuals representing GP models: (a) $x + 2$ and (b) 2. This representations should be read from left to right and bottom to top.

k-means extends the classic approach allowing each observation to belong to more than one cluster through a membership degree concept (Bezdek *et al.*, 1984).

One of the limitations of fuzzy k-means is the inability to distinguish between points very far from cluster centroids (extragrades) and those close to them. De Gruijter and McBratney (1988) proposed a modified method whereby a new extragrade class was introduced, leading to a membership degree dependent on the distance to cluster centroids.

The membership for the observation i in the j th class (m_{ij}) and the membership in the extragrade class (m_{i*}) are estimated using the following formulas:

$$m_{ij} = \frac{d_{ij}^{-2/(\phi-1)}}{\sum_{c=1}^k d_{ic}^{-2/\phi-1} + \left(\lambda \sum_{c=1}^k d_{ic}^{-2} \right)^{-1/(\phi-1)}} \quad (2.1)$$

$$m_{i*} = \frac{\left(\lambda \sum_{c=1}^k d_{ic}^{-2} \right)^{-1/(\phi-1)}}{\sum_{c=1}^k d_{ic}^{-2/\phi-1} + \left(\lambda \sum_{c=1}^k d_{ic}^{-2} \right)^{-1/(\phi-1)}} \quad (2.2)$$

where d_{ij} correspond to the Mahalanobis distance between the observation i and the centroid c of the j th class, k is the total number of classes (not including the extragrade class), ϕ the degree of fuzziness or overlap of clusters, and $\lambda = (1 - \alpha)/\alpha$ (α : mean value of extragrade class membership of the observations).

Data used to calibrate a PTF is a representative sample of an existent natural process and every re-sample generates a fluctuation around the true, unknown value of its properties, slightly changing the output of the model. Due to this phenomenon, it is more appropriate to predict a range than a single value. To estimate that range (from here on prediction interval (PI)), we used the fuzzy k-means with extragrades algorithm, as described by Tranter *et al.*, (2010). In the calibration process, the $\alpha/2$ and $1-\alpha/2$ (α : significance level) quantiles of the prediction residuals by class (cluster) were determined. Those values were weighted by the membership degree and added to the predicted value.

The key concept of this approach is that assumptions and outcomes of PTFs are only valid inside the domain of the calibration data. Any observation outside this domain is assigned to the extragrade class and its PI penalised (extended). The prediction for this observation is not necessarily wrong but the final user must be aware that the data used to train the PTF is different.

To run the fuzzy k-means algorithm, we set the fuzziness parameter ϕ to 1.5, using Mahalanobis distance metric. The α parameter was obtained by optimisation, aiming to reach an expected extragrade proportion of 5%.

Data was partitioned in different number of clusters (2 to 15) and the optimal number of them was determined calculating the prediction interval coverage probability (PICP, Eq. 2.3), which is the proportion of observations that lie within the PI, and the mean prediction interval (MPI, Eq. 2.4), both described by Shrestha and Solomatine (2006), using the following equations:

$$PICP = \frac{1}{n} \text{count}(a) \quad (2.3)$$

$$a : PL_i^L \leq p_i \leq PL_i^U$$

$$MPI = \frac{1}{n} \sum_{i=1}^n [PL_i^U - PL_i^L] \quad (2.4)$$

where n is the total number of observations, p_i is the i th observed value, PL_i^L and PL_i^U are the i th lower and upper prediction limit respectively. To select the optimal

number of clusters the value of PICP should be close to the confidence interval (95% in this case) and the value of MPI should be minimum (i.e. if two possible number of clusters have a similar PICP value, we must select the one with lower MPI value).

2.4 Results

2.4.1 Drained upper limit

We generated five PTFs to predict DUL ($\hat{\theta}_{DUL}$), using different input variables, where the less accurate ones correspond to simpler alternatives to be used (as described in Section 2.5) when data availability is limited. The resulting PTFs are:

$$\begin{aligned}\hat{\theta}_{DUL} (cm^3/100cm^3) = & 0.2739 + 0.005033 clay + 3.158 \times 10^{-5} sand CEC \\ & - 1.96 \times 10^{-5} sand^2 - 0.00256 clay BD\end{aligned}\quad (2.5)$$

with BD corresponding to bulk density, CEC to the cation exchange capacity, and $clay$ and $sand$ to the $< 2\mu m$ and $20-2000\mu m$ fraction of the soil respectively, with R^2 value of 0.76 and root mean square error (RMSE) of 4.39 (%);

$$\begin{aligned}\hat{\theta}_{DUL} (cm^3/100cm^3) = & 0.2358 + 0.002572 CEC + 0.001001 clay \\ & - 1.70 \times 10^{-7} sand^3\end{aligned}\quad (2.6)$$

with an R^2 of 0.75 and RMSE equal to 4.53 (%).

$$\begin{aligned}\hat{\theta}_{DUL} (cm^3/100cm^3) = & 0.374 + 0.01182 BD + 0.00365 clay \\ & + 6.09 \times 10^{-5} sand clay \\ & - 0.00339 sand - 0.00192 BD^2 clay\end{aligned}\quad (2.7)$$

with R^2 value of 0.75 and root mean square root (RMSE) of 4.63 (%);

$$\hat{\theta}_{DUL} (cm^3/100cm^3) = 0.2082 + 0.02757 OC + 0.002666 clay - 1.73 \times 10^{-7} sand^3 \quad (2.8)$$

where OC correspond to the soil organic carbon content in percentage, with R^2 and RMSE values of 0.71 and 4.92 (%) respectively; and

$$\hat{\theta}_{DUL} (cm^3/100cm^3) = 0.364 + 4.828 \times 10^{-5} sand clay - 0.00296 sand \quad (2.9)$$

with R^2 equal to 0.7 and RMSE of 4.98 (%). All the R^2 and RMSE values correspond to an internal validation.

2.4.2 Water content at -10 kPa

As with DUL, we generated PTFs with different input variables. The resulting PTFs are:

$$\hat{\theta}_{-10} (cm^3/100cm^3) = 0.5255 - 2.76 \times 10^{-5} sand^2 - 0.05195 BD^2 \quad (2.10)$$

with R^2 and RMSE values of 0.67 and 5.37 (%) respectively; and a simpler version using just particle size information:

$$\hat{\theta}_{-10} (cm^3/100cm^3) = 0.4795 - 3.873 \times 10^{-5} sand^2 - 6.701 \times 10^{-7} clay^2 sand \quad (2.11)$$

with an R^2 of 0.6 and RMSE equal to 5.96 (%).

2.4.3 Relation between DUL and -10

To establish the behaviour of these two “equivalent” water contents, we used Eq. 2.9 ($\hat{\theta}_{DUL}$) and Eq. 2.11 ($\hat{\theta}_{-10}$) and data described in Table 2.3. Both methods show a similar behaviour with particle size changes, with a general over-prediction and slightly higher dispersion of $\hat{\theta}_{DUL}$ (Fig. 2.2). Similar results were obtained when we calculated water content at the flux of $0.1 mm day^{-1}$ as proposed by Twarakavi *et al.*, (2009):

$$\hat{\theta}_{fc} (L^3/L^3) = n^{-0.6 \log_{10}(K_s)} (\theta_s - \theta_r) + \theta_r \quad (2.12)$$

where θ_{fc} is the water content at field capacity (flux=0.01 $cmday^{-1}$), n van Genuchten's shape parameter, K_s saturated hydraulic conductivity (in $cmday^{-1}$), θ_s the saturated water content and θ_{r} the residual water content (Fig. 2.2d). The results showed that water content at -10 kPa commonly used in Australia is too high for the DUL estimate and depends also on the hydraulic conductivity of the soil.

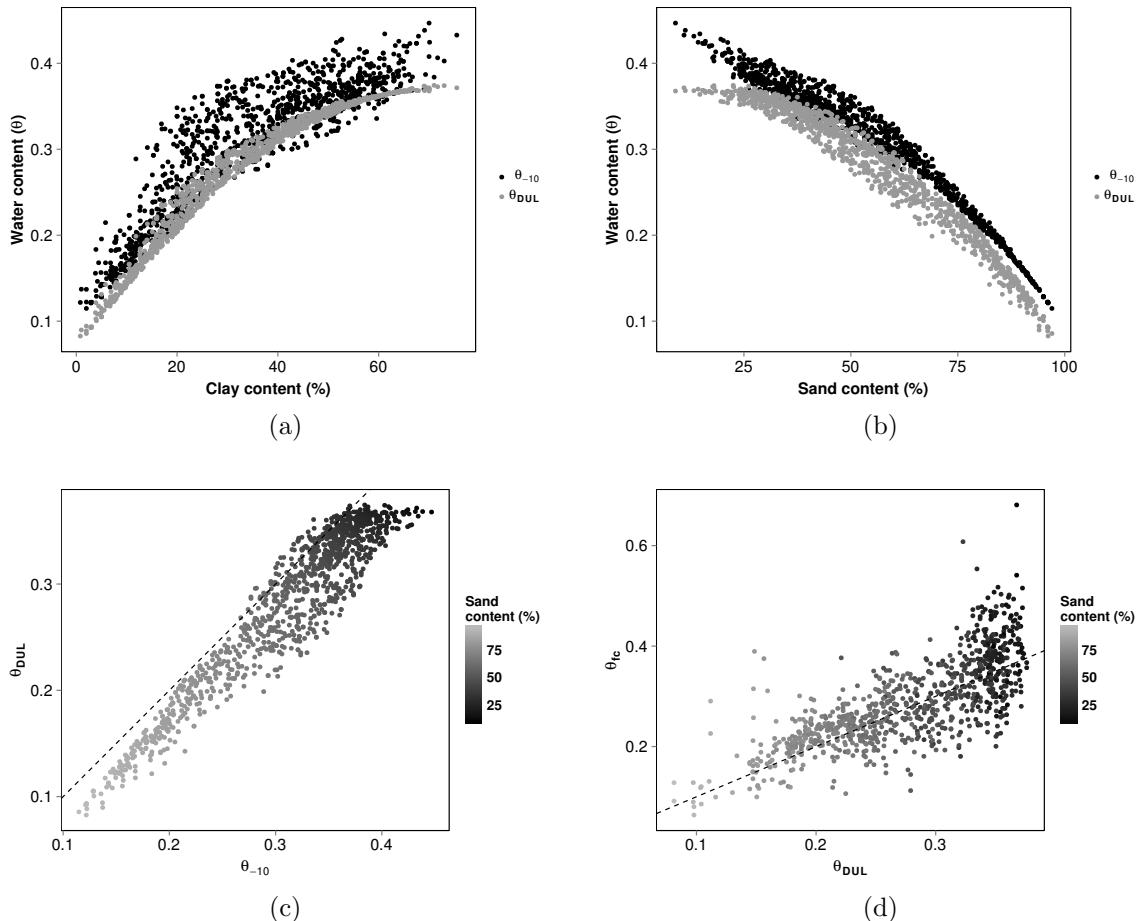


Fig. 2.2: Prediction comparison between: **a-c** $\hat{\theta}_{DUL}$ and $\hat{\theta}_{-10}$ using field measurements dataset (Table 2.3). Predictions were made using Eq. 2.9 and Eq. 2.11; **d** $\hat{\theta}_{DUL}$ and predictions made with PTF proposed by Twarakavi *et al.*, (2009).

2.4.4 Crop lower limit

Following the same procedure than with the previous properties, we generated PTFs to predict CLL ($\hat{\theta}_{\text{CLL}}$). The resulting PTF is:

$$\begin{aligned}\hat{\theta}_{\text{CLL}} \left(\text{cm}^3 / 100\text{cm}^3 \right) = & 0.1476 + 9.002 \times 10^{-5} \text{ clay}^2 \\ & - 0.00115 \text{ sand} \\ & - 9.752 \times 10^{-7} \text{ clay}^3\end{aligned}\quad (2.13)$$

with R^2 and RMSE values of 0.65 and 4.37 (%) respectively. We also derived a relation with DUL due that it is easier to measure than CLL:

$$\hat{\theta}_{\text{CLL}} \left(\text{cm}^3 / 100\text{cm}^3 \right) = 0.6151 \theta_{\text{DUL}} - 0.02192 \quad (2.14)$$

were θ_{DUL} corresponds to the measured value of DUL, with a R^2 value of 0.61 and RMSE of 4.65.

2.4.5 Water content at -1500 kPa

For laboratory measurement at -1500 kPa we generated the following PTF:

$$\begin{aligned}\hat{\theta}_{-1500} \left(\text{cm}^3 / 100\text{cm}^3 \right) = & 0.1766 + 0.00255 \text{ clay} \\ & - 0.001487 \text{ sand}\end{aligned}\quad (2.15)$$

with R^2 and RMSE values of 0.71 and 4.84 (%) respectively. As in the case of field measurements, we also generated a PTF for laboratory measurements, and obtained:

$$\hat{\theta}_{-1500} \left(\text{cm}^3 / 100\text{cm}^3 \right) = 0.814 \theta_{-10} - 0.07996 \quad (2.16)$$

with a R^2 value of 0.65 and RMSE of 5.43 (%).

2.4.6 Relationship between CLL and -1500

As in the previous comparison, laboratory measurements generated an over-prediction of soil water content estimated in-situ (Fig. 2.3). This indicates that this measures also depends on the plant and some of the plants here can survive at potentials dryer than -1500 kPa.

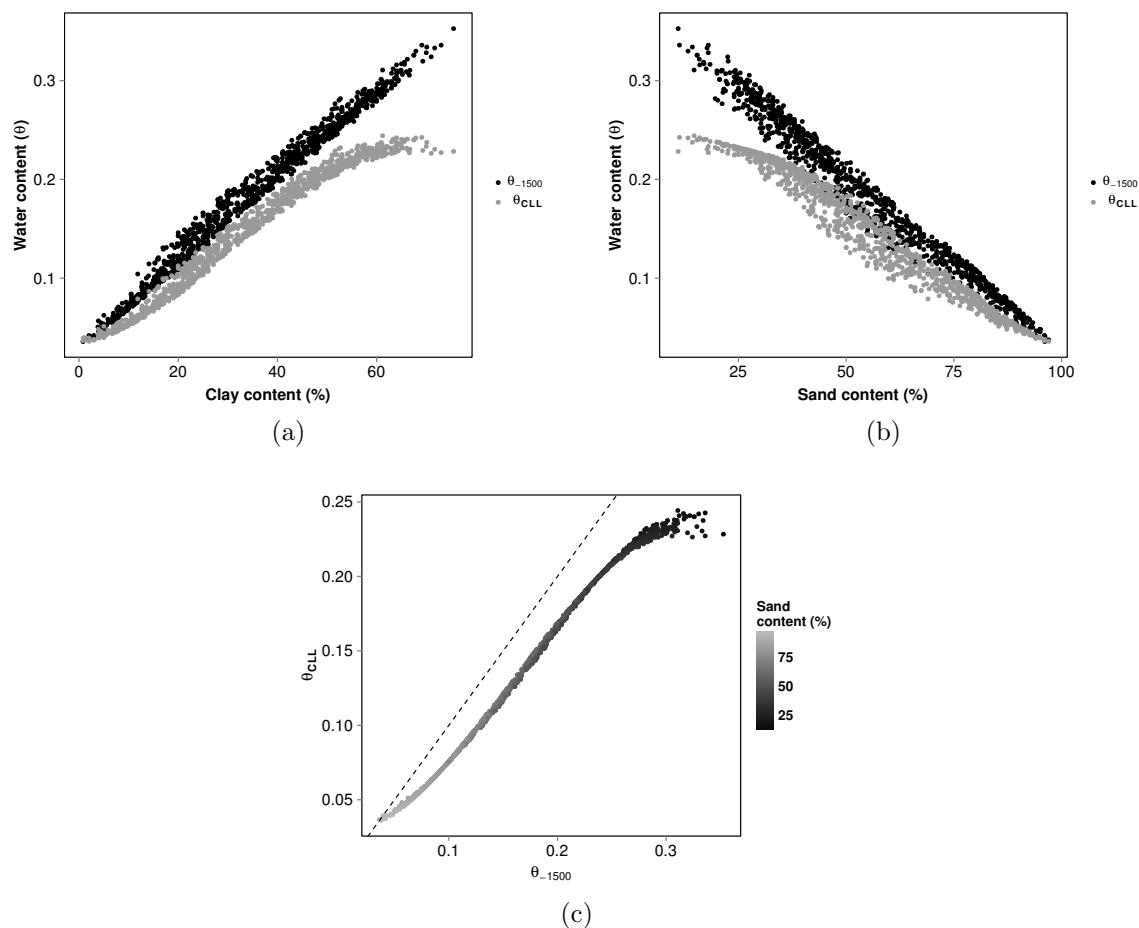


Fig. 2.3: Prediction comparison between $\hat{\theta}_{\text{CLL}}$ and $\hat{\theta}_{-1500}$ using field measurements dataset (Table 2.3). Predictions were made using Eq. 2.13 and Eq. 2.15.

2.4.7 Uncertainty estimation

Following the workflow specified in the section 2.3.3, to assess the uncertainty, we need to obtain the optimum number of clusters and prediction interval (PI) for every PTF. This is achieved by calculating PICP and MPI for different number of clusters. As an example, Fig. 2.4 shows values of PICP and MPI for as function of the number of cluster for Eq. 2.9. As we attempt to predict 95% prediction interval, we select the number of clusters that are the closest to the 95% confidence interval and the minimum MPI value correspond. In this case 13 clusters seems appropriate.

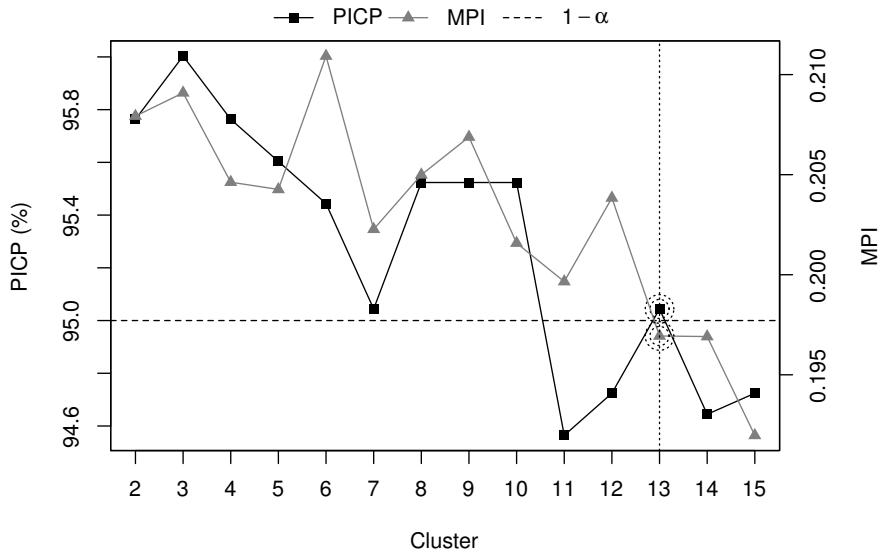


Fig. 2.4: PICP and MPI behaviour with different number of cluster for Eq. 2.9. Dotted circles highlight the optimum number of clusters.

2.4.8 External validation

Using the data described in Table 2.5 we performed an independent or external validation of our PTFs. Table 2.6 shows the performance of different PTFs for DUL and CLL. As in the prediction, we observed a decrease in RMSE when more relevant inputs are used. For example, in DUL, Eq. 2.5 that used sand, clay, BD and CEC as predictors has an $R^2 = 0.84$, while Eq. 2.9 that only used sand and clay has a poorer

result ($R^2 = 0.45$). However the uncertainty is a bit narrow, as the PICP for Eq. 2.5 is only 79% (expected to be 95%) and the PICP values also become smaller with simpler models. It is worth noting that PICP depends on the number of observations (the larger the number of observations, the closer the PICP to the nominal 95% confidence interval) and the standard deviation of the prediction (Hwang and Ding, 1997).

Table 2.6: External validation statistics of prediction quality

	R^2	RMSE	PICP	MPI
	%			
$\hat{\theta}_{\text{DUL}}$ (Eq. 2.5)	0.84	7.89	78.97	23.84
$\hat{\theta}_{\text{DUL}}$ (Eq. 2.6)	0.79	8.56	66.55	22.96
$\hat{\theta}_{\text{DUL}}$ (Eq. 2.7)	0.77	8.36	55.18	18.28
$\hat{\theta}_{\text{DUL}}$ (Eq. 2.9)	0.45	8.75	63.55	20.56
$\hat{\theta}_{\text{CLL}}$ (Eq. 2.13)	0.44	5.83	91.97	18.53
$\hat{\theta}_{\text{CLL}}$ (Eq. 2.14)	0.78	5.91	83.55	16.18
$\hat{\theta}_{-1500}$ (Eq. 2.15)	0.64	6.49	91.56	21.50
$\hat{\theta}_{-1500}$ (Eq. 2.16)	0.84	5.88	77.70	20.08

2.5 Making predictions with new data

In order to utilise the PTFs, first is necessary to evaluate how much information is available to perform predictions (step (1) in Fig. B.1). In an ideal case, many soil properties would be available and it would be possible to utilise the PTF with lower error. In this example we explore the use of a PTF when data is limited and just soil sand and clay fraction are available to predict DUL (thus selecting Eq. 2.9).

Once a PTF has been selected, is necessary to calculate the membership of an observation to each of the clusters using Eq. 2.1 and Eq. 2.2. This calculation also determine whether the inputs belong to the domain of the data used to generate the PTFs (step (2) in Fig. B.1). After calculating the memberships, the prediction with the PTF could be performed. Similarly, the values of the lower PI (PI^L) and the upper PI (PI^U) of the residuals of each cluster are weighted by the membership values to obtain the corresponding prediction limits (PL, step (3) in Fig. B.1). As an

illustration, Fig. 2.5 shows the plot of sand and clay content of the input variables for three observation. The domain of the input variables used to generate the PTF (Eq. 2.9) is illustrated by the convex hull, and the mean or centroid of 2 clusters are also depicted.

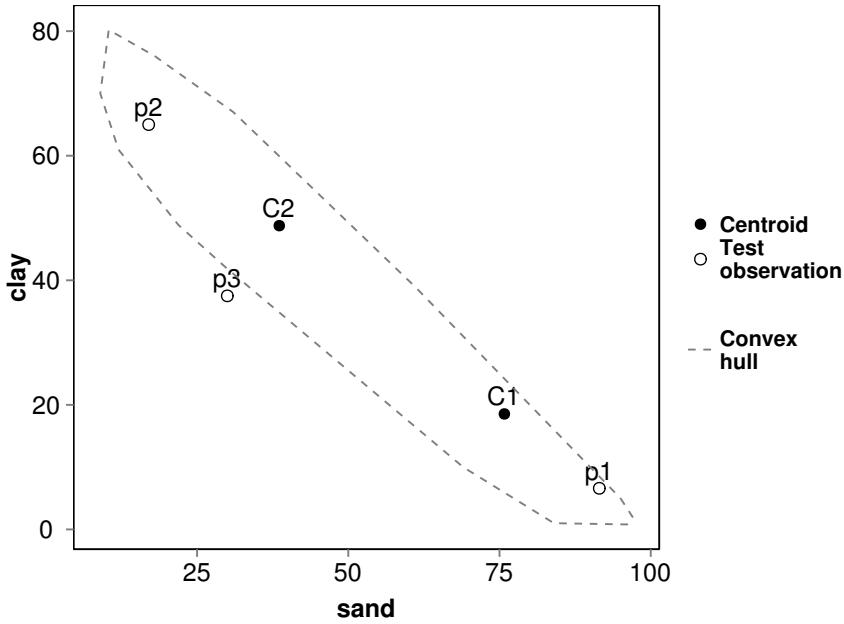


Fig. 2.5: Relative positioning of observations in relation of class centroids. Convex hull represents the limit to consider an observation as an extragrade.

Table 2.7 shows the calculation in this example, where the memberships of the observations in the two classes and the extragrade which was calculated using Eq. 2.1 and Eq. 2.2 respectively. Next, the prediction was calculated using Eq. 2.9 (in a example with only 2 clusters) and the values of PI^L and PI^U were weighted and added to the prediction to obtain the corresponding PL. The results are shown in Table 2.7. In a real case, the values of PI^L and PI^U can be obtained from Appendix C.

The first observation (p1) is closer to the centroid of cluster 1 (point C1 in Fig. 2.5) and it was effectively assigned to that class, as confirmed with the m_{C1} membership value of 0.99 (Table 2.7). Likewise, the second observation (p2) was assigned to the second class, represented by the centroid C2. The third point (p3) lies outside of the data domain (represented by the convex hull in Fig. 2.5) therefore its membership with the extragrade class is higher and its PI wider (0.36 compared with 0.22 and 0.18 for

Table 2.7: Membership (m) in clusters C1, C2 and extragrade (*), prediction intervals (PI), prediction limits (PL) and DUL prediction for example observations, using Eq. 2.9.

Obs.	sand	clay	m_{C1}	m_{C2}	m_*	PI^L	PI^U	PL^L	DUL	PL^U
p1	91.50	6.60	0.99	0.01	0.00	-0.09	0.13	0.03	0.12	0.25
p2	17.00	65.00	0.02	0.98	0.00	-0.09	0.09	0.28	0.37	0.46
p3	30.00	37.50	0.06	0.12	0.82	-0.19	0.18	0.14	0.33	0.51

p1 and p2 respectively).

2.6 Conclusions

We presented the need of soil data for Australia, reviewing biophysical models currently used. We identified eight key soil properties consistently used in these models, including bulk density, drainage upper limit, crop lower limit, water content at -10 and -1500 kPa, organic carbon, saturated hydraulic conductivity, and USLE soil erodibility factor.

We used a genetic programming technique to generate pedotransfer functions (PTFs) specifically designed to be used in the Australian context. We also used the fuzzy k-means algorithm to estimate their prediction intervals and to identify observations outside of the calibration data domain. The latest gives the possibility to use the PTFs in other locations with soils with properties within the range of Australian soils properties.

We also proposed to present PTFs along with uncertainty levels and information about the data used in the training process. Published PTFs usually lack this information and we believe it is crucial to provide it, independent of the method used to obtain it, to avoid PTFs misuse and extrapolation of the model to another data domain where prediction validity is not guaranteed.

2.7 References

- Arnold, J., Williams, J., Srinivasan, R., King, K., and Griggs, R. 1994. SWAT: Soil water assessment tool. *US Department of Agriculture, Agricultural Research Service, Grassland, Soil and Water Research Laboratory, Temple, TX.*
- Bezdek, J., Ehrlich, R., and Full, W. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10 (2): 191–203.
- Bouma, J. 1989. Using soil survey data for quantitative land evaluation. *Advances in soil sciences*. 9: 177–213.
- Buchmann, N. 2000. Biotic and abiotic factors controlling soil respiration rates in *Picea abies* stands. *Soil Biology and Biochemistry* 32 (11–12): 1625–1635.
- Comins, H. and McMurtrie, R. 1993. Long-Term Response of Nutrient-Limited Forests to CO₂ Enrichment; Equilibrium Behavior of Plant-Soil Models. *Ecological Applications* 3 (4): 666–681.
- Dalgliesh, N., Cocks, B., and Horan, H. 2012. APSoil-providing soils information to consultants, farmers and researchers. In: *16th Australian Agronomy Conference, Armidale, NSW.*
- Dalgliesh, N., Foale, M., and McCown, R. 2009. Re-inventing model-based decision support with Australian dryland farmers. 2. Pragmatic provision of soil information for paddock-specific simulation and farmer decision making. *Crop and Pasture Science* 60 (11): 1031–1043.
- Davis, J. and Farley, T. 1997. CMSS: policy analysis software for catchment managers. *Environmental Modelling & Software* 12 (2-3): 197–210.
- De Grujter, J. and McBratney, A. 1988. A modified fuzzy k-means method for predictive classification. In: *1. Conference of the International Federation of Classification Societies*. Ed. by H. Bock. Elsevier Science, Amsterdam: pp. 97–104.
- Dunne, K. A. and Willmott, C. J. 1996. Global distribution of plant-extractable water capacity of soil. *International Journal of Climatology* 16 (8): 841–859.
- Entekhabi, D., Rodriguez-Iturbe, I., and Castelli, F. 1996. Mutual interaction of soil moisture state and atmospheric processes. *Journal of Hydrology* 184 (1): 3–17.
- Evans, W., Gilfedder, M., and Austin, J. 2004. *Application of the Biophysical Capacity to Change (BC2C) model to the Little River (NSW)*. Canberra: CSIRO Land and Water Technical Report 15/04, CSIRO.

- Frank, S. A. and Slatkin, M. 1990. Evolution in a variable environment. *American Naturalist*: 244–260.
- Gardner, E., Shaw, R., Smith, G., and Coughlan, K. 1984. Plant available water capacity: concept, measurement and prediction. In: *The Properties and Utilization of Cracking Clay Soils*. Ed. by J. McGarity, E. Hoult, and H. So. University of New England, Armidale: pp. 164–175.
- Garrels, R. and Mackenzie, F. 1967. Origin of the Chemical Compositions of Some Springs and Lakes. In: *Equilibrium Concepts in Natural Water Systems*. Chap. 10: pp. 222–242.
- Gordon, H. 2002. *The CSIRO Mk3 climate system model*. CSIRO Atmospheric Research.
- Han, G., Zhou, G., Xu, Z., Yang, Y., Liu, J., and Shi, K. 2007. Biotic and abiotic factors controlling the spatial and temporal variation of soil respiration in an agricultural ecosystem. *Soil Biology and Biochemistry* 39 (2): 418–425.
- Haxeltine, A. and Prentice, I. 1996. BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types. *Global Biogeochemical Cycles* 10 (4): 693–709.
- Hearn, A. 1994. OZCOT: A simulation model for cotton crop management. *Agricultural Systems* 44 (3): 257–299.
- Hochman, Z., Dalgliesh, N., and Bell, K. 2001. Contributions of soil and crop factors to plant available soil water capacity of annual crops on Black and Grey Vertosols. *Crop and Pasture Science* 52 (10): 955–961.
- Hwang, J. G. and Ding, A. A. 1997. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association* 92 (438): 748–757.
- Ines, A. V., Honda, K., Gupta, A. D., Droogers, P., and Clemente, R. S. 2006. Combining remote sensing-simulation modeling and genetic algorithm optimization to explore water management options in irrigated agriculture. *Agricultural Water Management* 83 (3): 221–232.
- Jakeman, A., Littlewood, I., and Whitehead, P. 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of hydrology* 117 (1): 275–300.
- Jenkinson, D., Hart, P., Rayner, J., and Parry, L. 1987. Modelling the turnover of organic matter in long-term experiments at Rothamsted. *Intecol. Bull.* (15): 1–8.

- Johari, A., Habibagahi, G., and Ghahramani, A. 2006. Prediction of Soil-Water Characteristic Curve Using Genetic Programming. *Journal of Geotechnical and Geoenvironmental Engineering* 132 (5): 661–665.
- Jones, J., Hoogenboom, G., Porter, C., Boote, K., Batchelor, W., Hunt, L., Wilkens, P., Singh, U., Gijsman, A., and Ritchie, J. 2003. The DSSAT cropping system model. *European Journal of Agronomy* 18 (3–4): 235–265.
- Kaplan, J., Bigelow, N., Prentice, I., Harrison, S., Bartlein, P., Christensen, T., Cramer, W., Matveyeva, N., McGuire, A., and Murray, D. 2003. Climate change and Arctic ecosystems: 2. Modeling, paleodata-model comparisons, and future projections. *J. Geophys. Res* 108 (D19): 8171.
- Koza, J. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. *The MIT Press*.
- Koza, J. 1994. Genetic Programming II: Automatic Discovery of Reusable Subprograms. *The MIT Press*.
- Koza, J., Bennett, H., Andre, D., and Keane, M. 1999. Genetic Programming III: Darwinian Invention and Problem Solving. *Morgan Kaufmann Publishers*.
- Kuo, S. and Baker, A. 1980. Sorption of copper, zinc, and cadmium by some acid soils. *Soil Science Society of America Journal* 44 (5): 969–974.
- Landsberg, J. and Waring, R. 1997. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management* 95 (3): 209–228.
- Li, C., Frolking, S., and Frolking, T. 1992. A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. *J. Geophys. Res* 97 (D9): 9759–9776.
- Li, C., Aber, J., Stange, F., Butterbach-Bahl, K., and Papen, H. 2000. A process-oriented model of N_2O and NO emissions from forest soils: 1. Model development. *J. Geophys. Res* 105 (D4): 4369–4384.
- Littleboy, M., Silburn, D., Freebairn, D., Woodruff, D., and Hammer, G. 1989. PERFECT - A computer simulation model of Productivity Erosion Runoff Functions to Evaluate Conservation Techniques. *Queensland Department of Primary Industries Bulletin* (QB89005).

- Makkeasorn, A., Chang, N., Beaman, M., Wyatt, C., and Slater, C. 2006. Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. *Water Resources Research* 42 (9) W09401: W09401.
- McBratney, A. B., Minasny, B., Cattle, S. R., and Vervoort, R. 2002. From pedotransfer functions to soil inference systems. *Geoderma* 109 (1–2): 41–73.
- McCown, R., Hammer, G., Hargreaves, J., Holzworth, D., and Freebairn, D. 1996. APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. *Agricultural systems* 50 (3): 255–271.
- Minasny, B. and McBratney, A. 2000. Evaluation and development of hydraulic conductivity pedotransfer functions for Australian soil. *Aust. J. Soil Res.* 38: 905–926.
- Minasny, B., McBratney, A., and Bristow, K. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma* 93 (3–4): 225–253.
- Moorhead, D. and Reynolds, J. 1991. A general model of litter decomposition in the northern Chihuahuan Desert. *Ecological Modelling* 56: 197–219.
- Nachabe, M. 1998. Refining the definition of field capacity in the literature. *Journal of irrigation and drainage engineering* 124 (4): 230–232.
- Nemes, A., Pachepsky, Y., and Timlin, D. 2011. Toward improving global estimates of field soil water capacity. *Soil Science Society of America Journal* 75 (3): 807–812.
- Pampura, T., Groenenberg, J., Loftis, S., and Priputina, I. 2007. Validation of transfer functions predicting Cd and Pb free metal ion activity in soil solution as a function of soil characteristics and reactive metal content. *Water, Air, & Soil Pollution* 184 (1): 217–234.
- Parasuraman, K., Elshorbagy, A., and Si, B. C. 2007a. Estimating Saturated Hydraulic Conductivity Using Genetic Programming. *Soil Sci. Soc. Am. J.* 71 (6): 1676–1684.
- Parasuraman, K., Elshorbagy, A., and Carey, S. K. 2007b. Modelling the dynamics of the evapotranspiration process using genetic programming. *Hydrological Sciences Journal* 52 (3): 563–578.
- Parton, W., Schimel, D., Cole, C., and Ojima, D. 1987. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Sci. Soc. Am. J.* 51 (5): 1173–1179.

- Rab, M., Chandra, S., Fisher, P., Robinson, N., Kitching, M., Aumann, C., and Imhof, M. 2011. Modelling and prediction of soil water contents at field capacity and permanent wilting point of dryland cropping soils. *Soil Research* 49 (5): 389–407.
- Ratliff, L., Ritchie, J., and Cassel, D. 1983. Field-measured limits of soil water availability as related to laboratory-measured properties. *Soil Science Society of America Journal* 47 (4): 770–775.
- Rawls, W., Brakensiek, D., and Saxton, K. 1982. Estimation of soil water properties. *Trans. Asae* 25 (5): 1316–1320.
- Rawls, W., Gish, T., and Brakensiek, D. 1991. Estimating soil water retention from soil physical properties and characteristics. *Adv. Soil Sci* 16: 213–234.
- Richards, G. 2001. The FullCAM carbon accounting model: development, calibration and implementation for the National Carbon Accounting System. *National Carbon Accounting System Technical Report* 28.
- Richards, G. and Evans, D. 2000a. CAMAg National Carbon Accounting System. *Australian Greenhouse Office, Canberra*.
- Richards, G. and Evans, D. 2000b. CAMFor User Manual v 3.35. *National Carbon Accounting System Technical Report* (26).
- Rodriguez-Iturbe, I., Entekhabi, D., and Bras, R. 1991. Nonlinear Dynamics of Soil Moisture at Climate Scales: 1. Stochastic Analysis. *Water Resour. Res.* 27 (8): 1899–1906.
- Romano, N., Palladino, M., and Chirico, G. 2011. Parameterization of a bucket model for soil-vegetation-atmosphere modeling under seasonal climatic regimes. *Hydrology and Earth System Sciences* 15 (12): 3877–3893.
- Schimel, D., Braswell, B., Holland, E., McKeown, R., Ojima, D., Painter, T., Parton, W., and Townsend, A. 1994. Climatic, edaphic, and biotic controls over storage and turnover of carbon in soils. *Global Biogeochemical Cycles* 8 (3): 279–293.
- Selle, B., Wang, Q., and Mehta, B. 2011. Relationship between hydraulic and basic properties for irrigated soils in southeast Australia. *Journal of Plant Nutrition and Soil Science* 174 (1): 81–92.
- Sharma, D. K. and Jana, R. 2009. Fuzzy goal programming based genetic algorithm approach to nutrient management for rice crop planning. *International Journal of Production Economics* 121 (1): 224–232.

- Shrestha, D. and Solomatine, D. 2006. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks* 19 (2): 225–235.
- Sitch, S., Smith, B., Prentice, I., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J., Levis, S., Lucht, W., and Sykes, M. 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology* 9 (2): 161–185.
- Torri, D., Poesen, J., and Borselli, L. 1997. Predictability and uncertainty of the soil erodibility factor using a global dataset. *Catena* 31 (1-2): 1–22.
- Tranter, G., Minasny, B., McBratney, A., Murphy, B., McKenzie, N., Grundy, M., and Brough, D. 2007. Building and testing conceptual and empirical models for predicting soil bulk density. *Soil Use and Management* (23): 437–443.
- Tranter, G., Minasny, B., and McBratney, A. 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades. *Soil Sci. Soc. Am. J.* 74 (6): 1967–1975.
- Tuteja, N., Vaze, J., Murphy, B., and Beale, G. 2004. *CLASS—Catchment scale multiple landuse atmosphere soil water and solute transport model*. Department of Infrastructure, Planning, Natural Resources, and Cooperative Research Centre for Catchment Hydrology, Technical Report 04/12. New South Wales, Australia.
- Twarakavi, N. K., Sakai, M., and Šimůnek, J. 2009. An objective analysis of the dynamic nature of field capacity. *Water Resources Research* 45 (10): W10410.
- Veihmeyer, F. and Hendrickson, A. 1949. Methods of measuring field capacity and permanent wilting percentage of soils. *Soil science* 68 (1): 75–94.
- Viscarra Rossel, R., Minasny, B., Roudier, P., and McBratney, A. 2006. Colour space models for soil science. *Geoderma* 133 (3): 320–337.
- Walker, J. and Houser, P. 2001. A methodology for initializing soil moisture in a global climate model: Assimilation of near-surface soil moisture observations. *Journal of Geophysical Research* 106 (11): 11.
- White, R. E. 2009. *Principles and practice of soil science: the soil as a natural resource*. John Wiley & Sons.
- Wilkinson, S., Henderson, A., Chen, Y., and Sherman, B. 2004. SedNet user guide. *Client Report, CSIRO Land and Water*.
- Williams, J. 1995. The EPIC model. In: *Computer models of watershed hydrology*. Ed. by V. Singh. Water Resources Publications: pp. 909–1000.

- Wösten, J., Pachepsky, Y., and Rawls, W. 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology* 251 (3–4): 123–150.
- Zinn, Y., Lal, R., and Resck, D. 2005. Texture and organic carbon relations described by a profile pedotransfer function for Brazilian Cerrado soils. *Geoderma* 127 (1): 168–173.

Appendix A - Models investigated

PERFECT: Productivity Erosion Runoff Functions to Evaluate Conservation Techniques (Littleboy *et al.*, 1989). It predicts the effect of climate, soil type, crop sequence and fallow management on the water balance, erosion, and productivity. Developed for sub-tropical grain growing areas of Queensland.

SWAT: Soil and Water Assessment Tool (Arnold *et al.*, 1994). It predicts the impact of land management practices on water, sediment and agricultural chemical yields in large complex watersheds.

SedNet: Sediment River Network (Wilkinson *et al.*, 2004). It constructs sediment and nutrient (phosphorus and nitrogen) budgets for regional scale river networks to identify patterns in the material fluxes.

FullCAM: It is an activity-driven carbon accounting model capable of dealing with multiple carbon pools for the National Carbon Accounting System established by the Australian Government (Richards, 2001). It is a compendium of models like: the physiological growth model for forests, 3PG (Landsberg and Waring, 1997); the carbon accounting model for forests (CAMFor) developed by the Australian Greenhouse Office (Richards and Evans, 2000b); the carbon accounting model for cropping and grazing systems (CAMA_g) (Richards and Evans, 2000a); the microbial decomposition model GENDEC (Moorhead and Reynolds, 1991); and the Rothamsted Soil Carbon Model (RothC) (Jenkinson *et al.*, 1987).

APSIM: Agricultural Production Systems Simulator (McCown *et al.*, 1996). It simulates biophysical processes in farming systems, in particular where there is interest in the production, economic and ecological outcomes of management practice.

Mk3.5: Climate System Model (Gordon, 2002). A model developed by CSIRO, which contains a comprehensive representation of the four major components of the climate system (atmosphere, land surface, oceans and sea-ice). It is used to investigate the dynamical and physical processes controlling the climate system, for multiseasonal predictions, and for investigations of natural climatic variability and climatic change.

CENTURY: Is an agroecosystem model which simulates long-term changes in soil organic carbon and nitrogen, nutrient cycling, and plant production for soil–plant ecosystems. It was originally developed for its use in the U.S. Great Plains grasslands (Parton *et al.*, 1987) and has been ported to various ecosystems around the world.

DNDC: Denitrification-Decomposition model (Li *et al.*, 2000; Li *et al.*, 1992). A general model of carbon and nitrogen biogeochemistry in agricultural ecosystems which assesses trace gas emissions of scenarios like changes of land use, agricultural activities, mitigation options, etc.

CLASS: Catchment scale multiple Landuse Atmosphere Soil water and Solute transport model (Tuteja *et al.*, 2004). It consist in a group of tools for physically based eco-hydrological modelling. It was designed for investigation of the effects of landuse and climate variability on both paddock scale as well as the catchment scale. It includes a Crop Growth Model (CGM), Pasture Growth Model (PGM), Spatial Analysis (SA) and an Unsaturated Moisture Movement Model (U3M-1D).

DSSAT: Decision Support System for Agro Technology Transfer (Jones *et al.*, 2003). Developed to facilitate the application of crop models in a systems approach to agronomic research, to integrate knowledge about soil, climate, crops, and management for making better decisions about transferring production technology from one location to others where soils and climate differed.

OZCOT: Originally developed by Hearn (1994) to assess the performance of cotton crops under different environmental and management conditions.

GDAY: Generic Decomposition And Yield (Comins and McMurtrie, 1993). It is a plant-soil model describing fluxes of carbon and nitrogen plant, litter and soil compartments. It uses a simplified physiology-based canopy assimilation model to calculate carbon uptake, and the CENTURY model for soil carbon decomposition and nitrogen cycling.

BIOME4: It is a coupled carbon and water flux model that predicts global steady state vegetation distribution, structure, and biogeochemistry, taking account of interactions among these aspects (Kaplan *et al.*, 2003). Its simulation are based

on BIOME3 (Haxeltine and Prentice, 1996) and LPJ (Sitch *et al.*, 2003) models. It uses internal PTF to predict water retention parameters.

LPX: Land surface Processes and exchanges. It combines process-based, large-scale representations of terrestrial vegetation dynamics and land-atmosphere carbon and water exchanges. It is based on LPJ model (Sitch *et al.*, 2003), adding simulation of wildfires started by lightning ignition. It uses internal PTF to predict water retention parameters.

BC2C: Biophysical Capacity to Change (Evans *et al.*, 2004). It is a conceptual mass balance model designed to simulate the long-term average salt and water yield of whole catchments. Includes internal relations between water retention parameters and evapotranspiration.

CMSS: Catchment Management Support System (Davis and Farley, 1997). Designed to provide long term, broad area prediction of the impacts of different nutrient management strategies on water quality in Australian catchments.

IHACRES: Identification of unit Hydrographs And Component flows from Rainfall, Evaporation and Streamflow data (Jakeman *et al.*, 1990). It is a catchment-scale rainfall-streamflow modelling methodology whose purpose is to characterise the dynamic relationship between rainfall and streamflow. It uses water retention parameters to internally derive evapotranspiration.

Appendix B - PTF use diagram

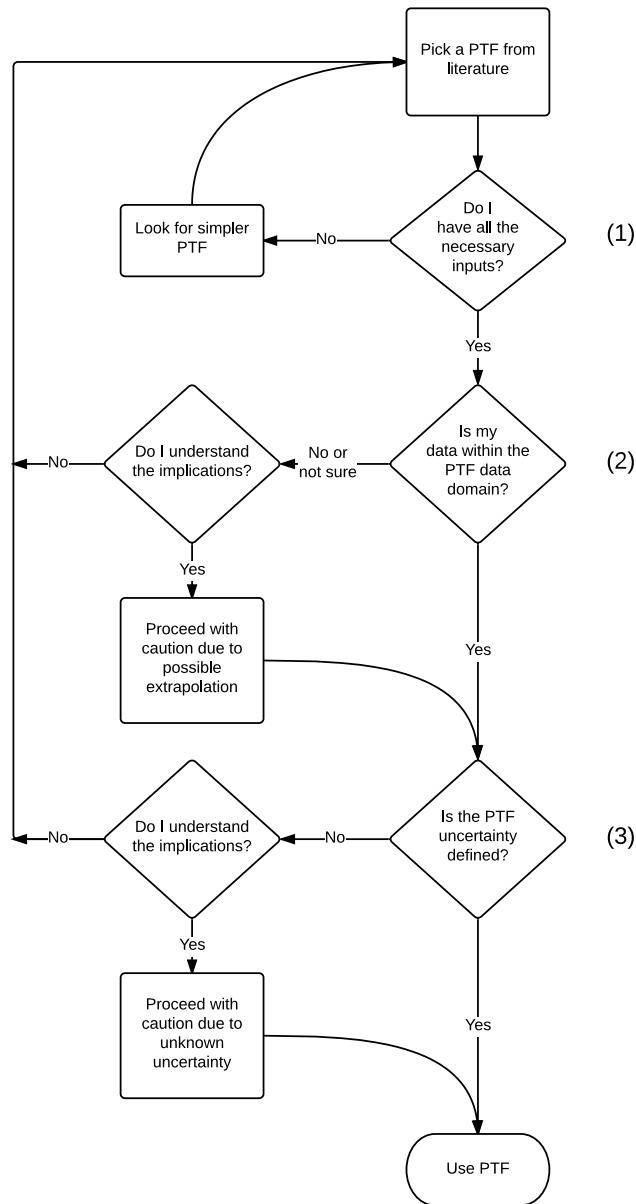


Fig. B.1: PTF use diagram

Appendix C - PTF cluster information

Table C.1: Calibration cluster information for use with Eq. 2.5.

(a) Class centroids and prediction intervals (PI)

	clay	sand	cec	bd	PI^L	Mean	PI^U
Cluster	Centroids				Cluster residuals		
1	14.92	81.19	5.94	1.69	-0.08	0.00	0.11
2	34.07	45.75	15.15	1.55	-0.07	0.00	0.08
3	43.38	47.01	28.94	1.40	-0.10	0.00	0.09
4	51.72	36.83	19.95	1.49	-0.12	0.00	0.07
Ex	—	—	—	—	-0.18	-0.01	0.22

(b) Variance-covariance matrix

	clay	sand	cec	bd
clay	0.03	0.02	-0.01	-0.26
sand	0.02	0.02	-0.00	-0.16
cec	-0.01	-0.00	0.03	0.75
bd	-0.26	-0.16	0.75	64.87

Table C.2: Calibration cluster information for use with Eq. 2.6.

(a) Class centroids and prediction intervals (PI)

	cec	clay	sand	PI^L	Mean	PI^U
Cluster	Centroids			Cluster residuals		
1	22.58	25.48	69.46	-0.10	0.00	0.13
2	22.71	57.16	32.94	-0.12	0.00	0.06
3	9.69	23.63	62.89	-0.08	0.00	0.11
4	3.79	11.18	86.07	-0.07	0.00	0.12
5	11.05	33.02	61.20	-0.16	0.01	0.09
6	12.50	30.74	46.03	-0.08	0.00	0.08
7	18.12	51.66	32.47	-0.05	0.00	0.07
8	34.04	53.26	37.10	-0.09	-0.01	0.06
9	24.66	45.10	40.60	-0.08	0.00	0.08
Ex	—	—	—	-0.19	0.00	0.19

(b) Variance-covariance matrix

	cec	clay	sand
cec	0.02	-0.01	-0.00
clay	-0.01	0.03	0.02
sand	-0.00	0.02	0.02

Table C.3: Calibration cluster information for use with Eq. 2.7.

(a) Class centroids and prediction intervals (PI)

	bd	clay	sand	PI^L	Mean	PI^U
Cluster	Centroids			Cluster residuals		
1	1.54	25.98	52.30	-0.10	0.00	0.08
2	1.44	57.45	32.76	-0.04	0.01	0.03
3	1.59	10.42	86.96	-0.09	0.00	0.11
4	1.49	51.73	33.98	-0.10	0.00	0.12
5	1.32	54.01	35.06	-0.08	0.02	0.03
6	1.33	36.71	50.30	-0.07	0.02	0.16
7	1.64	45.85	44.11	-0.09	0.01	0.03
8	1.77	13.71	83.40	-0.16	-0.01	0.04
9	1.51	35.82	59.62	-0.11	-0.01	0.08
10	1.60	42.85	36.75	-0.09	0.00	0.04
11	1.53	24.52	63.06	-0.06	0.00	0.11
12	1.31	21.49	74.54	-0.12	0.00	0.10
Ex	—	—	—	-0.23	0.01	0.24

(b) Variance-covariance matrix

	bd	clay	sand
bd	42.35	0.09	-0.04
clay	0.09	0.03	0.02
sand	-0.04	0.02	0.02

Table C.4: Calibration cluster information for use with Eq. 2.8.

(a) Class centroids and prediction intervals (PI)

Cluster	oc	clay	sand	PI^L	Mean	PI^U
	Centroids			Cluster residuals		
1	0.50	15.77	80.63	-0.09	0.00	0.12
2	0.19	29.26	47.85	-0.07	0.00	0.09
3	0.15	49.47	42.24	-0.05	-0.01	0.11
4	0.20	47.00	32.39	-0.11	0.01	0.06
5	1.05	16.62	80.34	-0.09	0.00	0.12
6	0.15	10.82	86.80	-0.11	-0.02	0.08
7	0.19	56.60	29.11	-0.08	0.01	0.08
8	0.27	22.56	65.11	-0.11	-0.01	0.09
9	0.28	60.58	30.10	-0.06	0.00	0.06
10	0.75	48.45	39.33	-0.08	-0.01	0.13
11	0.99	25.60	55.68	-0.04	-0.01	0.17
12	1.38	35.92	51.89	-0.06	0.01	0.05
13	0.31	31.91	63.90	-0.08	-0.02	0.14
14	0.23	41.26	45.84	-0.10	0.00	0.06
Ex	—	—	—	-0.22	0.00	0.23

(b) Variance-covariance matrix

	oc	clay	sand
oc	5.05	0.05	0.03
clay	0.05	0.03	0.02
sand	0.03	0.02	0.02

Table C.5: Calibration cluster information for use with Eq. 2.9.

(a) Class centroids and prediction intervals (PI)

	sand	clay	PI^L	Mean	PI^U
Cluster	Centroids		Cluster residuals		
1	58.84	38.17	-0.16	0.00	0.10
2	76.52	21.31	-0.12	0.00	0.07
3	59.88	32.42	-0.05	-0.01	0.09
4	33.96	44.99	-0.09	0.02	0.07
5	52.08	26.85	-0.06	0.00	0.13
6	61.29	24.74	-0.07	-0.01	0.11
7	35.06	52.25	-0.11	0.00	0.13
8	44.92	41.26	-0.07	0.00	0.06
9	77.20	15.43	-0.10	0.00	0.09
10	28.74	61.74	-0.08	0.00	0.12
11	22.70	60.09	-0.08	-0.01	0.05
12	40.94	50.70	-0.10	0.00	0.07
13	88.75	9.29	-0.09	0.01	0.15
Ex	—	—	-0.22	0.00	0.22

(b) Variance-covariance matrix

	sand	clay
sand	0.02	0.02
clay	0.02	0.03

Table C.6: Calibration cluster information for use with Eq. 2.10.

(a) Class centroids and prediction intervals (PI)					
	sand	bd	PI^L	Mean	PI^U
Cluster	Centroids		Cluster residuals		
1	70.31	1.43	-0.10	-0.01	0.12
2	32.19	1.69	-0.12	0.03	0.09
3	41.00	1.26	-0.10	0.01	0.10
4	71.88	1.61	-0.10	-0.01	0.11
5	87.94	1.58	-0.10	0.02	0.08
6	21.66	1.13	-0.14	0.00	0.08
7	41.79	1.55	-0.06	0.01	0.09
8	78.79	1.30	-0.08	0.01	0.12
9	26.66	1.33	-0.11	-0.02	0.10
10	55.55	1.49	-0.11	0.01	0.06
11	19.33	1.49	-0.11	0.00	0.13
12	61.54	1.70	-0.08	-0.01	0.12
Ex	—	—	-0.25	—	0.26

(b) Variance-covariance matrix		
	sand	bd
sand	0.00	-0.09
bd	-0.09	23.77

Table C.7: Calibration cluster information for use with Eq. 2.11.

(a) Class centroids and prediction intervals (PI)

	sand	clay	PI^L	Mean	PI^U
Cluster	Centroids		Cluster residuals		
1	64.25	17.35	-0.12	0.00	0.11
2	87.03	8.18	-0.13	0.00	0.13
3	45.54	43.45	-0.11	0.02	0.07
4	48.66	23.83	-0.08	0.03	0.14
5	23.64	65.36	-0.13	-0.01	0.08
6	44.25	35.92	-0.14	0.01	0.09
7	64.84	23.91	-0.10	0.00	0.17
8	75.62	11.21	-0.11	-0.02	0.08
9	22.89	39.51	-0.11	-0.01	0.11
10	26.37	56.15	-0.08	-0.01	0.14
11	25.57	46.27	-0.10	0.00	0.12
Ex	—	—	-0.32	—	0.32

(b) Variance-covariance matrix

	sand	clay
sand	0.01	0.01
clay	0.01	0.02

Table C.8: Calibration cluster information for use with Eq. 2.13.

(a) Class centroids and prediction intervals (PI)

	clay	sand	PI^L	Mean	PI^U
Cluster	Centroids		Cluster residuals		
1	26.70	52.29	-0.06	-0.01	0.09
2	37.81	59.23	-0.03	-0.01	0.05
3	40.25	45.55	-0.12	-0.01	0.04
4	47.98	39.56	-0.06	0.00	0.09
5	50.44	41.40	-0.05	0.01	0.06
6	58.60	23.75	-0.06	0.00	0.08
7	44.37	34.32	-0.09	0.01	0.11
8	24.37	61.72	-0.07	-0.01	0.09
9	31.98	60.39	-0.10	0.00	0.08
10	9.27	88.80	-0.10	0.00	0.07
11	62.39	29.57	-0.07	0.01	0.08
12	21.22	76.64	-0.09	0.00	0.05
13	15.25	77.50	-0.07	0.00	0.14
14	58.15	29.72	-0.09	-0.01	0.07
Ex	—	—	-0.21	0.00	0.26

(b) Variance-covariance matrix

	clay	sand
clay	0.03	0.02
sand	0.02	0.02

Table C.9: Calibration cluster information for use with Eq. 2.14.

(a) Class centroids and prediction intervals (PI)				
	dul	PI^L	Mean	PI^U
Cluster	Centroids	Cluster residuals		
1	0.44	-0.10	-0.01	0.09
2	0.21	-0.05	0.00	0.07
3	0.41	-0.10	-0.01	0.11
4	0.35	-0.07	-0.01	0.07
5	0.12	-0.07	0.00	0.08
6	0.49	-0.11	-0.01	0.10
7	0.25	-0.03	0.00	0.06
8	0.32	-0.06	-0.01	0.09
9	0.38	-0.10	-0.01	0.09
10	0.16	-0.11	0.00	0.11
11	0.29	-0.10	-0.01	0.10
Ex	—	-0.12	0.00	0.11

(b) Variance-covariance matrix	
	dul
dul	75.02

Table C.10: Calibration cluster information for use with Eq. 2.15.

(a) Class centroids and prediction intervals (PI)					
	clay	sand	PI^L	Mean	PI^U
Cluster	Centroids		Cluster residuals		
1	25.84	61.22	-0.08	-0.01	0.10
2	22.01	59.19	-0.09	0.00	0.08
3	10.83	71.92	-0.05	0.01	0.11
4	38.25	41.58	-0.09	-0.01	0.10
5	65.36	23.39	-0.10	0.02	0.12
6	43.86	45.21	-0.08	0.01	0.14
7	24.38	48.29	-0.05	0.01	0.02
8	11.81	75.89	-0.09	0.00	0.11
9	6.02	89.77	-0.11	0.00	0.13
10	56.12	26.62	-0.10	0.00	0.09
11	19.35	72.95	-0.10	-0.01	0.11
12	46.57	25.53	-0.05	-0.02	0.11
13	40.62	23.23	-0.03	-0.01	0.06
Ex	—	—	-0.25	—	0.21

(b) Variance-covariance matrix		
	clay	sand
clay	0.02	0.01
sand	0.01	0.01

Table C.11: Calibration cluster information for use with Eq. 2.16.

(a) Class centroids and prediction intervals (PI)				
	dul	PI^L	Mean	PI^U
Cluster	Centroids	Cluster residuals		
1	0.24	-0.08	0.00	0.06
2	0.45	-0.15	0.01	0.07
3	0.39	-0.13	0.00	0.10
4	0.17	-0.04	0.00	0.05
5	0.34	-0.11	0.01	0.10
6	0.28	-0.10	0.01	0.08
Ex	—	-0.29	—	0.11

(b) Variance-covariance matrix	
dul	
dul	108.98

Chapter 3

Predicting and mapping the soil available water capacity of Australian wheatbelt

Summary

Soil available water capacity (AWC) is the main source of water for vegetation and it is the potential amount of water available for atmospheric exchange. Studying its spatial distribution is crucial for agricultural planning and management and for use in biophysical modelling. The aim of this work is to obtain a continuous spatial prediction of AWC over Australia's wheatbelt, using digital soil mapping techniques. We used a data set of 806 soil profiles which have field measurements of drainage upper limit (DUL) and crop lower limit (CLL). We mapped AWC at five depth intervals (0-5, 5-15, 15-30, 30-60, and 60-100 cm) with the help of different combinations of environmental information (topographic, climatic, soils, Landsat imagery, gamma-ray spectrometry) as covariates. The modelling techniques used were symbolic regression (GP), Cubist, and support vector machines (SVM). We also tried two averaging methods to generate an ensemble model. We observed decreasing RMSE values with the addition of extra covariates and also an expected performance decrease with soil depth. In general, SVM produced the best accuracy. We were able to improve the predictions using one of the ensemble techniques, based on a weighted average of GP, Cubist and SVM models. The

map generated with the optimal ensemble model was an unrealistic representation of AWC therefore we decided to present a sub-optimal model as the final map. We stress the need to not only focus on the numerical performance in order to obtain a flexible and stable model, and a realistic visual representation of it.

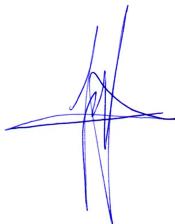
Statement of Contribution of Co-Authors

This chapter has been written as a journal article. The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit; and
5. they agree to the use of the publication in the student's thesis and its publication on the Australasian Research Online database consistent with any limitations set by publisher requirements.

In the case of this chapter, the reference for this publication is:

Padarian, J., Minasny, B., McBratney, A. B., Dalgleish, N. 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. Geoderma Regional (in press).

Contributors	Statement of contribution
José Padarian Signature:  Date: October 1, 2014	Data analysis; Writing
Budiman Minasny	Data analysis; Writing
Alex McBratney	Data analysis; Writing
Neal Dalgleish	Data collection

3.1 Introduction

Soil available water capacity (AWC) is defined as the amount of water soil can store between field capacity or drainage upper limit (DUL) and wilting point or crop lower limit (CLL). It is the main source of water for vegetation development and is related to the potential amount of water a soil could make available for the atmosphere through evapotranspiration (Dunne and Willmott, 1996). Information about its distribution in space is crucial for planning and management in agriculture, and for ecological modelling.

To model the spatial distribution of AWC, digital soil mapping has been proposed (McBratney *et al.*, 2003). The scorpan model describes that soil properties can be predicted from its predicting factors in the form of empirical regression equations. The general steps in the modelling process involve: collection of a dataset of soil observations over the chosen area of interest; compilation of relevant covariates for the area; calibration or training of a spatial prediction function based on the observed dataset; interpolation and/or extrapolation of the prediction function over the whole area of interest; calculation of uncertainty; and finally validation using existing or independent datasets.

Despite the importance of AWC, not many studies present a mapping methodology at national scale. Hong *et al.*, (2013) successfully predicted AWC for Korea based on detailed soil series maps and modal profiles, also recognising the shortcomings due to variability within mapping units. Poggio *et al.*, (2010) used morphological features as covariates, obtaining an optimal model selecting covariates using generalised additive mixed models, to map AWC in Scotland. Ugbaje and Reuter (2013) used two different covariates combinations (remote sensing data; terrain, climate, and vegetation attributes) and pedotransfer functions (PTFs) to map AWC in Nigeria, not finding a clear effect of number of covariates on model accuracy. Most of these studies used PTFs to predict the AWC. Thus the uncertainty of the map depends also on the accuracy of the PTFs.

In digital soil mapping, the visual representation of the product (map) depends on the covariates and the models used. Several studies that looked at the selection and parsimony of the covariates, and also studies have compared different data mining prediction. However no work has looked at the effect of both covariates and models on

the visual representation of the map.

A good digital soil map should have a balance of model parsimony (number of covariates), accuracy (numerical performance) and realism of the visual representations (maps). The aim of this work is to obtain a continuous spatial prediction of AWC over Australia, based on field measured data, that reconciles these three aspects, exploring the use of different covariates combinations and modelling techniques, and visually inspecting the generated maps.

3.2 Materials and methods

3.2.1 Data sets and study area

The data set used is a CSIRO Ecosystem Science (APSRU) compilation of 806 soil profiles, containing field measured DUL and CLL values, mainly distributed in productive soils (Dalgliesh *et al.*, 2012), as shown in Fig. B.1. The details of the measurement and data can be found in <http://www.asris.csiro.au/>.

A bioregion classification by Thackway and Cresswell (1995) was used to limit the study area, selecting the bioregions which contained observations of the APSRU data set. This selection, usually referred as “wheatbelt”, is represented as the greyed area in Fig. B.1 and it is equivalent to about 1.75 million km².



Fig. B.1: Location of soil profiles from APSRU database. Greyed area represents the bioregion subset where predictions were made.

3.2.2 Digital soil mapping model

In this study we used the scorpan approach (McBratney *et al.*, 2003) as an empirical quantitative descriptions of relationships between soil and other spatially referenced factors. It is represented as $S = f(s, c, o, r, p, a, n) + \epsilon$, where S : is the variable of interest (DUL and CLL), s : stands for soil (other properties of the soil at a point), c : climate (climatic properties of the environment at a point), o : organisms (vegetation or fauna or human activity), r : topography (landscape attributes), p : parent material (lithology); a : age (the time factor); n : space (spatial position); and ϵ correspond to the spatially modelled residuals (usually by kriging).

Soil attribute: S

We predicted soil properties related to water holding capacity of a soil. DUL represents the volumetric water content an initially saturated soil holds after draining for 2-3 days (Veihmeyer and Hendrickson, 1949). On the other hand, CLL corresponds to the volumetric soil water remaining in the soil after a healthy crop, with uninterrupted root development, has reached maturity under soil water-limited conditions (Hochman *et al.*, 2001). Both properties are measured in the field independently and were governed by different processes, hence different sources of error, thus we decided to model them separately.

Statistics of DUL and CLL measurements are presented in Table 3.1. We used the equal-area spline function (Bishop *et al.*, 1999) to convert the soil profile data into standard depths (0-5, 5-15, 15-30, 30-60, and 60-100 cm).

Table 3.1: Statistics of soil samples used for model generation.

	Mean	S.D.	Min.	Median	Max.
DUL (%)	30.20	11.54	3.00	32.00	56.00
CLL (%)	16.90	8.61	0.40	18.00	53.00

By definition, the predicted soil attribute S is delivered along with an uncertainty measure. This point is further elaborated in Section 3.2.3.

Factors: s, c, o, r, p, a, n

Environmental covariates are intended to explain scorpan factors and for each factor there is an extensive list of possible covariates to use. The covariates used in this work include: *a*) digital elevation model (DEM) and associated derivatives: slope (percentage), topographic wetness index (TWI) and multi-resolution valley bottom flatness (MRVBF) to try to explain factor r , as these attributes were found to explain variation in soil moisture and texture (Malone *et al.*, 2011); *b*) air maximum temperature and rainfall (summer and winter means); potential evapotranspiration (annual mean); and Prescott Index (Prescott, 1950) to try to explain factor c ; and *c*) remote sensing data (Landsat 7 imagery (2012 percentile composite), gamma-ray spectrometry (^{40}K , ^{232}Th , and ^{238}U), and weathering index (Wilford, 2012)) to try to explain o and p factor.

We also used a bioregion classification (IBRA *v6.1*) to stratify the modelling process. The use of this covariate might also be rationalised as the inclusion of information to try to explain changes in all the scorpan factors, constraining the spatial position n .

Function: f

The function f represents the connection between the soil attribute S and the scorpan factors. In this study we used three modelling techniques (described below) with different complexity. A randomly selected subset of the data (80%) was used for calibration and the remaining 20% for cross-validation. The subsampling process (cross-validation) was repeated 20 times to obtain an average error to compare the different modelling techniques.

Support vector machines (SVM) A method originally proposed by Cortes and Vapnik (1995) that looks for an optimal separating hyperplane between two classes by maximising the margin between the closest points of each class (Meyer, 2012). In a ε -regression case (as used in this study), the observations lie in between the two borders of the margin (supporting vectors), which are separated from the hyperplane by $\pm\varepsilon$ (maximum error). More detailed explanations about SVM can be found in Smola and Schölkopf (2004).

We used the R package *e1071 v1.6.1* (Meyer *et al.*, 2012), training the model with the default options.

Cubist A decision rule-based method where a tree is grown based on a succession of rules and where the terminal nodes (leaves) represent linear regression models. Originally proposed by Quinlan (1992) and Quinlan (1986), it has been widely used to model soil properties (Henderson *et al.*, 2005; Minasny *et al.*, 2008).

We used the R package *Cubist v0.0.13* (Kuhn *et al.*, 2013), training the model with the default options.

Genetic programming (GP) A machine-learning method for evolving computer programs, following the concepts of natural selection and genetics, to solve problems (Koza, 1992; Koza, 1994; Koza *et al.*, 1999). In this specific case, we used symbolic regression to fit a function to a specific data set. Its application to soil science is recent but expanding rapidly (Johari *et al.*, 2006; Makkeasorn *et al.*, 2006; Parasuraman *et al.*, 2007).

We used the software Formulize v0.98.2b, setting a stopping criteria of 25,000 generations and $F = \{+, -, *, \%\}$ as the function set to be used in the regression functions (i.e.: building-blocks).

Residuals: ϵ

The scorpan approach implies the addition of the spatial correlation structure of the model residuals to the predictions. The assumption of kriging is the stationarity in the process (the residual have the same mean = 0 everywhere). At local scale, data density is usually high and the distribution is more-or-less homogeneous, so kriging of the residuals does not present further complications. This is not necessarily true for a continental scale using legacy data. Data is usually clustered in space leaving extensive areas without information. When the kriging method is applied to clustered data, the resulting map usually presents artefacts due to interpolation between distant clusters or extrapolation. In addition the stationary process may not hold. For these reasons we decided to omit this step, which could be addressed in future studies.

3.2.3 Prediction and mapping

Covariates selection

Due to complexity of the interactions between the soil attribute S and the *scorpan* factors, the selection of covariates is an important part of the modelling process.

In general, the addition of covariates to the modelling process improves predictions, but that depends on which and how many covariates are being included. For instance, when covariates are highly correlated to each other, the addition of new covariates does not produce better results (Ugbaje and Reuter, 2013). Another possible problem is model overfitting, where the resulting model represents not only real relationships but also relationships occurring by chance in the training data that may be absent in other observations (Carr, 1988). An example of the latter could be found in Table 5 of McBratney *et al.*, (2000), where adding all the available covariates generated a model with higher error compared with a model using a sub-set of them.

Two groups of method for covariate selection exist, the “filter” and the “wrapper” approach (John *et al.*, 1994). The filter approach consists in pre-processing the covariates, identifying the relevant ones, via statistical procedure or expert knowledge (Lark *et al.*, 2007), and using selected covariates for the model calibration. On the other hand, the wrapper approaches uses different sub-sets of covariates to calibrate the model and the accuracy of the predictions using different combination of covariates are compared.

To evaluate the effect of covariate selection on prediction quality, we used the wrapper approach, grouping the covariates in sub-sets by categories, using the concept of *scorpan* model, and to add them sequentially to the model, gradually increasing the complexity. The data groups used correspond to: *a*) bioregion classification, *b*) topographic attributes (slope, TWI and MRVBF), *c*) landsat imagery, *d*) gamma-ray spectrometry, *e*) weathering index, and *f*) climate data (air temperature, rainfall, evapotranspiration, Prescott Index).

The data was split into 80:20 for prediction and validation. The 3 models (GP, Cubist and SVM) were calibrated for each of the covariate selection, and for each model, 20 model cross-validation fittings were performed. An ANOVA/Tukey analysis was computed to determine the difference of performance as a function of covariate combinations.

Ensemble Model

The “best model” is a subjective concept, usually defined as the model which produced the lowest error (or RMSE) in a validation dataset. As an alternative to selecting this “best model”, we suggest the use of model averaging. Model averaging consists of creating multiple models and combining them to obtain a single final model. The advantage of this method is that most of the time, the combined model performs better than any of the individual models. It is a method used for almost 200 years as pointed in an interesting review by Clemen (1989).

We tested the idea of model averaging for the 3 different prediction models (GP, Cubist and SVM).

We used two averaging techniques, with different weighting criteria. First, an equal weight average (EWA), where the final prediction is obtained assigning the weights $\hat{\beta}_{\text{EWA}} = \{\frac{1}{k}, \dots, \frac{1}{k}\}$, where $\hat{\beta}$ is vector of weights and k is the number of models. The second corresponds to an averaging technique used by Granger and Ramanathan (1984), where the weights correspond to the ordinary least squares estimates of a multiple linear regression, $\hat{\beta}_{\text{GRA}} = (X^T X)^{-1} X^T y$, where X and y stand for the matrix of predicted values of the different models and the vector of observed values, respectively (Diks and Vrugt, 2010).

Mapping and validation

All the covariates were available as raster files. As a pre-process, they were re-sampled (average) to match a 500 m grid. In order to complement the numerical evaluation of the models with a visual evaluation, we applied the fitted models of DUL and CLL to the whole extent, for a grid spacing of 500 m, using all the covariate combinations and modelling techniques. To validate the models we used the remaining observations (20% randomly selected from the APSRU dataset) from the 20 iterations of the modelling process.

Uncertainty assessment

As we mentioned in Section 3.2.2, the soil attribute S should have an associated uncertainty level. This is a measure to evaluate the risk involved in using the predictions for a decision-making process (Goovaerts, 2001). We estimated the uncertainty of the

predictions using the fuzzy k-means with extragrades algorithm (Tranter *et al.*, 2010). This method classifies the covariates values at the observed points (observations used in model training) in clusters. Each cluster has a central value (centroid), and an associated range of error estimated from the $\alpha/2$ and $1 - \alpha/2$ (α : significance level) quantiles of the prediction residuals. When a new value is predicted, the distance between values of its covariates and the centroids of the clusters is estimated and a membership grade assigned (grade of “belongingness” to each cluster). Finally, the error ranges of the clusters are weighted by the membership grades and added to the prediction. The advantage of this method is that if the covariates of a new prediction are too dissimilar to the ones used in the model training process (high distance from centroids), the prediction is assigned to the extragrade class and its error penalised.

3.3 Results

3.3.1 Covariates selection

We generally observed that adding groups of covariates decreased the magnitude of the cross-validated error (Fig. B.2).

As we mentioned in Section 3.2.3 in some studies the addition of extra covariates does not yield better results. In other cases, a smaller number of covariates is preferred, following parsimony and Occam’s razor principle (Blumer *et al.*, 1987). In this study we decided to select all the covariates for two reasons: there is not significant loss of accuracy when using the maximum number of covariates, and all the covariates are already available for future improvements of the models (with the addition of more observations).

To clarify the analysis, we defined the analysis with three different combinations of covariates which reflect the complexity: *a*) using bioregions and topographic attributes (from hereon COV 1), *b*) using bioregions, topographic, weathering and climate data (from hereon COV 2), and *c*) using bioregions, topographic, gamma-ray, Landsat, weathering and climate data (from hereon COV 3).

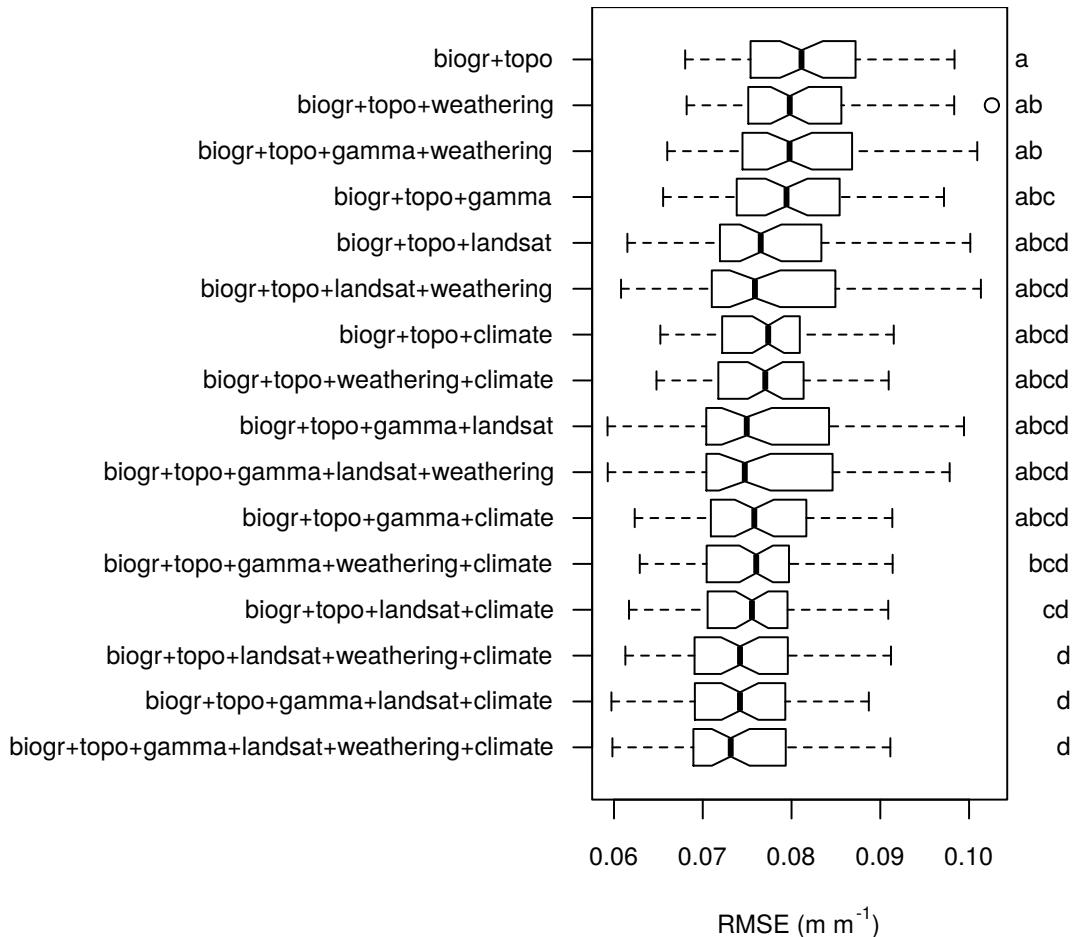


Fig. B.2: Boxplot of model cross-validated RMSE (20 iterations) trained with different combinations of covariates. The prediction corresponds to DUL at 0-5 cm depth. bio: bioregion; topo: slope, TWI and MRVBF; weathering: weathering index; gamma: gamma-ray spectrometry; landsat: Landsat 7 bands; climate: air temperature, rainfall, evapotranspiration, Prescott Index. Letter on the right margin represent mean groups after an ANOVA/Tukey analysis.

3.3.2 Ensemble Model

Performance of individual models was consistent with SVM showing the best results, followed by Cubist and GP (Table 3.2). Regarding the averaging method, the GRA method tends to have a better performance, showing results which are as good or better than the best of the three individual models. This error decrease in GRA indicates that all the models produced unbiased predictions (Draper, 1995) leading in a reduction of

the uncertainty. The superior performance of GRA is an expected result. This kind of model averaging is widely used in machine learning (Hashem, 1997).

Analysing the effect of the number of covariates used in the modelling process, this time with a different aggregation level (ANOVA/Tukey was performed on each modelling technique), it is possible to observe a change in the tendency observed in Fig. B.2. When we added all the covariates, the error tended to increase, a phenomenon known within the statistical literature as *Simpson's paradox* (Simpson, 1951). Due to the error increase being non-significant, we decided to continue with all the covariates.

Table 3.2: RMSE values of DUL validation between 0 and 5 cm depth. Mean of 50 iterations by model and covariate combinations. RMSE range between brackets. Units in m m^{-1} .

	GP	Cubist	SVM	EWA	GRA
COV 1	0.059 (0.051–0.07)	0.049 (0.044–0.055)	0.049 (0.044–0.054)	0.05 (0.045–0.055)	0.048 (0.043–0.053)
COV 2	0.052 (0.045–0.071)	0.046 (0.04–0.052)	0.044 (0.04–0.048)	0.045 (0.039–0.054)	0.043 (0.038–0.048)
COV 3	0.052 (0.045–0.063)	0.047 (0.042–0.053)	0.046 (0.041–0.05)	0.046 (0.041–0.052)	0.044 (0.04–0.049)

3.3.3 Visual evaluation

So far, we have explored several combinations of covariates, modelling techniques, and ensemble methods. Results showed that increasing the complexity of the model and complexity of the covariates will decrease the cross-validated error.

However, the approach that is usually overlooked is that the generated maps can be different for different combinations of covariates and models. Which model should be used? Should we aim for the model which produces the least error? Or should we subjectively choose a model which is more realistic?

We recommend using expert knowledge to visually evaluate the generated maps including but not limited to the criteria pointed in this section.

Artifacts

Artifact is the name assigned to any error observed in a digital signal. Remote sensing data and its derivatives correspond to representations of signal captured by electronic sensors and processed by algorithms. All the equipment and techniques used to generate this information is prone to introduce error, therefore is necessary to make sure that the covariates used are artifact-free.

Fig. B.3 shows the map produced using COV 3 and GRA. The map showed undesirable artifacts generated by one of the covariates (Fig. B.3a). Removing this covariate (^{232}Th data) did not significantly decrease the performance of the model and allowed us to obtain a more realistic representation (Fig. B.3b). Although ^{232}Th data is indicative of the parent material (Wilford, 2012) the continental data can be noisy in some parts of the country. We suggest that checking for anomalies and unrealistic representations is a recommended step.

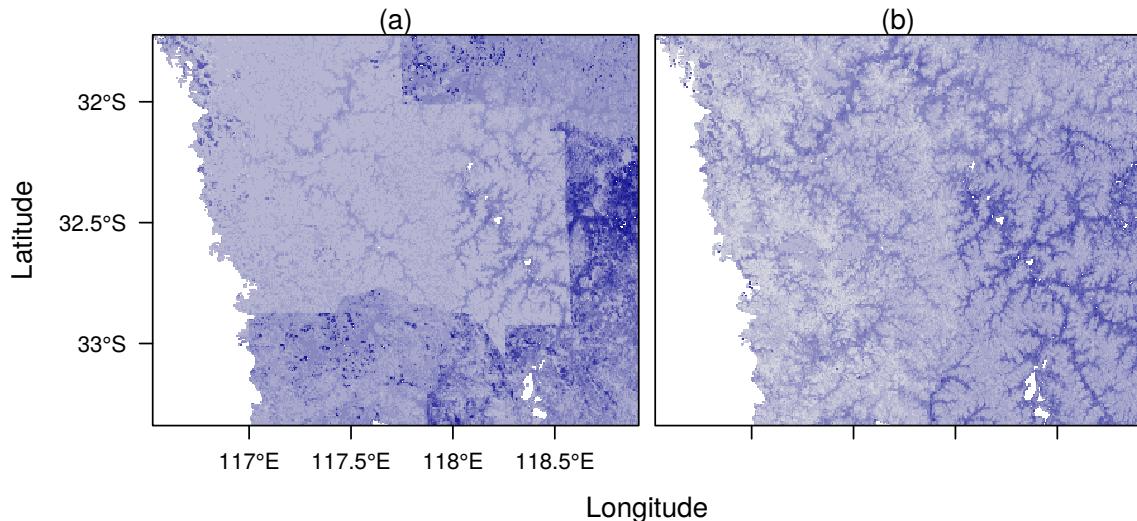


Fig. B.3: Subarea showing artifact caused by ^{232}Th data on DUL map. (a) Map with artifact, and (b) map without artifact.

Models concordance

If more than one modelling technique is evaluated, a measure of concordance between the models is a good indicator of areas where potential mapping problems could be

found.

We estimated and mapped the standard deviation of the three modelling techniques as a measurement of agreement between the models, using the three covariates combinations mentioned in Table 3.2. Using COV 1 combination we could observe an important discordance between the models, specifically by GP which only used the bioregion classification as a predictor, generating a sharp contrasts observed in Fig. B.4a. As we increased the number of covariates, the discordance between models also increased, especially because SVM (and Cubist to a lesser extent) captured more complex interactions in some bioregions. This was evident in areas like Cobar Peneplain, central NSW (distinguished from most of the surrounding bioregions which are relatively flatter landscapes); Esperance Plains, south-east WA, and South Eastern Queensland (which present strong marine influence and limited by abrupt ranges); or the highest areas of South Eastern Highlands, and Nandewar, NSW. Besides the distinctiveness of these areas, they are also poorly represented by soil samples in the dataset.

Out-of-range predictions

When the fitted models are applied to the whole extent, it is possible to find combination of covariates absent in the training process, due to non-sampled or poorly represented areas. This is the reason why a systematic sampling schema should be used when possible. The modelling procedure is based on point soil observations and if the modelling technique is not robust enough, that could lead to obtain predictions beyond the expected limits of a soil property.

For instance, at South Eastern Highlands and South Easter Queensland bioregions, GP predicts negative values of CLL in depth (60-100 cm), and the ensemble model produces maps with anomalies (Fig. B.5). In Section 3.3.2 we remarked that GRA ensemble-method generates the best numerical results, but after this final step, the “best model” was discarded. We continued the analysis with the second-best model, namely SVM with all covariates (excluding ^{232}Th data, see Section 3.3.3)

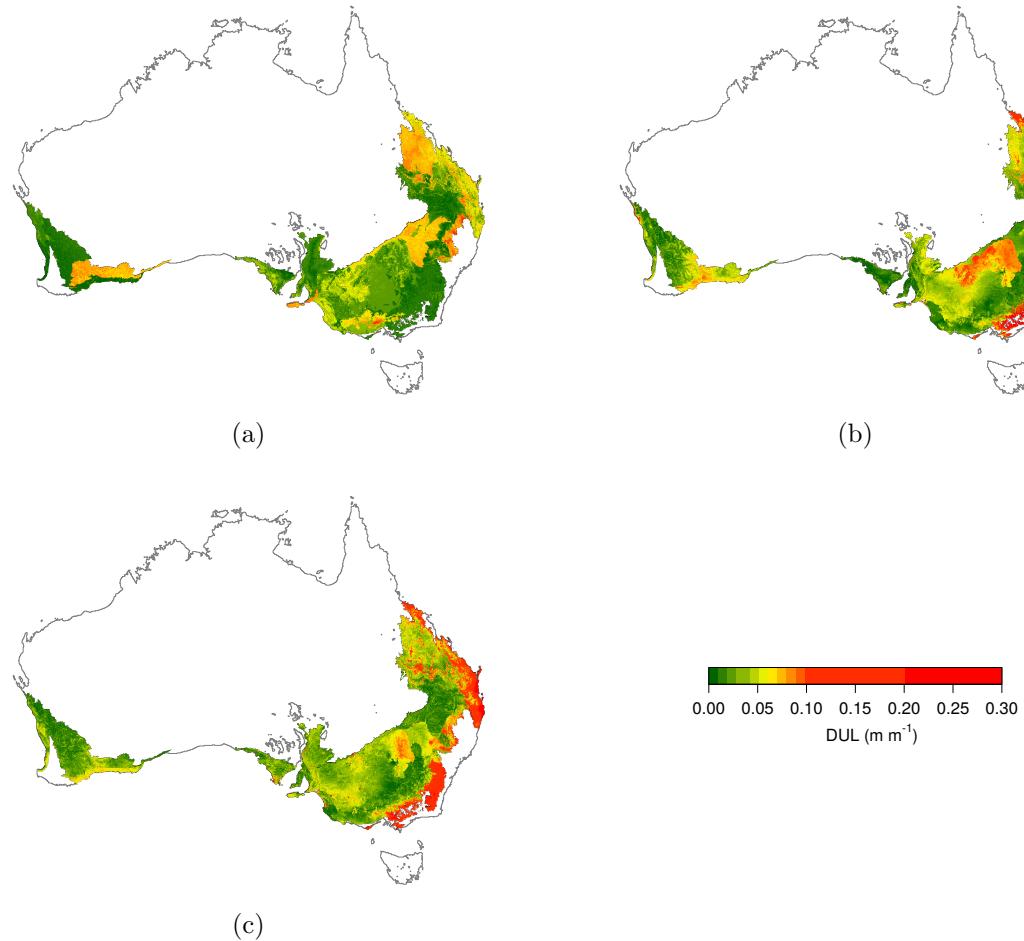


Fig. B.4: Standard deviation of DUL predictions, between 0 and 5 cm depth, of SVM, Cubist, and GP models, with different combinations of covariates. (a) COV 1, (b) COV 2, and (c) COV 3.

3.3.4 Validation

Based on the numerical and visual evaluation, we selected the DUL and CLL models generated with SVM, using COV 3 and excluding the covariate ^{232}Th , as our final map. Table 3.3 shows model validation errors in depth. As expected, the performance of the models decreased with depth. This has also been observed by Malone *et al.*, (2011) who predicted AWC in the agricultural district Edgeroi, NSW, Australia (30.32S, 149.78E), and for other soil properties predictions like organic carbon (Minasny *et al.*, 2006; Jobbág and Jackson, 2000). Deeper layers of soil are not as exposed to weathering

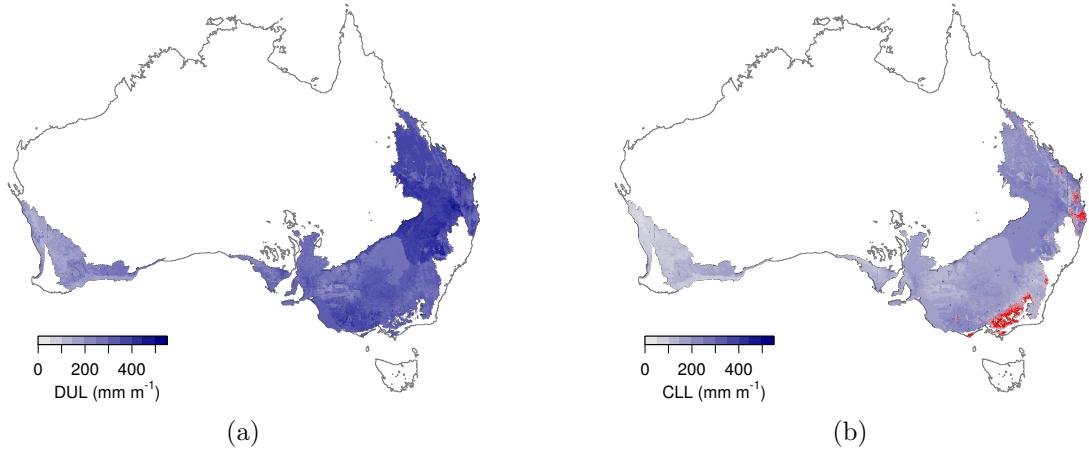


Fig. B.5: Water content to 1 meter depth ($mm\ m^{-1}$) based on ensemble model, using COV 3. Red colour represents negative values. (a) Drainage upper limit, and (b) Crop lower limit.

factors as the top layers, therefore the correlations with climatic and remote sensing data (that mainly reflect the surface condition) tend to be lower.

Table 3.3: R^2 values for validation of DUL and CLL in depth, using SVM and COV 3.

	DUL	CLL
0-5 cm	0.6710	0.6906
5-15 cm	0.6558	0.6772
15-30 cm	0.6090	0.6214
30-60 cm	0.5725	0.5640
60-100 cm	0.4906	0.4915

3.3.5 AWC map

The final step in the calculation of AWC is to estimate the difference between DUL and CLL. We subtracted both values and used the lower and higher prediction limit of CLL and DLL respectively to represent the uncertainty (Fig. B.6).

As we pointed in the visual evaluation (Section 3.3.3), models tended to have lower performance in poorly sampled areas, specially the highest areas of the landscape. The

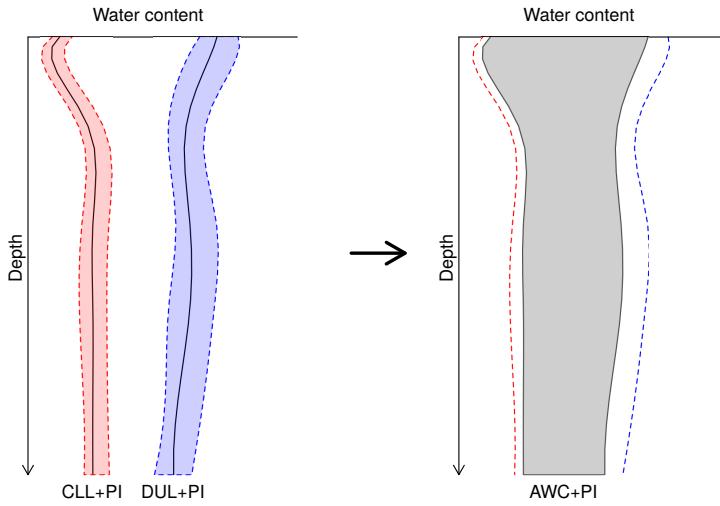


Fig. B.6: Diagram of estimation of AWC based on DUL and CLL values with their respective uncertainty levels.

uncertainty follows the same trend, presenting wider prediction intervals in these areas (Fig. B.7).

The final map (Fig. B.8) shows a good representation of the AWC distribution over the continent. The WA region presents the lower water contents, in concordance with the sandy soil located in the area. The soil-heterogeneous VIC area seems well represented as well, with the appearance of Vertosols in the southern region with higher AWC associated with them. In the northern area of VIC, close to the border with NSW, the sharp transition between the Cobar Peneplain, dominated by Kandosols, and the north-eastern bioregions, dominated by Vertosols, is clearly represented. QLD area generally presents high values of AWC in concordance with clayey textural classes, but areas in Brigalow Belt North and South Eastern Queensland have a small number of observations thus limited by the higher uncertainty levels.

3.4 Conclusions

We explored the use of digital soil mapping approach to model AWC in Australia, balancing three important aspects of it, which are not discussed in previous studies: model parsimony, accuracy and realism of the visual representations. We also explored

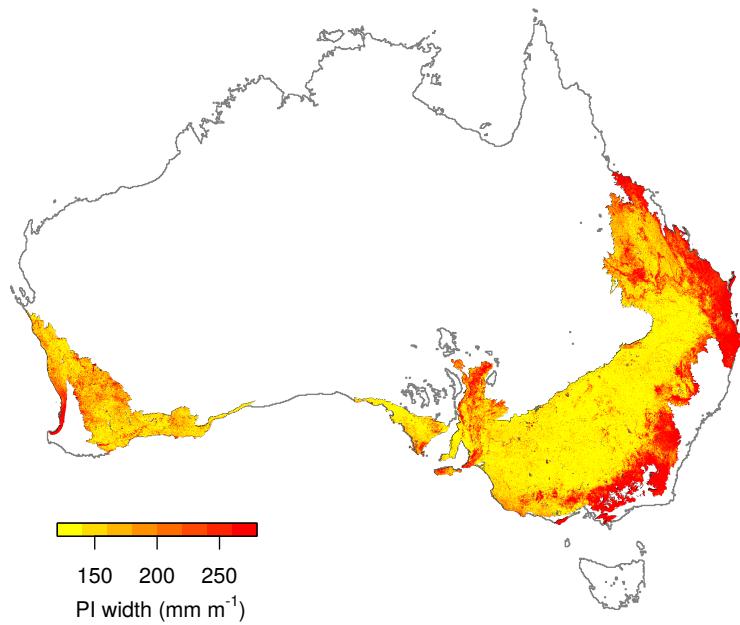


Fig. B.7: Prediction interval width (mm m^{-1}) based on SVM predictions, using COV 3, to 1 meter depth.

the use of ensemble methods (i.e.: model averaging) as an alternative to single model selection.

We used different combinations of environmental covariates to represent the various process involved in soil formation. In many studies the use of multiple covariates does not yield better results compared with simpler models. In our case the combination of all the available covariates showed the best accuracy.

We tried three different modelling techniques, namely symbolic regression (GP), Cubist and support vector machines (SVM). In general, SVM presented the best accuracy. We were able to improve the predictions generating an ensemble model (least squares or GRA method), based on a weighted average of GP, Cubist and SVM models.

After visually evaluating the generated maps, we decided to present a sub-optimal model, generated with SVM because it generated a more realistic representation compared with the optimal GRA model. The final AWC map is a good representation of the study area, except in poorly sampled areas, where the uncertainty levels increase considerably.

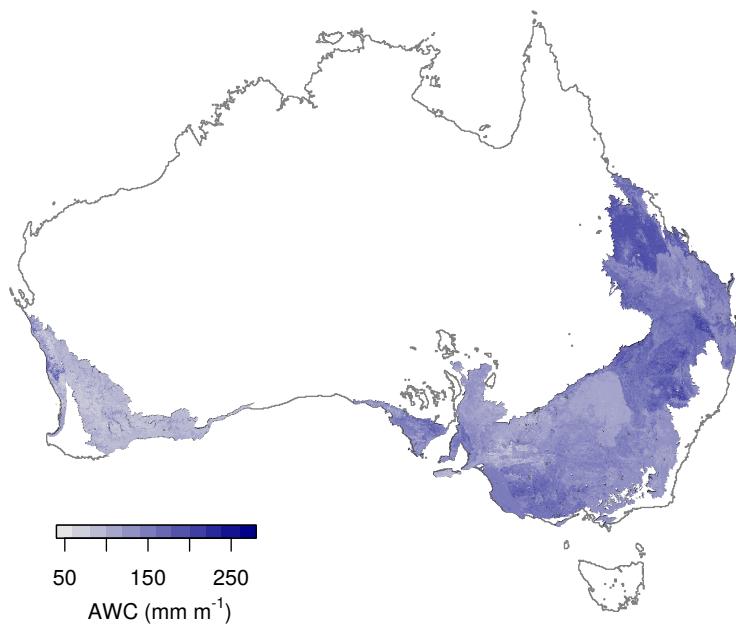


Fig. B.8: Available water content (mm m^{-1}) based on SVM predictions, using COV 3, to 1 meter depth.

Balance model parsimony, accuracy and realism of the visual representations is a sensible aspect of the digital soil mapping approach that could be achieved in different ways, but we stress the need to consider the knowledge about the modelled process and not only focus on the numerical performance in order to obtain a flexible and stable model, and a realistic visual representation of it. The main reason is that model evaluation or validation is usually only based on the smallest error or uncertainty. However the evaluation is based on point observations which do not reflect the spatial representation. We should have a more objective way to evaluate the spatial representation of digital soil maps.

3.5 References

- Bishop, T., McBratney, A., and Laslett, G. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma* 91 (1): 27–45.
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. 1987. Occam's razor. *Information processing letters* 24 (6): 377–380.
- Carr, M. B. 1988. Determining the optimum number of predictors for a linear prediction equation. *Monthly weather review* 116 (8): 1623–1640.
- Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5 (4): 559–583.
- Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20 (3): 273–297.
- Dalgliesh, N., Cocks, B., and Horan, H. 2012. APSoil—providing soils information to consultants, farmers and researchers. In: *16th Australian Agronomy Conference, Armidale, NSW*.
- Diks, C. G. and Vrugt, J. A. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment* 24 (6): 809–820.
- Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*: 45–97.
- Dunne, K. A. and Willmott, C. J. 1996. Global distribution of plant-extractable water capacity of soil. *International Journal of Climatology* 16 (8): 841–859.
- Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. *Geoderma* 103 (1): 3–26.
- Granger, C. W. and Ramanathan, R. 1984. Improved methods of combining forecasts. *Journal of Forecasting* 3 (2): 197–204.
- Hashem, S. 1997. Optimal linear combinations of neural networks. *Neural networks* 10 (4): 599–614.
- Henderson, B., Bui, E. N., Moran, C., and Simon, D. 2005. Australia-wide predictions of soil properties using decision trees. *Geoderma* 124 (3): 383–398.
- Hochman, Z., Dalgliesh, N., and Bell, K. 2001. Contributions of soil and crop factors to plant available soil water capacity of annual crops on Black and Grey Vertosols. *Crop and Pasture Science* 52 (10): 955–961.

- Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., and Lee, K. 2013. Predicting and mapping soil available water capacity in Korea. *PeerJ* 1: e71.
- Jobbágy, E. G. and Jackson, R. B. 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological applications* 10 (2): 423–436.
- Johari, A., Habibagahi, G., and Ghahramani, A. 2006. Prediction of Soil-Water Characteristic Curve Using Genetic Programming. *Journal of Geotechnical and Geoenvironmental Engineering* 132 (5): 661–665.
- John, G. H., Kohavi, R., Pfleger, K., et al., 1994. Irrelevant Features and the Subset Selection Problem. In: *ICML*. Vol. 94: pp. 121–129.
- Koza, J. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. *The MIT Press*.
- Koza, J. 1994. Genetic Programming II: Automatic Discovery of Reusable Subprograms. *The MIT Press*.
- Koza, J., Bennett, H., Andre, D., and Keane, M. 1999. Genetic Programming III: Darwinian Invention and Problem Solving. *Morgan Kaufmann Publishers*.
- Kuhn, M., Weston, S., Keefer, C., and Quinlan, N. C. C. c. f. C. b. R. 2013. *Cubist: Rule- and Instance-Based Regression Modeling*. R package version 0.0.13.
- Lark, R., Bishop, T., and Webster, R. 2007. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma* 138 (1): 65–78.
- Makkeasorn, A., Chang, N., Beaman, M., Wyatt, C., and Slater, C. 2006. Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. *Water Resources Research* 42 (9) W09401: W09401.
- Malone, B., McBratney, A., and Minasny, B. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160 (3): 614–626.
- McBratney, A., Mendonça Santos, M. d. L., and Minasny, B. 2003. On digital soil mapping. *Geoderma* 117 (1): 3–52.
- McBratney, A. B., Odeh, I. O., Bishop, T. F., Dunbar, M. S., and Shatar, T. M. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma* 97 (3): 293–327.
- Meyer, D. 2012. Support Vector Machines. *The Interface to libsvm in package e1071. e1071 Vignette*.

- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. 2012. *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien. R package version 1.6-1.
- Minasny, B., McBratney, A., Tranter, G., and Murphy, B. 2008. Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties. *European Journal of Soil Science* 59 (5): 960–971.
- Minasny, B., McBratney, A. B., Mendonca-Santos, M., Odeh, I., and Guyon, B. 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Soil Research* 44 (3): 233–244.
- Parasuraman, K., Elshorbagy, A., and Carey, S. K. 2007. Modelling the dynamics of the evapotranspiration process using genetic programming. *Hydrological Sciences Journal* 52 (3): 563–578.
- Poggio, L., Gimona, A., Brown, I., and Castellazzi, M. 2010. Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. *Geoderma* 160 (2): 175–188.
- Prescott, J. A. 1950. A climatic index for the leaching factor in soil formation. *Journal of Soil Science* 1 (1): 9–19.
- Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1 (1): 81–106.
- Quinlan, J. R. 1992. Learning with continuous classes. In: *Proceedings of the 5th Australian joint Conference on Artificial Intelligence*. Vol. 92. Singapore: pp. 343–348.
- Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*: 238–241.
- Smola, A. J. and Schölkopf, B. 2004. A tutorial on support vector regression. *Statistics and computing* 14 (3): 199–222.
- Thackway, R and Cresswell, I. 1995. An interim biogeographic regionalisation for Australia: a framework for establishing the national system of reserves. *Australian Nature Conservation Agency, Canberra*.
- Tranter, G., Minasny, B., and McBratney, A. 2010. Estimating Pedotransfer Function Prediction Limits Using Fuzzy k-Means with Extragrades. *Soil Sci. Soc. Am. J.* 74 (6): 1967–1975.

- Ugbaje, S. U. and Reuter, H. I. 2013. Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. *Vadose Zone Journal* 12 (4).
- Veihmeyer, F. and Hendrickson, A. 1949. Methods of measuring field capacity and permanent wilting percentage of soils. *Soil science* 68 (1): 75–94.
- Wilford, J. 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. *Geoderma* 183: 124–142.

Chapter 4

General discussion, conclusions and future research

4.1 General discussion

PTFs maintenance and centralisation

During the development of this project it was evident that a large number of PTFs are currently available. PTFs try to represent a natural system, dynamic by nature. Another evolving matter is the amount of information available for creating PTFs. Due to these two factors, it is expected that PTFs also evolve. In Chapter 1, I generated an improved version of a PTF first proposed by Minasny and McBratney (2001). This improvement was possible using a bigger dataset. One may argue the novelty of this approach, but it is undoubtedly important to “update” PTFs at some point, as new data comes to hand.

This high availability of PTFs may be confusing for the end-users. It is the end-user’s responsibility to: look for the appropriate PTF; compare the data used in the PTF calibration process (when available) with the data he/she wants to predict; and/or if the PTF is applicable in the region of interest. Presumably, the end-user is capable of performing this process, but why not give him/her the means to facilitate this process?. McBratney *et al.*, (2002) have already discussed the potential benefits of an inference system to predict other soil properties selecting the PTF with better performance. I see a great benefit behind this idea, not just as an “oracle” but as a

knowledge organisation system. It is the responsibility of the soil science community (or any scientific community) to organise the information generated by them and prevent its misuse. The framework proposed in Chapter 2 a step closer to standardisation of information, and it could be easily implemented in an inference system like the one described by McBratney *et al.*, (2002).

“Big mapping” challenge

Initiatives like GSM and TERN aim to provide soil information to a wide audience in form of soil maps. The efforts to map big areas date back to China (4,000 years ago) or the Roman empire based on soil suitability for plant growth or other uses (McDonald, 1994). Modern efforts are based in more complex principles, aiming at detail never seen before, which undoubtedly presents new challenges for the soil community.

As we mentioned in Chapter 3, methodologies to provide national/continental maps are relatively new. Traditional approaches like Monte Carlo for uncertainty assessment are still valid, but they become more restrictive due to the computation time/resources used. Working with big maps implies working with more efficient methods, and optimised code, things not always present in the “toolbox” of a soil scientist. Besides this scale change, working with increasingly larger areas, the detail of the studies is also an evolving matter. Increasing the details not only implies collecting more data, which is a challenge in itself, but also increasing the computer resources. Duplicate the detail of a digital map (e.g.: from 500 to 250 m resolution) means quadruplicate the number of pixels and the time/resources associated.

Performance of extragrades class in PTFs

The extragrade class corresponds to marginal observations of each cluster, considering a n -dimensional space defined by all the soil properties used to calibrate a PTF. The concept is similar to the widely used bioclimatic envelope model, which uses associations between aspects of climate (climate variables forming a n -dimensional space) and known occurrences of species across landscapes of interest to define sets of conditions under which species are likely to maintain viable populations (Araújo and Peterson, 2012). The model defines a marginal bioclimate where the conditions are not favourable for the development of the species (usually with fewer specimens compared

with the core bioclimate). The limit between both bioclimate is defined by a threshold (usually between 90-97%) that must be defined generally combining expert knowledge and absence records (Carpenter *et al.*, 1993), and, of course, varies within species.

This comparison presents interesting differences. In the envelope model, the difference between marginal and the core bioclimate are notorious (with fewer specimens in the marginal bioclimate). In this study, we could not find a significant difference in the prediction error of observations in the extragrade class (marginal) compared with the observations of any other cluster (core). PTFs are governed by other processes, and some soil properties are strongly driven by physics, which tend to be more universal, which may lead to differences compared with the envelope model. Considering this, it is necessary to expand the concept of extragrade class to define how far is possible to extrapolate the predictions of PTFs.

4.2 Overall research conclusions

The research presented here has been successful at creating a framework to address soil data needs, using as example, soil properties related to water holding capacity (drainage upper limit (DUL) and crop lower limit (CLL)) within Australia.

In the case when additional soil information is available, the use of pedotransfer functions (PTFs) is recommended. I successfully generated a group of them, using symbolic regressions, based on soil data availability. I used the fuzzy k-means with extragrades algorithm to solve two important information delivery issues related to PTFs, namely assessment of uncertainty levels and delineation of data domain.

When no extra soil information is available, the spatial location should be enough to obtain information about DUL and CLL in the area of interest. I generated a spatial model of AWC using a digital soil mapping approach. I balanced three important aspects of it, which are not discussed in previous studies: model parsimony, accuracy and realism of the visual representations. The predictions of this model are also delivered with uncertainty levels and domain delineation, extending the approach applied to PTFs generation to a spatial context.

The implementation of this framework should automatically deliver predictions with uncertainty levels and, when using PTFs, information about the data domain used in the training. Despite the specific methods used to make predictions and to assess

uncertainty levels, I remark that the aim of the framework is not to enforce the use of specific methodologies but to deliver detailed information to the end-users to avoid erroneous interpretation of predictions.

4.3 Future work

There are many opportunities for future work and some of these have been briefly mentioned in previous chapters. Opportunities include:

- (i) *Uncertainty levels for particle-size classification systems transformation:* Chapter 1 had an exploratory mission, specifically for the use of symbolic regression. I extended the PTFs generation in Chapter 2 addressing uncertainty level, thus this step has to be applied to the PTFs to transform from the Australian to the USDA/FAO soil particle-size classification system.
- (ii) *Further development of uncertainty propagation:* In Chapter 2 I estimated DUL and CLL with their respective uncertainty levels to finally calculate AWC (DUL - CLL). I kept the lower prediction interval of CLL and the upper prediction interval of DUL as a measure of the uncertainty of AWC. This procedure assumes that the total uncertainty calculus is fully compositional, thus it can be calculated systematically from the uncertainty of its components. This simplification is not necessarily true, and it is important to assess this issue, specially when including PTFs on inference systems where using predicted variables as predictors is a tempting alternative.
- (iii) *Further development of PTFs and maps for key properties:* In this project I focused in soil water retention properties but the need of soil data goes beyond this group. As mentioned in Section 2.2.1, eight key soil properties are consistently used by biophysical models in Australia, thus the obvious necessity of model the remaining ones, namely BD, OC, K_{sat} , and $K_{erosion}$.
- (iv) *PTFs transportability:* I mentioned that PTFs should not be used beyond the geomorphic region or soil type from which it was developed, since they may lose their validity. With the identification of the data domain proposed in Chapter 2, using the fuzzy k-means with extragrades algorithm, it would be possible to

identify if soils from another regions are within the data domain of the generated PTFs. An evaluation of the predicting capabilities of the PTFs with information of similar soils but from completely different geographic areas seems a logic future step.

- (v) *Further development of methodology to add kriged error ϵ to scorpan model at continental scale:* As mentioned in Section 3.2.2, I omitted the step of adding the spatial correlation structure of the model residuals to the predictions due to the clustering observed in data at continental scale. It is complicated to overcome the fact that big areas does not count with soil samples but a estimations of error per cluster (subareas) could be considered, in addition to an estimate about the behaviour of the error in poorly sampled areas.
- (vi) *Further development of visual evaluation of maps:* A few general considerations were followed in the development of this project, specifically in Section 3.3.3, but it is still a methodology under development. It is necessary to condense the traditional expert knowledge used in map visual evaluation, and find patterns of error to create general rules and make the procedure more objective.

4.4 References

- Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. *Ecology* 93 (7): 1527–1539.
- Carpenter, G., Gillison, A., and Winter, J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity & Conservation* 2 (6): 667–680.
- McBratney, A. B., Minasny, B., Cattle, S. R., and Vervoort, R. 2002. From pedotransfer functions to soil inference systems. *Geoderma* 109 (1–2): 41–73.
- McDonald, P. 1994. *The literature of soil science*. Vol. 4. Cornell University Press.
- Minasny, B. and McBratney, A. 2001. The australian soil texture boomerang: a comparison of the australian and USDA/FAO soil particle-size classification systems. *Aust. J. Soil Res.* 39 (6): 1443–1451.