Steps Toward A Universal Soil Classification

Philip Andrew Hughes

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy.

2016

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 fulfillment of the requirements for the
Doctor of Philosophy.

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

Signature: Philip A. Hughes

III
Thesis summary

This thesis comprises of an introduction and literature review (chapter one), followed by original research in chapters two to seven. Chapter eight contains the final conclusions as well as description of future work required.

Chapter one covers the literature regarding the historical importance of soils and agriculture, and discusses briefly the disconnection that has occurred between people and soil as a result of urbanization. The story commences during the 1900s, where the inception of soil as a discipline is highlighted. From here the discourse moves to taxonomic schemes focusing on the segmented nature of many schemes around the world and the requirements for a unified system. This leads to explanations of the function of several global taxonomic systems. Two of the most significant taxonomic systems, the World Reference Base (WRB) and Soil Taxonomy (ST) are compared, and the practicality of using one of these major systems for a global taxonomy is discussed. The highlighted strengths, weaknesses and general suitability of these world systems leads to the inevitable conclusion that despite the efforts of various agencies, there is still no unified soil taxonomic system that covers every situation. The introduction outlines the pertinent points that have been discovered in the course of four years of research.

Chapter two discusses the initial steps of improving/creating a world taxonomic system. It sets the scene of taxonomy in Australia, champions the place of numerical methods in soil classification schemes, and suggests that many of the philosophical problems of soil classification can be solved if mathematical methods are adopted. To that end a new algorithm, akromeson, is formulated. This algorithm is adapted from fuzzy k means with extragrades (FKME). In this new technique, extragrades, data with extreme characteristics (not unlike outliers), are mitigated in a different manner. Instead of relegating this extreme data to a single extragrade group, it nominates key extragrade points, turning them into the focal point of clusters. Data that is not extreme is free to resolve into typical fuzzy k means (FKM) clusters within the bounds of these end point clusters. The
method was tested on a data set from Edgeroi, NSW and the results identified geological features in
the landscape despite the fact that there was no direct geological information residing within the
Edgeroi data set. This was seen as good evidence for successful clustering.

Chapter three takes the concept of akromeson and applies it to data sourced from the United States
Department of Agriculture (USDA) in Lincoln Nebraska. In the pursuit of better definition for ST
Ochric horizons, soil surface horizon data was partitioned via the akromeson algorithm. This chapter
discusses the implications of this new method of horizon partitioning in a generalized world
classification such as the US based ST. Implications of this work include the assessment of soil
horizon thickness in ST and the role of end point clusters in detecting and monitoring anthropogenic
changes to soil. This analysis is an important step in the assessment of ST, determining if it truly is a
global scheme, and identifying ideas or methods by which it could be made more acceptable to other
classification schemes.

Chapter four builds on the idea that numerical methods can aid classification and addresses the
question of how to use mathematical approaches to bridge the gap between two different taxonomies;
the US based ST, and the Australian Soil Classification (ASC). To this end the ideas of principal
component analysis and taxonomic distances are put to the test. Algorithms regarding next nearest
neighbour analysis at the Order level for both taxonomies enables a new statistical test, for taxonomic
accuracy. This is a measure of the overall dispersion of Soil Taxonomy compared to the ASC. The
implications of this comparison are determined. Of great importance in this section is the
harmonization of data pertaining to both ST and the ASC. Any given soil taxonomic description can
be summed up as the mean value of 23 distinct properties at 18 standardized depth increments. This
data harmonization enables several mathematical steps that will eventually lead to the creation of a
global taxonomy scheme.

Chapter five consists of work on the mathematical methods required for a global taxonomy. It
becomes apparent that placing all the soil taxonomic allocations from both ST and ASC in the same
taxonomic space is not sufficient if a global taxonomy is to be created. The harmonized data needs to be tested to ensure that both taxonomies have been accurately converted into the same taxonomic space. This is not easy to determine as there are a myriad of differing protocols and standards by which the soils of each classification scheme have been allocated. In order to range the two taxonomies, two Orders, almost identical in pedological description, Vertisols (Vertosols) and Spodosols (Podosols) are selected for analysis. They are converted to principal component space, the convex hulls calculated and overlaid. This procedure is then performed along with mean taxonomic distances for each Order to gain an understanding on how each Order from each classification system compares to Orders of the opposing taxonomic system. This analysis highlights the ST Order Andisols. Its comparison with the ASC Orders was then studied. Next nearest neighbour analysis is then further developed to determine which tiers of which taxonomy are on a similar operational level. This first of a kind analysis has broad implications, which require further study.

Chapter six makes use of the standardization and taxonomic analysis performed in chapters three and four. Soil Taxonomy and ASC together creates a cluttered taxonomic space. There are taxa within each soil system that are so close in properties and attributes as to be considered near duplicates. The duplication is related to the operational levels of the tiers of each taxonomy. Operational levels are defined by the mean nearest neighbour distance, calculated from a distance matrix of the combined data set. This concept requires the development of a new method, nearest neighbour analysis. The tiers of ST are matched by nearest neighbour analysis to the closest tier of the ASC. These Operational levels are harmonized and a culling algorithm devised to remove duplicate soil taxa, then the two taxonomies are combined and assessed. This is a vital step in the creation of a combined taxonomy and nomenclature.

Chapter seven brings all the elements of the previous chapters together. ST has been assessed, compared and matched with the ASC. This combined data has a more even distribution of taxa than either ST or ASC. Using a Ward dendrogram, the combined data set is split into groups. More mathematical analysis determines the properties responsible for the splitting of the groups, which are
then combined into a numerical taxonomic description scheme. The groups are assessed and it is noted that the majority of soil taxonomic units tend to converge in groups that are considered to be taxonomically similar. Some emergent soil taxonomic allocations emerge. This is evidence of a property that has a higher priority than is recognized in the taxonomic systems. The role of Andisols in the ASC is discussed as data pertaining to these soils is incongruous, the same issues may be reallocating taxa from the Spodosol group. The creation of bland designations that successfully identify tiers, yet are not easy for a human operator to distinguish is noted. The role of a functional nomenclature is therefore discussed.

Finally, in chapter eight it is concluded that it is possible to harmonize soil information and use this to create a taxonomic scheme, which has less potential for conflict. Future directions are highlighted. In this instance there are taxa in the ASC that need to be reassessed and re-introduced to the data scheme. This may reallocate some taxa into more homogeneous groups within the combined classification, making a more consistent taxonomic scheme.
Acknowledgements

This thesis is the culmination of not only my effort but the effort and patience of a cloud of people that have surrounded me. From an academic perspective, I have had nothing but the highest quality help from my supervisors. Alex McBratney and Budiman Minasny are people of intense intellect, but extremely different in their approaches. Alex has this paradox in his style, to demand a very specific outcome, yet gives the freedom to do it any way you see fit. Budi by contrast provides very sound advice while uttering no sound at all. Both these people have been able to extract some remarkable work from me, I have no idea how I would have written this thesis without them. Much of my work has had international help. I thank Erika Micheli, Jon Hempel and Vince Lang for their support and data.

Even though they were not technically supervisors, there are many people in the department who I have mined wisdom from on sometimes a daily basis. Many of the methodologies I used were learned from Brendan Malone. He assisted me in my honours year and when the chips were down, he helped me again in my PhD. I feel truly honoured to have had the assistance of a man of his calibre for both my theses. Similarly there are lecturers such as Balwant Singh, Stephen Cattle, Brett Whelan and Damien Field who would on occasion help me with whatever I was struggling with without rolling their eyes (too much). Other lecturers and generally people with more important things to do than listen to me would regularly sit with me and help me through whatever I was dealing with. Willem Vervoort, Floris Van Otrop, Inakwu Odeh and Thomas Bishop. I thank you from the bottom of my heart. A special thank you to Uta Stockmann, who provided vital assistance to a key part of my thesis-explaining to me exactly what Matroshka dolls are. Kanika Singh, your encouragement means a lot to me.

The environment in which I wrote my thesis has to be seen to be believed. There is a diverse collection of students who are in their own way extremely smart and outrageously funny. I started my thesis with Sebastian Campbell, his help and friendship was invaluable. When he moved on I was given assistance on various projects by Jose Paderion, and there has been a steady stream of capable people who have been able to work with me and provide support. Ed Jones, you are a legend. Thank you Mario Fajardo. Patrick, I’m sure you are useful for something. Jason Morris, your poster of procrastination haunts me. All the other level-one-ites, you know who you are. It has been a privilege working with you.
On the home front there is a family who have been encouraging me over the last few years. They don’t necessarily understand what I do, but they have learned to nod at the correct times when I talk. Mum and Dad, I’m glad you lived to see this day. My brother and sister, John and Emma (Eidel), thank you for being you, and just so you know, I have a PhD now. I am therefore correct in any argument. Many thanks to my parents-in-law, Helen and Trevor. They spent a lot of time sifting through the debris of the house, looking after the kids while I was off soil taxonomy-ing. A big thank you to my Aunt Jan. She selflessly offered her editorial help despite her lack of soil knowledge, I took her up on this offer and as a result, I’m sure she will have a twitch every time someone mentions soil for decades to come. What she lacked in technical expertise she made up for with grammatical ability. It somehow felt wrong to ask her advice on how to word the acknowledgements, so she is stuck with this for thanks.

It has been said that it takes a village to raise a child. Vikings would say that it takes a child to raze a village. In the case of my family, it’s a little bit of both. My two kids have been a paragon of patience, and they are of an age where patience is hard to come by. They love me and I love them. Emma Joy, you have no idea how much light you brought into my life when you arrived. Please know that I am only human and am imperfect. I get angry, I sometimes don’t know what to say, and as you get older, it is a father’s responsibility to be as embarrassing as possible. Please have the foresight to see through all this, to a man who cares deeply for you and would sacrifice anything to make sure you were ok. Olivia, you are the second child which has benefits and pitfalls. Please know I love you dearly, that it makes me the happiest man on the planet to be your daddy, even when I am coaxing you off the roof or sitting with you in the casualty ward on Christmas Eve.

Finally to my wife Vanessa. This thesis has been a true test of character. Thank you for putting up with me. Thank you for the editing help, thank you for the sleepless nights. I can definitely say that our relationship is truly unique, you have been a great friend and it has been a fun ride so far. Here’s to the next big project and the upheaval that brings.

If I have forgotten someone, well… That’s typical isn’t it? The academy music is playing now, the compare is moving up to me so it’s time to leave the stage. Try the fish. It’s delicious.
Beautiful

Let her go, can she stay?
No. Slipping away in the undertow,
It’s not about you or me, it was unbearable, it was holy,
So don’t you tell me what to say.

Cause every look was a smile and every moment was balm to me,
And every word was a song, and every minute was beautiful.
And every look was a smile and every moment was balm to me,
And every word was a prayer and every image is beautiful.

Will of iron, heart of gold,
Such power love, grace, never cold. Great in mind,
I was never too tall for her to sing to me:
“Whose little boy are you?”
So don’t you tell me how I feel.

Cause every look, beautiful, and every moment was balm to me,
And every word, was a symphony and every image is beautiful.
And every look was a smile and every moment beautiful and every word beautiful,
And every image….

So don’t you tell me how I’m feeling and don’t you tell me what to say,
Cause every image is beautiful.

(In loving memory of Doreen Whitefield, 1918 - 2013)
Chapters of this thesis have been submitted or published in scientific journals

Chapter 1


Chapter 2

Hughes, P.A., McBratney, A.B., Minasny, B., and Hempel, J., Creation of soil surface horizon classes from the USDA soil characterization database. Submitted to Geoderma regional.
1.2.1.3 Australian Soil Classification system
1.2.1.4 New Zealand Soil Classification system
1.2.2 Issues with current taxonomies and definition of soil
  1.2.2.1 Soil forming factors
  1.2.2.2 Issues of unification between taxonomies
1.2.3 Interpreting between soil taxonomic systems
1.2.4 Creating a taxonomy
1.3 Soil data and its implications in classification
  1.3.1 Mathematical methods and numeric classification
1.4 Conclusions
1.5 References

Chapter 2

End members, end points and extragrades in numerical soil classification

Summary
2.1 Introduction
2.2 Algorithms
  2.2.1 Fuzzy-k-means
  2.2.2 Fuzzy-k means with extragrades
2.2.3 Akromeson
  2.2.3.1 Identification of end points (Akro)
  2.2.3.2 Creation of individual end point centroids (Meson)
  2.2.3.3 Nomenclature
2.3 The problem with extragrades and a possible solution
2.4 Application in the Edgeroi area
2.5 Results
  2.5.1 Weighting factor (w)
2.6 Discussion
2.6.1 End point positions
2.6.2 Cluster types
  2.6.2.1 The weighting factor in akromeson
  2.6.2.2 The overall deviation of all the extragrades
2.6.3 Significance
2.7 Conclusions
2.8 Acknowledgments
2.9 References

Chapter 3

Creation of soil surface horizon classes from the USDA soil characterization database

Summary

3.1 Introduction
  3.1.1 Development of soil classification
  3.1.2 The soil horizon
  3.1.3 Previous work

3.2 Method
  3.2.1 Developing a soil horizon database
  3.2.2 Soil properties
  3.2.3 Filling in the gaps
  3.2.4 Algorithm

3.3 Results
  3.3.1 Horizon classes and material classes created
  3.3.2 Comparison between material classes and horizon classes
  3.3.3 Using dendrograms
  3.3.4 Selection of data set
  3.3.5 Geographic representation

3.4 Discussion
3.4.1 What are the soil classes? 76
3.4.2 Central classes 77
3.4.3 Kappa 78
3.4.4 End point classes 79
3.4.5 End point class “G” 81
3.4.6 Comparison with ochric horizons from Soil Taxonomy 81
3.4.7 Ochric horizons 82
3.4.8 Differences between systems 82
3.4.9 Identifying “ochric like” horizons 82
3.5 Conclusions 83
3.5.1 Numerical analysis 83
3.5.2 Class selection 83
3.5.3 Probability map created 84
3.5.4 Classes used for assessment of environmental damage 84
3.5.5 Colour as a defining characteristic 84
3.6 Further work 84
3.6.1 Applying weightings 84
3.6.2 Analysing sub surface horizons 85
3.6.3 Creating numerical pedons 85
3.7 Acknowledgments 86
3.8 References 86

Chapter 4 91

Comparative taxonomy between the USDA Soil Taxonomy and the Australian Soil Classification System 91

I: Data harmonization, calculation of taxonomic distance and inter-taxa variance 91

Summary 91

4.1. Introduction 92
4.1.1 Comparison of differing soil classification schemes 92
   4.1.1.1 The soil profile description 92
   4.1.1.2 Soil profile description’s around the world 92
   4.1.1.3 Soil Taxonomy and the Australian Soil Classification 93
4.1.2 Development of a universal soil classification system 95
   4.1.2.1 On the application of taxonomic distance calculations in soil classification 95
4.1.3 Objectives 96

4.2 Methods 97

4.2.1 The Soil Taxonomy Great Group centroids 97
   4.2.1.1 USDA data 97
   4.2.1.2 Properties 97
   4.2.1.3 Database structure 97
   4.2.1.4 Depth function using splines 98
4.2.2 Australian Soil Classification System Great Group centroids 99
   4.2.2.1 The database 99
   4.2.2.2 Extracting properties 99
4.2.3 Properties that require processing 100
   4.2.3.1 Processing texture 100
   4.2.3.2 Processing colour 101
   4.2.3.3 Processing Evidence of water and ice 101
   4.2.3.4 Processing bulk density 101
   4.2.3.5 Depth-function harmonization 103
   4.2.3.6 Calculation of centroids 104
   4.2.3.7 Direct profile comparison 106
   4.2.3.8 Principal component analysis and Euclidean distances 113
   4.2.3.9 Missing data 114
   4.2.3.10 Density analysis 114

4.3 Results of Comparative Taxonomy 116

   4.3.1 Centroid comparison 116
4.3.1.1 Distance between and mapping of centroids in the two systems 116
4.3.1.2 Soil Taxonomy in Principle Component space 117
4.3.1.3 Australian soil classification Principle Component Analysis plot 118

4.3.2 Taxonomic precision and accuracy of Soil Taxonomy and the Australian Classification System 121
4.3.2.1 Inter taxa variation 121
4.3.2.2 Adding Australian Soil Classification System Great Group centroids to Soil Taxonomy 124
4.3.3 Principal components and soil properties 125
4.3.3.1 Soil properties in relation to depth 134

4.4 Discussion 135
4.4.1 Data size 135
4.4.2 Comparing taxonomies 135
4.4.3 Taxonomic accuracy 136
4.4.4 Important variables 136

4.5 Conclusions 137

4.6 Further work 137
4.6.1 Visualization 138
4.6.2 Properties in relation to differing taxonomies 138
4.6.3 Creating comparable groups 138

4.7 References 139

Chapter 5 143

Comparative taxonomy between the USDA Soil Taxonomy and the Australian Soil Classification System 143

II: Comparison of Order, suborder and Great Group taxa for Soil Taxonomy and Australian Soil Classification 143

Summary 143

5.1. Introduction 144
5.1.1 Grouping taxa numerically 145
5.1.1.1 Intra Distance analysis 145
5.1.1.2 Inter Distance analysis 146
5.1.1.3 Intra hull analysis 146
5.1.1.4 Inter hull analysis 147

5.2 Aims 148

5.3. Method 148

5.3.1 Soil Profile Description centroids 148
5.3.2 Nearest neighbour distances 149
5.3.3 Convex hulls 149
  5.3.3.1 Hull sizes to taxonomic level 150
  5.3.3.2 Estimation 150
  5.3.3.3 Overlap 150

5.4. Results 150

5.4.1 All Soil Profile Descriptions 150
5.4.2 Nearest neighbour distances 151
  5.4.2.1 Nearest Neighbour Distance (NND) tables 152
  5.4.2.2 Nearest neighbour distances Australian Soil Classification to Soil Taxonomy 153
  5.4.2.3 Nearest neighbour distances Soil Taxonomy to Australian Soil Classification 154

5.4.3 Hulls 155
  5.4.3.1 Hull overlay analysis 155
  5.4.3.2 Hull Tables 155
  5.4.3.3 Similar Orders 157
  5.4.3.4 Different Orders 158
  5.4.3.5 Similar but different Orders 158

5.4.4 Concentricity 159

5.5 Discussion 160

5.5.1 Hulls 160
5.5.2 Hydrosols and Aridisols 161
5.5.3 Andisols in Australia 162
Chapter 7

Creation and Nomenclature for a New Classification System

Summary

7.1 Introduction and previous work

7.1.1 Creation and use of a combined database

7.1.2 Taxonomic keys and computer aided classification

7.1.3 Distance comparison

7.1.4 Dendrograms and cladistics

7.1.4.1 Studying relationships between tiers and individual units of a taxonomy

7.1.4.2 Studying relationships between taxonomic systems

7.1.4.3 Understanding the tier structure of taxonomic systems

7.1.5 Taxonomic designations

7.1.5.1 Naming convention

7.2 Aims

7.3 Methods

7.3.1 Ward dendrogram of Principle Component Analysis

7.3.2 Finding groups and determining property clusters

7.3.3 Determining the property differences between groups

7.3.4 Systematic name

7.3.5 Link back to original taxonomy

7.3.6 Depth

7.4 Results

7.4.1 Dendrogram

7.4.2 Properties that separate groups

7.4.2.1 G1 and G2

7.4.2.2 G21 and G22

7.4.2.3 G211 and G212

7.4.2.4 G2121 and G2122

XXI
7.4.2.5 G21211 and G21212  205
7.4.2.6 G212121 and G212122  206
7.4.2.7 GX and GY  207
7.4.2.8 GY1 and GY2  207
7.4.2.9 GY2X  207
7.4.2.10 GY21 and GY22  207
7.4.3 Creating a classification  208
7.4.4 Organizing the data  210
  7.4.4.1 The taxonomic key  211
7.5 Discussion  213
  7.5.1 Grouping  213
  7.5.2 Groups and pedological/property processes  214
  7.5.3 Allocation issues  215
  7.5.4 Andisols  215
  7.5.5 Spodosols  216
  7.5.6 Mapping other taxonomies  216
  7.5.7 Reduction in conflict  217
7.6 Conclusion  217
  7.6.1 Future work  217
    7.6.1.1 Addition of other taxonomies  217
    7.6.1.2 Fixing Andisols  218
    7.6.1.3 Fixing Spodosols/Podosols  218
7.7 References  219

Chapter 8  221

8.1 Conclusions  221
8.2 Future directions  223
  8.2.1 Andisols  224
  8.2.2 Spodosols/Podosols  224
8.3 Closing Remarks
List of Figures

Figure 1. How the akromeson algorithm finds points and sets them as centroids for clustering. 33
Figure 2. Notional data set from Table 1 displayed graphically. 40
Figure 3. Fuzzy K Means applied to Figure 1 data. 43
Figure 4. Fuzzy K Means with Extragrades applied to Figure 1 data. 45
Figure 5. Dataset from Figure 1 corrected using the akromeson algorithm. 46
Figure 6. Resolution of clusters in principle component space using Fuzzy K Means with extragrades, alpha set at 1/(k+1). 49
Figure 7. Resolution of clusters in principal component space using akromeson, w=60. 51
Figure 8. Resolution of clusters in principal component space using akromeson, w = 30. 52
Figure 9. Comparison of weight factor distribution within clusters. 54
Figure 10. A dendrogram demonstrating the priority set in the cluster analysis. 73
Figure 11. Soil Classes plotted on a texture triangle. 74
Figure 12. Density plot of soil data based on a 600km window. 76
Figure 13. Probable distributions of surface horizon classes. 78
Figure 14. Probability plot for the soil horizon class Kappa. 79
Figure 15. Probable locations of soil surface end point classes. 80
Figure 16. Depth function harmonisation. 104
Figure 17. Flow diagram demonstrating the pre-processing of data prior to PCA analysis. 107
Figure 18. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. 108
Figure 19. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. 109
Figure 20. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. 110
Figure 21. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. 111
Figure 22. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. 112
Figure 23. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines.

Figure 24. Density plot for carbon comparing Soil Taxonomy and Australian Soil Classification.

Figure 25. Density plot for Soil Taxonomy and Australian Soil Classification demonstrating a systematic error in the clay fraction of the Australian dataset.

Figure 26. Convex Hulls of every Order in Soil Taxonomy.

Figure 27. Convex Hulls of every Order in the Australian Classification System.

Figure 28. Data dispersion of Australian Soil Classification versus Soil Taxonomy.

Figure 29. Soil Taxonomy Soil Taxonomic Allocations plotted against Australian Soil Classification Soil Taxonomic Allocations in 3d PCA space.

Figure 30. Cumulative variance of gravel, clay, silt, sand, red and green on Principal Component one calculated from rotational values.

Figure 31. Cumulative variance of blue, water, ice, bulk density, organic carbon and carbonates on Principal Component one calculated from rotational values.

Figure 32. Cumulative variance of pH, CEC, Ca, Mg, Na and K on Principal Component one calculated from rotational values.

Figure 33. Cumulative variance of acid saturation, base saturation, Exchangeable Sodium Percentage, Electrical Conductivity and gypsum on Principal Component one calculated from rotational values.

Figure 34. Cumulative variance of gravel, clay, silt, sand, red and green on Principal Component two calculated from rotational values.

Figure 35. Cumulative variance of blue, water, ice, bulk density, organic carbon and carbonates on Principal Component two calculated from rotational values.

Figure 36. Cumulative variance of pH, CEC, Ca, Mg, Na and K on Principal Component two calculated from rotational values.

Figure 37. Cumulative variance of acid saturation, base saturation, Exchangeable Sodium Percentage, Electrical Conductivity and gypsum on Principal Component two calculated from rotational values.

Figure 38. All Australian SPDs overlaid with US centroid data in 2D (left) and 3D (right).

Figure 39. Mean nearest neighbour distances.

Figure 40. Hull areas according to taxonomy natural (left) and logged (right).

Figure 41. Overlap of convex hulls.
Figure 42. Overlaid convex hulls. 157

Figure 43. Overlaid convex hulls. 158

Figure 44. Positions of convex hulls of Orders over their native taxonomic systems in PC space. 159

Figure 45. Matroscka plot. 160

Figure 46. Convex hull plots of two Orders overlaid. 173

Figure 47. Mean nearest neighbor distances according to classification system. 173

Figure 48. Comparison of thinned data from ST and ASC with un-thinned data from both taxonomic systems. 178

Figure 49. The biological tree of life. 192

Figure 50. Demonstration of nomenclature of groups prior to group names. 199

Figure 52. Dendrogram groupings. 203
List of Tables

Table 1. A notional data set which is plotted in figure 1.

Table 2. Centroids calculated from Fuzzy K Means (FKM) applied to data from Figure 1 and plotted in Figure 3.

Table 3. Memberships to the clusters in Table 2 from Fuzzy K Means as applied to data from Figure 1.

Table 4. Positional data pertaining to centroids calculated using Fuzzy K Means with Extragrades

Table 5. Memberships from FKM with extragrades applied to Figure 1 data.

Table 6. Properties of Edgeroi centroids as portioned by Fuzzy K Means with extragrades.

Table 7. Initial centroids that were input into the akromeson algorithm.

Table 8. Cluster centroids as determined by akromeson algorithm.

Table 9. Centroid table for AM 5,6 with a w of 30.

Table 10. Central material classes organised by sand fraction, then clay fraction.

Table 11. End point material classes sorted by sand fraction, followed by clay fraction.

Table 12. Central horizon classes organized by sand and clay fraction.

Table 13. End point horizon classes organized by sand and clay fraction.

Table 14. Example of the layout of one of the properties-fragments % for vertosols.

Table 15. Example of how all the properties are concatenated into a single vector for each taxa profile description.

Table 16. Soil Taxonomy Orders and their variance.

Table 17. Australian Soil Classification Orders and their variance.

Table 18. Cumulative variance calculated from rotational values of principal components.

Table 19. Average nearest neighbour distances between orders of the ASC and ST.

Table 20. Average next nearest neighbour distances between Orders of ST and the ASC.

Table 21. Hull overlap areas between Orders of ST and Orders of ASC.

Table 22. Priorities of properties in reference to tier level of ST and ASC.

Table 23. List of soil taxa from both ASC and ST which are a minimum of 5 STDU apart. ASC taxa are coded but ST have full names provided.

Table 24. First table of property means.

Table 25. Second table of property means.
List of Abbreviations

AM    Acromeson
ASC   Australian Soil Classification
BD    Bulk Density
CEC   Cation Exchange Capacity
CSIRO Commonwealth Scientific and Industrial Research Organisation
EC    Electrical Conductivity
FAO   Food and Agriculture Organization of the United Nations
FKM   Fuzzy K Means
FKME  Fuzzy K Means with Extragrades
FPI   Fuzziness Performance Index
IUPAC International Union of Pure and Applied Chemists
IUSS  International Union of Soil Services
MaxCls Maximum cluster partition
MIR   Mid InfraRed
MPE   Modified Partition Entropy
NIR   Near InfraRed
NND   Next Nearest neighbour Distance
NRCS  Natural Resource Conservation Service
NSW   New South Wales (Australia)
OC    Organic Carbon
OM    Organic Matter
PCA   Principal Component Analysis
SPD   Soil Profile Description
ST    Soil Taxonomy
STA   Soil Taxonomic Allocation
STDU  Soil Taxonomic Distance Units
TERN  Terrestrial and Ecological Research Network
US    United States
USA   United States of America
USDA  United States Department of Agriculture
VIR   Visible InfraRed
w     Weighting factor
WRB   World Reference Base
XRD   X-Ray Diffraction
XRF   X-Ray Fluorescence
Introduction

A soil taxonomic system is an anthropological construct which attempts to create understanding regarding the variability of soils around the world, which are a complex component of ecosystems based on climate, temporal and geological factors. Learning taxonomy has sometimes been considered a chore, and writing papers for international journals in which soil types need to be translated is an activity sometimes loathed by those who are writing for differing disciplines such as agronomy or hydrology and are not specifically concerned with the soil type the study is based upon. The reason for this negative perception is largely due to the difficulty in understanding and successfully conveying meaning between the many disconnected classifications in existence. There are two main internationally recognized systems, World Reference Base (WRB) and Soil Taxonomy (ST), as well as numerous smaller regional systems. The terminology of any one of these classification systems can seem disjointed, confusing or misleading to a user versed in a different taxonomic system. There is a philosophical conflict in the literature, with experts advocating one system over another (Haggard and Weindorf, 2012). Because of this, a unified soil classification has been a challenge that could be deemed too difficult to surmount. The real issue is not that soil classification is siloed or conflicted, but that there is a lack of communication between systems. With new mathematical methods, the issues separating the taxonomies of the world needn’t be an impediment, harmonization methods are possible. Additionally, the individual taxonomies should not be viewed as competitors in the ring of soil science, but more accurately as players in a combined scheme that is bigger than any single classification system. The purpose of this body of work is primarily to assess two taxonomic systems; the Australian Soil Classification (ASC) and the US based ST. This thesis also looks at methods of unifying all the taxonomies of the world, using these two taxonomic systems as a test bed for harmonization techniques.

World taxonomies

The taxonomic systems of the world are numerous and have been developed to fulfill a specific taxonomic function. ST is, along with WRB, considered to be a global system for which regional
classifications should be translated for the purpose of international research. ST is a hierarchical system that utilizes pedological processes in its classification, and this approach has advantages and disadvantages (Brevik et al., 2016; Mazhitova, 2008; Weindorf et al., 2009; KrasilNikov et al., 2009). WRB’s methodology differs in that it concentrates on properties specifically while ST uses the properties as an indicator of processes. The resulting two taxonomies are remarkably similar, yet it is difficult to find coherence between the two. Using some numerical methods and the ASC as a neutral method of comparison, it may be possible to create understanding between these systems.

**Mathematical methods**

Soil taxonomies use a few specific properties or processes to create individual taxonomic designations. There has, however, been a huge amount of amassed data that has been collected. This means that each Soil profile description (SPD) which has already been classified to a particular taxon probably has much more data associated than the two or three attributes required, such as texture and base saturation in the case of Mollisols, for its taxonomic allocation. This extra data can be collected and collated in a larger data pool in which several other ancillary properties can be linked. Extra properties can therefore be inferred in these soil taxonomic allocations (STA). Mathematical techniques such as fuzzy k means (FKM) can be adapted to create better understanding of a taxonomic systems horizons(McBratney and DeGruitjer, 1992). Other methods may be able to go a step further for example; spline functions (Bishop et al., 1999) and Principal Component Analysis (PCA) could assist in the standardization of properties and depths.

**Making taxonomies talk**

If data pertaining to horizons can be standardized and compared, then there is no reason why a more coherent comparison between taxonomies cannot be made. Understanding the similarities, differences, strengths and weaknesses of every taxonomy in relation to every other one is a significant step in the creation of a world taxonomy and the understanding of all the worlds’ soil properties and processes.
World taxonomy
Methods such as Ward’s minimum variance and dendrograms (Murtagh and Legendre, 2014) may be tested for comparing taxonomies and placing each individual taxonomic allocation in a defined part of a world scheme. The properties that associate or separate could be established making an overarching taxonomic key by which the soils of any nation or system could be allocated.

Aims of the Thesis
The aims of this thesis are to:

- Create new methods and algorithms for the analysis of soil profiles and horizons.
- Harmonize data between two differing taxonomic systems to enable comparisons between them.
- Use the harmonized data to determine similarities, differences, strengths and weaknesses between two taxonomic systems.
- Use harmonized data to determine common points of two differing taxonomies that can be used to create appropriate groupings.
- Combine harmonized data in such a way that there is good coverage between the two taxonomies but the data set has little or no duplication.
- Use mathematical methods to link taxonomic designations from two different classification systems and create common groups.
- Identify the processes and properties that distinguish between groups identified.
- Create from harmonised data and subsequent groupings a workable taxonomy and key.
References


Chapter 1

Literature review: An overview of previous studies regarding the synchronization of soil taxonomic systems around the world and numerical analysis of soil for the creation of a world taxonomic system.

1.1 Introduction

Perhaps surprisingly, soil is one of the most important components of life, especially human life on the earth. It underpins the supply and use of water, construction and food production, is sensitive to global changes and its stability and utility is the lynchpin for the success and failure of civilizations (Olson, 1981). Recently, this nexus of human activity which is produced by soil has been recognized. More importantly, the fact that soil is not eternal and degradation can lead to serious world problems has also been recognized (Koch et al., 2013). This understanding has led to various bodies around the world coining and promoting the concept of “Soil Security” (McBratney et al., 2014). Over the next few decades, soil security and the issues surrounding it will become more prominent in world strategies concerning economics and the supply of food and water. Despite its emerging importance, the study of soil is poorly understood. Part of this deficit in understanding is due to a lack of synchronization between soil taxonomic systems around the world (Hempel et al., 2013). The reasons for this are complex and varied. The impetuses behind the gaps and miscommunication between world systems require some explanation as well, as societies have not always been unaware of soil related issues. The circumstances that transpired to reduce soil based knowledge in the community need to be understood, as it is often quoted: “Those who cannot remember the past are condemned to repeat it” (Santayana, 1905). It is somewhat appropriate to begin with a general explanation of what soils are (Simonson, 1957), then move to a brief historical overview before addressing some of the modern intricacies of soil classification systems and taxonomies.
1.1.1 Soil definition and divergence

Soil is defined by the second edition of Soil Taxonomy (1999) as:

“A natural body comprised of solids (minerals and organic matter), liquid, and gases that occurs on the land surface, occupies space, and is characterized by one or both of the following: horizons, or layers, that are distinguishable from the initial material as a result of additions, losses, transfers, and transformations of energy and matter or the ability to support rooted plants in a natural environment.”

From this relatively simple definition, the nature of the world’s soils begins to emerge. It is obvious that the world is not a uniform mass, and that the biota (both macro and micro) and mineral compositions of regions change. Where there is difference, there are ecological niches, and as a result soils vary according to patterns in geology, climate and time. Defining individual soil types is therefore as complex as the communities that arise from them. Solids, minerals and organic matter create diversity in chemical composition (Gupta and Germida, 1988), while liquid and gases impart physical properties and limitations on organisms that reside within. Horizons (Kew, 2016) or layers add further vertical and lateral dimensions which can be limited by a confining boundary, which subsequently affects the processes by which the organisms function. Material additions, losses and changes are facilitated by time and partially by temperature. This has a flow-on effect, especially with the evolution of microorganisms within the horizons. It can be seen, therefore, that this relatively simple definition can result in a multitude of differing combinations of materials, spaces, fluids, energy and time. Understanding this flux of properties is essential when determining soil capabilities and will become more important as population increases and climate changes. Understanding of soil has always been ingrained in human culture, so it is more of a case of reclaiming understanding rather than learning anew.

1.1.2 Soil in antiquity

The perception of soil complexity and of soils as individuals with defining characteristics has been understood since prehistory (Brevik and Hartemink, 2010). This ancient understanding stems from the fact that for millennia, human survival has been based on agriculture, and because agriculture is dependent on soil, survival to some extent was dependent on understanding soil. Methods of
describing soils have therefore emerged from antiquity (McNeill and Winiwarter, 2004). Historical texts from the past indicate that the lives of people were intertwined with the health of the soil. Religious metaphors were based upon soil and agriculture, such as the Christian parable of the sower (Matthew 13), describing as a metonymy of religious truth, four different types of soil - a path (compacted soil), rocky ground, weed strewn soil and “good soil”. These four soil types did not require further explanation when the story was first disseminated; images such as this had a resonance that does not exist today. The concepts of seasons and even contemporary festivals typically have their philosophical roots in the soil (Williams, 2004). As humanity progressed and cities absorbed the bulk of human population, its link with the earth was lost, and it has only been in the last two centuries that science has moved to reclaim concepts that for such a long time had been second nature to the people that relied on them. The definitions of soils that have subsequently emerged have diverged owing to location, communication and time. With the shift to globalization, and emerging pressures on the soil resource, it is necessary to mend the fractures. Creating a useful soil language, i.e. a universal soil classification is therefore imperative, but not easy. This is because of the very nature of soil classification (Sanchez et al., 2003).

1.2 Soil classification

The basis of soil classification is the soil horizon (Bridges, 1997). This is a layer of soil that can be differentiated from other layers by various physical and chemical properties such as changes in texture, colour or mineral inclusions. Once a horizon is identified, its depth and properties can be described by a series of tests. A soil profile description is composed of all component horizons. The number of horizons and their properties are usually recorded from the surface of the soil to the lowest practical depth or its termination by a confining layer, or lithic contact (Kew, 2016). Multiple soil descriptions that share important diagnostic properties, and have similar behaviours and functions can be grouped together by criteria such as chemical parameters, mineralogy, etc. to form classes or taxa. A series of soil profile descriptions are linked, typically using expert opinion, from these, a typical soil profile is chosen from the set of descriptions allocated to the class as being “representative” of the group. These are compiled with other typical soil profiles - a soil classification system. Navigating
this system is usually done with a taxonomic key. This is essentially a binomial decision tree by which the inclusion or exclusion of certain diagnostic properties can be used to determine the soil type. Although the world’s soils can be rightly considered as vastly different, at the same time they can also be considered as an unbroken skin, which stretches over the entire face of the earth. This view conceptualizes soil as a continuum (McBratney et al., 2003; Nachtergaele, 2000). Whether viewed as vastly different or as part of a continuum, soils have needed to be analysed. Scientists have utilised taxonomies for this purpose to direct research and organise results. Taxonomies are useful because they provide conceptual models for extraordinarily complex processes (Bockheim and Gennadiiyev, 2000). Mapping and classification of soils has, is, and will be an incremental exercise which creates advances which inevitably fuel other advances. An incremental increase in knowledge and systematic application has been observed in the soil classification systems that have emerged all over the world over the last century (Brevik et al., 2016). However, there are now many taxonomic systems that exist, and to a certain extent, compete in a global data space (Shi et al., 2006) with varying levels of success (Habarurema et al., 1997). The challenge facing contemporary soil scientists is similar to that faced in traditional soil management (Barrera-Bassols and Zinck, 2003) and revolves around the transference of information. In more recent times, the scope of this challenge has increased to a global scale.

1.2.1 Soil classification systems of the world

As already alluded to, soil has more than one function. This review focuses on soil classification in regard to agricultural systems, rather than schemes which are primarily concerned with mining, geology, archaeology and engineering such as the unified soil classification system (Stevens, 1982) – although, of course, there will be some crossover. It should also be noted that there are various terms used to describe the levels of a hierarchical taxonomy. Most commonly the main three levels are named “Order”, “Suborder” and “Great Group”, but this is by no means a definite convention. For this review, these levels will be referred to as tiers. The main period when soil classification schemes began to emerge was at the turn of the previous century (Simonson 1968; Bockheim, 2005). This was a time when communication was difficult - databases could not be shared and mathematical analysis
was done by hand and was, therefore, a tedious process. Documents pertaining to this subject are not readily available. Of the few documents that emerge, examples such as Carmago et al., (1987) are rare and endangered, and it is difficult to access complete copies. More recently, Australia (Isbel, 2016), Austria (Nestroy et al., 2000), Brazil (Embrapa, 2006), Canada (Committee, 1978), China (Gong, 2001), Cuba (Hernandez-Jiminez et al., 1999), France (Agronomique, 2009), Germany (Wittman et al., 1997), Hungary (Micheli et al., 2006; Lang et al., 2013), Japan (Obara et al., 2011), New Zealand (Hewitt, 2010), Russia (Shishov et al., 2001) and Switzerland (Krasilnikov et al., 2009) have all been identified as having developed their own independent classification schemes (Brevik et al., 2016; Krasilnikov et al., 2009), usually as successors to their existing traditional systems. New Zealand and Australia are of considerable interest, being two islands which are in proximity - even sharing geological strata - yet with taxonomic systems that are vastly different. However, it is the World Reference Base (WRB) and Soil Taxonomy (ST) that share the limelight as globally accepted soil classification schemes. It should be noted that there are things to be learned from the contrasts between these four taxonomies. As they were never created to be a universal system, the differences between Australia and New Zealand demonstrate the difficulties in creating a global taxonomy on a case-by-case basis, while differences between Soil Taxonomy and WRB demonstrate a universal clash of methodologies.

1.2.1.1 World Reference Base (WRB)

WRB is a globally accepted taxonomic system (Buol et al., 2011). The goals stated by the various founders of WRB included the creation of “a framework through which ongoing soil classification could be harmonized” (Nachtgerale et al., 2000). In order to do this, agreement on key soil groupings was sought. As the groupings were defined and redefined, an independent soil classification scheme emerged. WRB classifies soils based on the ideas of quantitatively defined diagnostic horizons along with diagnostic properties and materials (Brevik et al., 2016). This system has definitions for 36 diagnostic horizons, 18 diagnostic properties and 17 diagnostic materials. There is an emphasis that these properties and materials should ideally be able to be measured in the field, although in the case of certain soils and materials, such as Andisols, this is not always feasible. Additionally, in the WRB
system, the materials and properties should suggest pedogenesis, although pedogenesis is not in itself a defining criterion of WRB groups. The hierarchy of WRB is different to other taxonomies in that it has two major levels or tiers for its classifications. Considering that the majority of taxonomies in the world have a level of convergence at the third tier (most commonly called Great Groups) and this is a distinctive, possibly exclusive, feature. The initial tier of WRB is typically associated with parent material and climate, using properties and materials (prefix qualifiers) that may suggest a climate regime, but as with pedogenesis, climate itself is not a differentiating criterion (Gray et al., 2011). The rationale behind the use of properties and materials in this way is that properties are more easily measured in the field (Bockheim, 1999; Arnold, 1983). The second tier is related to what are referred to as “secondary soil forming processes”. These secondary soil forming characteristics combined with qualifiers create a soil classification system that is flexible and adaptive, especially when coping with human interference (Brevik et al., 2016; Kabala and Zapart, 2009). The resulting taxonomy has proved itself to be useful in the communication of soil related issues. However, although this system is a success, it has in its own right become “yet another” taxonomy, and the system by nature of the definition of world soils is extraordinarily complex. This means that it is another system that needs to be learned if communication between agencies is to be achieved. The goal of providing neutral ground between world taxonomies is therefore still elusive.

1.2.1.2 USDA Soil Taxonomy

Soil Taxonomy, or ST (Soil survey staff, 1977) is one of the major taxonomic systems that is used the world over. The majority of the literature ascribes its conception to at or around the 1950s but in reality, its foundations are much older, dating back to the conclusion of the American civil war (Smith, 1983). Allegedly a prominent American of the time, Patrick Henry was reported to have said: “Since the achievement of our independence, he is the greatest patriot who stops the most gullies.” (NRCS historical note 1). Soil conservation was, therefore, of utmost importance from the late 1800s, and was officially endorsed with the establishment of the Soil Conservation Service during the dustbowl of the 1930s, and eventually became the Natural Resource Conservation Service (NRCS) in 1994 (Ditzler and Ahrens, 2006). It was at this time that the concept of a specific system such as Soil
Taxonomy arose as an attempt to harmonize two concepts: the concept of zonality, borrowed from Russian taxonomies of the time, and the refinement of great soil groups, a common starting point for many of the world’s taxonomies (Bockheim et al., 2014; Hole and Campbell, 1985). Soil Taxonomy uses specific properties to define soil profiles. These profiles are subset into diagnostic horizons which typically describe specific pedological processes and climate (Anderson and Schaetzl, 2005). This sets it apart from WRB which pays more attention to specific material properties. This system consists of six tiers: Order, Suborder, Great Group, Subgroup, Family and Series with an extra localized designation called “phase” for soil maps. Typically, Orders are established by the properties of a diagnostic horizon, either one of eight surface horizons (epipedon) or a diagnostic subsurface horizon of which there are 12. There are 12 Orders, typically around 10 Suborders per Order, with dozens of specific Great Groups which slot in under the Suborders. The merger of all these Orders, suborders and Great Group designations create a large number of individual Great Group composites. Literature commonly cites the number of Great Groups in existence for ST at around 300 (Grunwald, 2016), with a further 1200 Subgroups and 17000 Series. All of these are approximate figures, but it provides a picture of the detail and complexity involved in creating a taxonomic designation for ST. Soils described to Order level in ST require some exacting measurements e.g. a specific base saturation requirement between horizons in a Mollisol, and the taxonomic key relies on definitions that are scattered throughout the document, making the soil classification scheme itself daunting to the uninitiated. The full version (Soil survey staff, 1999) comprises 860 pages and is highly detailed. The current taxonomic key (Soil survey staff, 2014) – an abridged version of the system is 338 pages long. This level of detail although ponderous, makes meticulous descriptions possible, hence its preferred status in scientific literature.

1.2.1.3 Australian Soil Classification system
One of the more familiar systems to the author is the Australian Soil Classification System (ASC) (Isbell, 2016). This soil classification system evolved from two preceding systems - great soil groups (Stace et al., 1968) and the factual key (Northcote, 1979). The resulting classification was prepared rapidly for a soil conference in 1996. The rapid preparation resulted in a rapid classification scheme
which had some unexpected serendipities, the most obvious being its speed of use. This system is composed of several tiers: Order, Suborder, Great Group, Subgroup and Family. Subgroups are rarely used and Families less so; indeed, they are not explicitly described in the ASC book, but the level of detail is not the major feature of this system. There are fourteen Orders, of which one Order is anthropic, with some 90 or so associated suborders. A soil profile description can usually be determined to the Suborder level with a minimum of laboratory equipment. The major advantage of this system, as already noted, is speed. The simple diagnostic criteria, combined with new spectroscopic techniques and phone apps which assess soil stability (Fajardo et al., 2016), allow a user to rapidly ascertain the order to which a soil profile description belongs. Then other properties of the soil can be ascertained in a matter of minutes - a feat which is difficult to perform with other contemporary taxonomies. With the Australian Soil Classification system a good general purpose idea of the soil can be achieved within a half hour. However, the speed at which classification occurs is probably offset by accuracy. Moreover, other classifications derived from hours or weeks of lab work can provide a user with more accurate results.

1.2.1.4 New Zealand Soil Classification system
The New Zealand Classification (Hewitt, 2010) has a hierarchical set of tiers. These tiers are divided into Orders, Group and Subgroup with the newer Family and Sibling tiers at the end of the classification series. The Orders themselves are divided into three main categories pertaining to soil development, and there are three Orders outside of these which are dominated by anthropic features, some kind of extreme water activity or large accumulations of organic matter. The major distinction that defines New Zealand soils and their classification system is the extreme youth of most of the soils, combined with attention paid to allophane (Parfitt, 1990). Management issues are also greatly divergent, primarily dealing with porous soils and wet climate which results in large nutrient leaching events, the primary concern of government agencies (Di and Cameron, 2002). This combination of factors assists in the superior description of Andisols (which are not classified at all in the ASC) and in the descriptions of emerging soils, which will become more useful as the glaciers retreat from northern Europe and Antarctica. These kinds of features are in stark contrast to a typical Australian
soil, despite the relative close proximity of the two nations, highlighting in a regional way the reason why these two soil classification systems developed in the way that they did, as adopting the Australian classification for these conditions would have proved futile.

### 1.2.2 Issues with current taxonomies and definition of soil

The major issues with most of the classification systems in existence are that they are either:

- not broad enough to be used internationally or
- not universally adopted or understood by all countries/organizations.

Also; legacy descriptions and some current descriptions do not account for extreme weathering or anthropic activity, which degrades some sand bearing soil classes, creating homogeneous horizons, making previous taxonomic allocations such as Podosols (Spodosols in ST) inaccurate. There are those who would create a case for sandy soil classifications (Harper and Gilkes, 1994), but such classifications are problematic for definitions which require pedological development (Bilzi and Ciolkosz, 1977).

Many other minor technical impediments exist in the creation of global standards, such as laboratory methods and definitions. The definition of silt in the United States, for example, requires a different particle size than the definition for silt in Australia (Padarian et al, 2012). This and other examples seem insignificant but can translate to massive systematic data inconsistencies.

### 1.2.2.1 Soil forming factors

Many soil classifications have been influenced by an emphasis on soil forming factors (Krasilnikov, 2009; Muir, 1962) It is this concept that guides many of the profile descriptions in ST, and many other soil classification systems, resulting in methodologies that have adopted a hierarchical approach (McBratney et al., 2002) such as the presence of natric horizons which are typically the result of eluviation down the profile, but are commonly evidenced by a high sand fraction overlying a clay pan. Classifying the circumstances in which the soil has come into being as much as the physical
properties of the soil itself. This idea was taken a step further with the combination of the quantification of these factors (Jenny, 1994) and refinement with the SCORPAN concept (McBratney et al., 2003), assisted by mathematical methods to view soils as more of a continuum of properties which change over time and space, with this change being guided by physical attributes of the environment. This continuum can be seen in the subtle changes in the pedological organization of sands in Western Australia. There is a distinct classification scheme in Western Australia which has a large emphasis on high sand soils (Schoknecht and Pathan, 2013). This classification is a good example where highly weathered soils which could be considered homogeneous can be differentiated, as they have a host of differing terrain positions, parent materials, weathering intensities and other soil forming factors which make them distinctive, along with a select number of diagnostic horizons. Mathematical methods could be used on these overdeveloped soils because the high content of sand, along with other, more subtle properties could be identified numerically.

1.2.2.2 Issues of unification between taxonomies
In the pursuit of better understanding of world soils, efforts to create a truly global classification system have begun (Hempel et al., 2013), but the challenges of such an enterprise are immense. As a consequence of the many regional taxonomies, a classification scheme for a given region may not be appropriate for another, making communication of concepts between nations sometimes impossible (Shi et al., 2006; Deckers et al., 1998). There has been a push in recent years to adopt either Soil Taxonomy (Sanchez et al., 2003; Yaalon, 1995) or WRB (Krogh and Greve, 1999; Rossiter, 2007) as a global standard. However, problems remain with various examples in the literature which describe failures by either systems, or at least situations which show weaknesses in one system where the other has strengths (Brevik et al., 2016; Mazhitova, 2008; Haggard, 2012; Krasilnikov et al., 2009). The solutions on offer involve the development of a global taxonomy that is understood by everyone and covers all soils. In doing this, a common series of properties and a common method of analysing these properties needs to be created, possibly projecting everything into the same data space. Soil information and utility needs to be simply understood and if possible, be communicated to users who adhere to a vast range of soil classification systems. These are lofty ideals which require a systematic
approach, which is now being attempted by numerous agencies (Micheli et al., 2016; Badia et al., 2013; Lang et al., 2013, Minasny et al., 2008 Golden et al., 2010).

1.2.3 Interpreting between soil taxonomic systems
The difficulty in soil taxonomic understanding is not in the taxonomies themselves, but in communication. Differences in definitions, terminology, experimental techniques and units have the effect of obfuscating commonalities and divergences between systems. Some taxonomies or soil types are specific to localized conditions, which may not exist in broader or differing systems (Niemeijer and Mazzucato, 2003). This means that there are some soil taxa that not only are not directly translatable in other systems, but are completely incomprehensible to them. ASC for example is excellent for describing old soils and soils with texture contrast, and is good with sodic soils (Rees et al., 2010). As well, because of its simplicity, modern prediction techniques can be used successfully in conjunction with the system (Viscarra Rossel and Webster, 2012). The drawback is that it lacks provision for Andisols (Morand, 2013). The New Zealand system has several different allophanic classifications which are perfect for Andisols and Andisol-like soils, but the New Zealand classification is almost completely centred on younger soils (Birkland, 1984). Such a taxonomic system is too narrow to be used elsewhere. There are many systems in existence that are ingenious, but similarly fractured. A number of ideas have been suggested to bridge the gap. Researchers such as Simonson (1959) suggested that simple communicative solutions such as replacing confusing words that are almost exclusively specific to soil science with words or terms that are more commonly used (Bockheim and Gennadiyev, 2000). This is only a partial solution, and it obviously was never adopted. With increased computer power comes the idea of translating soil concepts using mathematical methods (Bouma, 1989), but this also requires a more methodical approach to data collection and storage.

1.2.4 Creating a taxonomy
Other sciences and organizations have managed to adopt world standards (Hempel, 2013). Plant and animal researchers can refer to a universal taxonomic system. Similarly chemists can rely on the
system devised by the International Union of Pure and Applied Chemists (IUPAC) (Muller, 1994). Many soil and agricultural institutions such as the Food and Agriculture Organization of the United Nations (FAO) have recognized that there are world standards for other disciplines, and understand there is a level of disconnection between global soil systems. The solution to date has been to create “another” taxonomic system which overarches everything (Cline, 1963; Nachtergale, 1990, Alavi et al., 2010). This idea is not unique, but it is ambitious. Data compatibility issues and philosophical differences (Haggard and Weindorf, 2012) make the creation of a universal system a treacherous space in which to work. It is possible to create a new world taxonomy, but in doing so, it is necessary to negotiate the many agencies that have a stake in such a venture. It is therefore not so much the creation of a world taxonomy that is a problem (although this in itself is a difficult enough proposition), but its universal adoption. Even adopting parts of a global system, such as recognizing a standard for anthropological changes in soil is a complex exercise (Bryant and Galbraith, 2002). For these reasons, the goal of creating a common standard by which the world adheres has not come to fruition, despite the calibre of people who have invested their time into this goal.

1.3 Soil data and its implications in classification

Today, soil forming factors can be augmented with a multitude of data collection methods that would previously have been unheard of. New equipment that is easier to use on a mass scale or more portable is released every year (Young et al., 2000). New innovations surrounding remote sensing (Jackson et al., 1995; Ben-Dor; 2002, Doolittle, 1987), or other improvements in NIR analysis that could be used remotely under the correct circumstances (Ben-Dor and Banin, 1995) allow the collection of properties such as organic carbon (Henderson et al., 1989) from a distance. Proximal sensing has also seen a rapid rise. Visible, near and mid infra-red (VIR NIR and MIR) can be sensed with relatively inexpensive equipment and methodologies (Viscarra Rossel and Chen, 2011; Viscarra Rossel et al., 2006, Doolittle and Brevik, 2014). The increased importance of spectral data has been internationally recognised, and as a result, various world and national agencies have created a global spectral library by which many materials and properties can be inferred (Viscarra Rossel, 2009).
Other methods such as XRD, x-ray fluorescence (Weindorf and Chakraborty, 2016) and even slaking indication methods (Fajardo et al., 2016) can be employed to add data to the pool.

1.3.1 Mathematical methods and numeric classification
Numerical methods, existing in a data space that can be well above the three easily perceived dimensions, aided by computer analysis is an appropriate avenue to explore for soil classification/harmonization purposes (Bruce et al., 2002). Soils by numbers and digital methods have been around since the fifties (Moore and Russell, 1966). Unfortunately, while innovation was high, the algorithms and equipment with which to perform complex data analysis did not then exist, or was in its infancy. The invention of more powerful computers and statistical languages fuelled the developments of the next half century, creating immense terrestrial data sets and the means by which to study them. These ideas were pioneered by scientists such as Webster and Burrough (1972) and Hole and Hirokana (1960), and improved upon as the technology allowed it. With the increase in data collection, computer power can utilize new mathematical methods which can be employed to aid in soil classification (McBratney et al., 2002). In essence, the computer used in this way is an extension of the human mind, enabling inferences via modelling that were not previously possible. Procedures have been developed to reduce data dimensionality, typically referred to as numerical methods (Carre and Jacobson, 2009; Jenny et al., 1968; Raynar, 1966; Webster, 1977). Pedotransfer functions and splines can estimate gaps in the data collection (Bishop et al., 1999; Nemes 2003.), taxonomic distances can be calculated on variables of equal weight (Hughes et al., 2014; Verheyen et al., 2001; Lang et al., 2013) - a better concept when considering geostatistical and spatial elements of a taxonomy (McBratney and DeGrujiter, 1992; Triantafilis and McBratney, 1993). Clustering algorithms can aggregate soil groups by hitherto unseen connections between soil properties (Harsanyi and Chang, 1994), classifications can be made either hierarchically or non-hierarchically (Webster and Burrough, 1972; Crommelin, 1973; Hartigan and Wong, 1979). Fuzzy sets enable the recognition of intergrading soil properties and formation (Bezdek 1974; De Gruijter and McBratney, 1988; Kruse et al., 2007; Triantafilis and McBratney, 1993), which gives rise to fuzzy k means with extragrades (McBratney and De Gruijter, 1992) - an algorithm which enables the processing of soil
information in the presence of outliers and extragrades, which can then be tuned using the fuzziness performance index (FPI) (Odeh et al., 1990), amongst other classification performance indicators. This idea is improved on with the akromeson algorithm which allows clusters to resolve in data space by creating extragrade clusters that deal with outlying information (Hughes et al., 2014; Keshava and Mustard 2002; Bensaid et al., 1996) and tested on data sets in Edgeroi (Ward, 1999; Triantafilis et al., 2013). With these numerical and mathematical methodologies, research has turned to the issue of communication between taxonomies. Multidisciplinary and multi nation approaches have been sought as the urgency of unpicking the fractured nature of classification and understanding of soil has emerged (Sanchez et al., 2009). Information for the major, and most of the minor world taxonomies exist (Lang et al., 2013; TERN data set, landcare New Zealand data set), and the numeric methods described can be used to standardize soil profile descriptions and taxonomic descriptions for the various systems. Once standardized, these can be compared numerically, and more advanced techniques such as dendrograms and convex hull analysis performed (Xu et al., 1998) for either translation or participation in a larger global system.

1.4 Conclusions

The understanding of soils has been obscured by changes to society and lifestyle, and then has returned to civilization at an amazing pace. Over the last century especially, taxonomies and methods of data collection and analysis have flourished, creating a myriad of systems and technologies. Despite these frenetic advances, there is still no magic bullet of unified soil classification for an increasingly globalized world. The future directions of creating world understanding of soil related issues seem to hinge on either a truly global classification, or a method of translation between the myriad of taxonomic schemes in existence – or possibly a combined approach. There are regional temporal and philosophical factors that need to be addressed when creating a world taxonomic method. The next few years may herald amazing convergences between nations, should the challenges be overcome. This will have impacts on climate change, politics, economics and poverty. This review has described the deficiency in the communication and transference of soil property ideas between taxonomies. The solutions offered up to date have revolved around some kind of translation system
between taxonomies or possibly an overarching world standard by which all soil systems could subscribe. Today, there is a huge amount of data available and methods are coming online that enable standardization between taxonomic systems, so direct comparisons are possible. All of these factors combined could be used to create a world based taxonomy.
1.5 References


Isbell, R. 2016. The Australian Soil Classification, CSIRO Land and water, Canberra.


Österreichischen Bodensystematik 2000, OBS


Webster, R. 1977. Quantitative And Numerical Methods In Soil Classification And Survey. Quantitative And Numerical Methods In Soil Classification And Survey.


Chapter 2

End members, end points and extragrades in numerical soil classification

Summary

Soil classification has progressed with the introduction of computers in the mid-20th century to the point where algorithms can be used to organize soil information into clusters that correspond with soil classes. Algorithms such as fuzzy-k-means perform well, but can be biased by extreme data. Extragrades (data points that fill the extreme data space) were derived to accommodate this problem but estimating the number of extragrades can be challenging and lead to dubious classifications. The idea of end members is discussed and it is concluded that end points, observations that represent the most extreme parts of the soil continuum, are useful in identification of extragrades. This chapter presents and discusses a new clustering algorithm, *akromeson* which identifies extreme points in a given data set and converts them into pseudo clusters, which are then run concurrently with a semi-supervised fuzzy-k means algorithm. A synthetic data set was constructed in order to compare this new method to fuzzy-k means and fuzzy-k means with extragrades. It was able to correctly fix the positions of the centroids, (which was beyond the capacity of fuzzy-k means), and correctly estimated which of the data were genuine extragrades, outperforming fuzzy-k means with extragrades. The performance of this methodology was then evaluated using a data set from the Edgeroi region of New South Wales, Australia. The algorithm identified an extreme cluster on the periphery of the data, and a method was developed to use this new method to routinely find clusters. The ability to efficiently cluster data may provide an added advantage to stakeholders when they are assessing land use practices, especially in regard to areas which exhibit extreme soil properties that require careful management, which this algorithm is capable of detecting.

2.1 Introduction

Numerical soil classification was an idea-rich, data-poor movement which began in the late 1950s, soon after the invention of the digital computer which allowed the ready calculation of taxonomic
distances – distances between points in multi-dimensional spaces, the axes defined by the characters or attributes of soil profiles. This movement was based on a taxonomic principle from Adanson (Moore & Russell, 1966) – classifications should be based on many characters, and each should have equal weight. Similarities between profiles were then converted to distances, either Euclidean, Mahalanobis or of some other metric, which enabled the simplification of multi-dimensional space, a process known as ordination (Hole & Hironaka, 1960; Rayner, 1966; Jenny et al., 1968). For the first time, a view of the soil property universe was revealed. The numerous studies showed in pedological data for most classifications, local areas of higher data density (weak clusters) with areas of smaller density grading in all directions. Data based Classifications were created both hierarchically “inter alia” (Webster & Burrough, 1972) and non-hierarchically (Crommelin & de Grujiter, 1973) using various clustering algorithms, perhaps most notably the k-means algorithm (Hartigan, 1975). Clustering is defined by Kruse et al., (2007) as: “…an unsupervised learning task that aims at decomposing a given set of objects into subgroups or clusters based on similarity…”. The work (summarised in Webster, 1977) succeeded in unraveling the structure of the soil character space, i.e., weak clustering with intergrades, data with loose association within the data space, in all directions of the character space. The work also allowed the creation of multivariate polythetic classes using hierarchical and non-hierarchical schemes during the 1960s. Later, the intergrading nature of the soil population was recognised in a formal way using the concept of fuzzy sets. This allowed the creation of non-hierarchical continuous classes defined by their centroids. This was achieved using the FKM algorithm (Bezdek, 1971). In addition to intergrades (more correctly intragrades), extragrades were also recognised. These are soil individuals that fall outside the main cloud of soil individuals representing the edge of the soil universe. This allowed for more continuous soil mapping, which is more natural when using geostatistical methods (McBratney & de Grujiter, 1992; Triantafilis & McBratney, 1993; Triantafilis et al., 2001), and realized with the fuzzy-k-means with extragrades algorithm (de Grujiter & McBratney, 1988). In the field of geology, the end member concept is used when dealing with the edge of the object universe. In the context of soil texture this would correspond to soil materials with values such as 100% clay, 0% sand, 0% silt. The concept is also used in remote sensing using algorithms to recognize the shape of end members to identify minerals in the variance
of soil surface data. By including several end member signatures in the overall signal, the mineralogy of various soil types can be estimated at low cost (Keshava & Mustard, 2002). It may be useful to introduce another reference for the data at hand, the end point. It is slightly different, therefore, to the idea of the end member which is a theoretical concept. End points have the advantage of being tangible, as they are based on a data point that actually exists. End members are less descriptive as they may not be found in nature at all. This paper will address the question of whether it would be preferable to continue with the extragrade concept, or to move to the idea of end points. A new approach is presented for the first time here based on central intragrades and external grading to these end points. First, the fuzzy-k means with extragrades algorithm was applied, then a new algorithm; akromeson (etymology Greek, literally edge-centre) was introduced that creates classes from the edge of the data field and fuzzy classes in the center of the data. Its performance will be tested with a simple constructed dataset. Secondarily, the algorithm’s ability to handle real data will be evaluated using soil data, which was collected from the Edgeroi district in New South Wales, Australia. Finally, the results will be discussed and any implications or unexpected findings will be highlighted and scrutinized.

2.2 Algorithms

2.2.1 Fuzzy-k-means

An example of the fuzzy-k means clustering algorithm (FKM) is documented in (Carré & Jacobson, 2009; McBratney and de Grujter, 1992). It works by minimizing the distance between the values of all the data in a given cluster and its corresponding centroid, while maximizing the distance between all centroids in the data set. This is done by first allocating the positions of each data point to the defined number of centroids, determining the nearest data points to these centroids and calculating their mean values which then become the new centroids. This process is iterated until the closest possible configuration has been reached, usually determined by the distance each centroid ‘moves’ at each iteration. The “fuzzy” part of FKM is based on the fact that instead of hard partitions (a data point is either in a given class or not) a membership value in the range of 0 to 1 is assigned. This is
advantageous because in the example of soil classification, a profile can partially be from several different classes, allowing for a certain degree of interpretation.

2.2.2 Fuzzy-k means with extragrades
Fuzzy-k means with extragrades (McBratney & de Grujter, 1992) uses the same principle as FKM, but recognizes that some data may confound the clustering process. To correct this issue, an extragrade class (or cluster) is created which includes all the data in the periphery of the soil universe. FKM is iterated with memberships allocated to this extragrade cluster.

2.2.3 Akromeson
As described in the introduction, akromeson attempts to define the end members of the data and fix them as clusters. The akromeson algorithm works in two stages. The first is identification of end points (akro). These are selected as the most extreme data on the convex hull of the data cloud. Once determined, the end points are treated as fixed centroids. The second stage is the creation of additional non-fixed centroids from the data set (meson). These additional centroids are run concurrently with FKM cluster centroids of which the user defines the number of FKM clusters, but not their position within the data space. A simple diagram explaining the algorithms function is presented in Figure 1.
Figure 1. How the akromeson algorithm finds points and sets them as centroids for clustering. The steps are as follows: 1. (top left) Create a convex hull around the data set and calculate the Euclidean distances of all the hull points to the origin. 2. (Bottom left) Find the point of maximum distance to the origin on the convex hull (this will become an end point), eliminate hull points within the circle of radius r around the end point. 3. (Top right) Turn the end points into fixed centroids with a weighting factor. 4. (Bottom right) Run FKM using the fuzziness exponent (phi) and cluster numbers as determined by Odeh et al., (1990), allow additional centroids to resolve in the normal way.
2.2.3.1 Identification of end points (Akro)

End points are determined by the following process:

1. Decompose the soil data $\mathbf{X}$ with $n$ rows of observation and $p$ columns of variables using standardized principal component analysis (PCA), based on the correlation matrix, into $\mathbf{P}$ with $n$ rows and $p$ components. PCA is used because it decomposes soil data into uncorrelated components. This circumvents the need to use Mahalanobis distances, assuming the data can be represented as a continuous variable.

2. Create a convex hull ($\mathbf{C}$) from the PCA components ($\mathbf{P}$)

As the time complexity of convex hull calculation increases by the order of $\mathcal{O}(n^{\log_2})$, large numbers of components become increasingly time consuming to compute (Xu et al., 1998). As the number of components used equals the number of soil properties in the data, which can be greater than 10 components, a method was therefore derived in which the data was compared two dimensions at a time and the convex hulls from these dimensions used as an approximation of a $p$ dimensional hull;

   a) Let the components be an $n \times p$ matrix where each point is represented by a row and each dimension is represented by a column.

   b) Find all $\binom{p}{2}$ combinations of dimensions and combine these into $\frac{\binom{p}{2}}{2}$ 2-column matrices.

   c) Find the row indices of the points on the convex hull for each of these matrices (again, letting the columns be the dimensions of the points).

   d) The union of these sets of row indices are the indices of the simplified hull.

3. Calculate the Euclidean centre of the data, $\mathbf{c}$ (in the case of principal components it is usually zero), set $\mathbf{C}$ as a matrix of the convex hull data.
4. Establish the data point in \( C \) with the maximum distance from the centre or origin \( o \) which will be designated \( e \). Add \( e \) to the resulting end point matrix \( E \).

5. Calculate all the distances between \( e \) and all other points on the convex hull \( C \).

6. Establish the data point with the maximum distance from \( e \), called \( m \).

7. Find Euclidean distance \( d \) between \( m \) and \( e \).

8. Calculate \( r \) (the radius of an exclusion sphere in \( p \) PCA dimensions) as:

\[
r = d \times u, \quad 0 < u <
\]

(1)

9. For the matrix \( C \), exclude the data that is within a sphere of radius \( r \) around the vector \( m \).

10. The data from \( C \) that remains from the exclusion process in step 9 replaces \( C \).

11. Evaluate a stopping criterion \( s \) which is a percentage of the overall distance between the origin \( o \) and the furthest point of the original convex hull as:

\[
s = b \times t, \quad 0 < t < 1
\]

(2)

Where:

\( b = d \) from the first iteration

Therefore:

\[
\text{If } s > r, \text{ stop}
\]

(3)

Repeat from step 4 until the stopping criterion is greater than \( r \) or the list of vectors has been exhausted. The stopping criterion prevents the algorithm from selecting neighboring points. The size of \( u \) determines the number of end points. A large \( u \) will produce a small number of end points; conversely a small \( u \) will produce a larger number. This value can be optimized by multiple runs if it is necessary to achieve a specific number of end points. The end points calculated in this way are measured data points that exhibit the most extreme characteristics. They are tangible, and because they have been sampled, they may represent a larger group in the soil continuum, and can easily be
differentiated from what was previously a generalized extragrade class if it becomes apparent that these data need to be considered.

2.2.3.2 Creation of individual end point centroids (Meson)

The end point centroid algorithm is a modified FKM with partial supervision from Bensaid et al. (1996). In its simplest form, matrix $E$ which contains the end points is designated as data of known clusters. $E$ is combined with principal component data ($P$) and the semi supervised fuzzy-$k$ means algorithm is then run. Essentially the idea is to fix the end points as permanent clusters ($k_e$) and to find additional clusters ($k_d$) within the data space. The combined data, the end point $E$ and observation $P$ is then clustered into $k$ clusters ($k = k_e + k_d$) using the semi supervised FKM algorithm. $E$ needs to be weighted by a factor ($w$) and the centroids for each cluster are calculated as;

$$v_c = \frac{\sum_{i=1}^{n_e} w(u_{ic}^e)^\phi e_i + \sum_{i=1}^{n_p} (u_{ic}^p)^\phi p_i}{\sum_{i=1}^{n_e} w(u_{ic}^e)^\phi + \sum_{i=1}^{n_p} (u_{ic}^p)^\phi}, c = 1, \ldots, k$$ \hspace{1cm} (4)

Where:

$v_c$ is the centroid of cluster $c$

$\phi$ is the fuzzy exponent

$e_i$ is data of end member $i$, (element $i$ of matrix $E$) where $i = 1, \ldots, n_e$

$n_e$ is the number of end points

$w$ is the weight given to each end point

$u_{ic}^e$ is the membership of data $e_i$ to class $c$, which is fixed

$p_i$ is data point $i$, (element $i$ of matrix $P$) where $i = 1, \ldots, n_p$

$n_p$ is the number of observations $P$

$u_{ic}^p$ is the membership of data $p_i$ to class $c$, which is to be determined

The membership of $E$ is fixed, while the membership of the other data is determined by the conventional FKM algorithm:
\[ u_{ic}^p = \frac{d_{ic}^{-2/(\varphi - 1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\varphi - 1)}}, \quad i=1,\ldots,n_p, \quad c=1,\ldots,k \]  \hspace{1cm} (5)

Where:

- \( d_{ic} \) is the distance between centroid \( \nu_c \) and data point \( p_i \).

The algorithm iterates between equation 2 and 3 until a stable solution is reached. The parameters that need to be considered are \( \varphi \) (fuzziness exponent), \( w \) (weighting factor) and \( k \) (the number of clusters).

Weighting factor \( (w) \) should be proportional to the size of the data set, but can also be user defined. The \( \varphi \) and \( k \) can be determined using (Odeh et al., 1990). The algorithm determines isolated points on the extreme edge of the data field. These points are sometimes outliers which will be discussed in later chapters or can be the result of errors in data entry. An error is typically a single point allocated to a cluster. Such points can be singled out and individually assessed.

### 2.2.3.3 Nomenclature

The akromeson algorithm produces two distinct sets of clusters. One set, the extragrade clusters are defined by the training step, the second set, the intragrade clusters are defined by the fuzzy-k algorithm. The overall number of clusters in each instance can be set by the user. Nomenclature should therefore be similar to FKM (e.g. defining five clusters can be described as fuzzy-5 means, as demonstrated in papers such as McBratney and de Gruijter1992), but should highlight extragrade and intragrade clusters. The akromeson algorithm therefore will be expressed in the following way;

Akromeson \( x,y \) with;

- \( x \) being the end point or extragrade cluster number, and
- \( y \) being the intragrade cluster number.

Therefore, an akromeson algorithm that identifies nine end points and resolves 15 clusters making a total of 24 clusters would be referred to as or AM (9,15). These parameters would indicate that 24 centroids would end up in the final centroid table, nine fixed clusters in the periphery of the data and 15 in the center.
2.3 The problem with extragrades and a possible solution

Fuzzy-k means with extragrades is a proven method of identifying outliers. It is difficult however, to determine between outlying clusters and extragrade data. Let us now compare the FKM with extragrades method with the akromeson approach. Table 1 is a notional two dimensional data set comprising of 33 bivariate observations. The graphical representation of this table is shown in Figure 2.
Table 1. A notional data set which is plotted in figure 1. Cluster a, b, c, d and e represent typical data which is organised into clusters. Extreme points and outlying points, which tend to sit far away from the data field. They are representative of the data that needs to be incorporated into cluster analysis, yet have the potential to bias the procedure. X and y are coordinates of the data.

<table>
<thead>
<tr>
<th>Row</th>
<th>ID</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cluster a</td>
<td>46.94</td>
<td>24.23</td>
</tr>
<tr>
<td>2</td>
<td>Cluster a</td>
<td>48.19</td>
<td>27.82</td>
</tr>
<tr>
<td>3</td>
<td>Cluster a</td>
<td>46.86</td>
<td>27.78</td>
</tr>
<tr>
<td>4</td>
<td>Cluster a</td>
<td>47</td>
<td>25.63</td>
</tr>
<tr>
<td>5</td>
<td>Cluster a</td>
<td>49.58</td>
<td>29.1</td>
</tr>
<tr>
<td>6</td>
<td>Cluster b</td>
<td>27.81</td>
<td>50.67</td>
</tr>
<tr>
<td>7</td>
<td>Cluster b</td>
<td>25.63</td>
<td>47.84</td>
</tr>
<tr>
<td>8</td>
<td>Cluster b</td>
<td>24.82</td>
<td>50.2</td>
</tr>
<tr>
<td>9</td>
<td>Cluster b</td>
<td>22.15</td>
<td>51.77</td>
</tr>
<tr>
<td>10</td>
<td>Cluster b</td>
<td>21.41</td>
<td>49.95</td>
</tr>
<tr>
<td>11</td>
<td>Cluster c</td>
<td>51.48</td>
<td>49.57</td>
</tr>
<tr>
<td>12</td>
<td>Cluster c</td>
<td>50.33</td>
<td>49.24</td>
</tr>
<tr>
<td>13</td>
<td>Cluster c</td>
<td>48.31</td>
<td>49.07</td>
</tr>
<tr>
<td>14</td>
<td>Cluster c</td>
<td>49.94</td>
<td>49.68</td>
</tr>
<tr>
<td>15</td>
<td>Cluster c</td>
<td>49.19</td>
<td>47.26</td>
</tr>
<tr>
<td>16</td>
<td>Cluster d</td>
<td>71.51</td>
<td>44.08</td>
</tr>
<tr>
<td>17</td>
<td>Cluster d</td>
<td>73.86</td>
<td>50.94</td>
</tr>
<tr>
<td>18</td>
<td>Cluster d</td>
<td>78.13</td>
<td>48.54</td>
</tr>
<tr>
<td>19</td>
<td>Cluster d</td>
<td>77.13</td>
<td>52.72</td>
</tr>
<tr>
<td>20</td>
<td>Cluster d</td>
<td>75.47</td>
<td>45.3</td>
</tr>
<tr>
<td>21</td>
<td>Cluster e</td>
<td>50.72</td>
<td>73.73</td>
</tr>
<tr>
<td>22</td>
<td>Cluster e</td>
<td>49.35</td>
<td>76.58</td>
</tr>
<tr>
<td>23</td>
<td>Cluster e</td>
<td>50.8</td>
<td>74.21</td>
</tr>
<tr>
<td>24</td>
<td>Cluster e</td>
<td>50.88</td>
<td>76.07</td>
</tr>
<tr>
<td>25</td>
<td>Cluster e</td>
<td>50.89</td>
<td>74.26</td>
</tr>
<tr>
<td>26</td>
<td>Extreme point</td>
<td>4.97</td>
<td>6.09</td>
</tr>
<tr>
<td>27</td>
<td>Extreme point</td>
<td>99.64</td>
<td>1.58</td>
</tr>
<tr>
<td>28</td>
<td>Extreme point</td>
<td>6.86</td>
<td>90.37</td>
</tr>
<tr>
<td>29</td>
<td>Extreme point</td>
<td>95.11</td>
<td>95.72</td>
</tr>
<tr>
<td>30</td>
<td>Outlying point</td>
<td>82.21</td>
<td>16.83</td>
</tr>
<tr>
<td>31</td>
<td>Outlying point</td>
<td>81.91</td>
<td>79.08</td>
</tr>
<tr>
<td>32</td>
<td>Outlying point</td>
<td>21.74</td>
<td>15.59</td>
</tr>
<tr>
<td>33</td>
<td>Outlying point</td>
<td>18.37</td>
<td>81.83</td>
</tr>
</tbody>
</table>

39
Each symbol represents a data point in x,y format. These data are not symmetrical and they have been generated in such a way that they are slightly random and ultimately organize themselves into 5 discrete, equal clusters. Aside from these clusters, there are data representing extreme information. This data is organized into two layers; a layer which is close to, but not touching the edge of the data field, described as outlying points, and a layer situated at the extreme edge of the data field, described as extreme points.

A FKM analysis without extragrades was performed on this data set (Figure 3), $\varphi$ of 1.75 and $k=5$. This demonstrated the classical problem: the necessity of accounting for extreme data. The cluster centroids (represented by solid points), although in the general vicinity of the clusters, were influenced by the leverage of the extragrade points. If this were applied to real soil information, data in between the extragrades and the true centroids would be interpreted as having the shortest distance and data within the cluster would be calculated as having longer distances, leading to less accurate classifications. Centroid positions and memberships are presented (Table 2 and Table 3).
Table 2. Centroids calculated from Fuzzy K Means (FKM) applied to data from Figure 1 and plotted in Figure 3. Each letter is representative of a cluster centroid. Their positions in the data field are denoted by the x and y columns.

<table>
<thead>
<tr>
<th>id</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5a</td>
<td>50.14</td>
<td>48.86</td>
</tr>
<tr>
<td>5b</td>
<td>79.06</td>
<td>59.19</td>
</tr>
<tr>
<td>5c</td>
<td>22.09</td>
<td>38.38</td>
</tr>
<tr>
<td>5d</td>
<td>40.18</td>
<td>78.09</td>
</tr>
<tr>
<td>5e</td>
<td>61.19</td>
<td>19.66</td>
</tr>
</tbody>
</table>
Table 3. Memberships to the clusters in Table 2 from FKM as applied to data from Figure 1. The id column indicates the cluster that the data should reside. The positions of the clusters are shown in Table 2. The Maximum allocation to Cluster (MaxCls) column indicates the point with the maximum membership. The letters for each cluster were randomly generated which is why id “a” is represented by 5e, and id “c” is represented by 5a. It is important to note that both systems still identify the same cluster groups even though the letters are assigned randomly.

<table>
<thead>
<tr>
<th>id</th>
<th>MaxCls</th>
<th>Cl</th>
<th>5a</th>
<th>5b</th>
<th>5c</th>
<th>5d</th>
<th>5e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>5e</td>
<td>0.005</td>
<td>0.002</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0.997</td>
</tr>
<tr>
<td>a</td>
<td>5e</td>
<td>0.023</td>
<td>0.011</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0.988</td>
</tr>
<tr>
<td>a</td>
<td>5e</td>
<td>0.009</td>
<td>0.004</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0.995</td>
</tr>
<tr>
<td>a</td>
<td>5e</td>
<td>0.021</td>
<td>0.01</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0.989</td>
</tr>
<tr>
<td>a</td>
<td>5e</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>5c</td>
<td>0.002</td>
<td>0.001</td>
<td>0</td>
<td>0.999</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5c</td>
<td>0.007</td>
<td>0.003</td>
<td>0</td>
<td>0.997</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5c</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0.998</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5c</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5c</td>
<td>0.037</td>
<td>0.017</td>
<td>0</td>
<td>0.98</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>5a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>5a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>5a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>5a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>5a</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>5b</td>
<td>0.242</td>
<td>0.109</td>
<td>0.867</td>
<td>0</td>
<td>0</td>
<td>0.024</td>
</tr>
<tr>
<td>d</td>
<td>5b</td>
<td>0.002</td>
<td>0.001</td>
<td>0.999</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>5b</td>
<td>0.001</td>
<td>0</td>
<td>0.999</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>5b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>5b</td>
<td>0.023</td>
<td>0.01</td>
<td>0.987</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
</tr>
<tr>
<td>e</td>
<td>5d</td>
<td>0.004</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0.998</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>5d</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>5d</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0.998</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>5d</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0.999</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>5d</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0.998</td>
<td>0</td>
</tr>
<tr>
<td>EM1</td>
<td>5c</td>
<td>0.065</td>
<td>0.014</td>
<td>0.001</td>
<td>0.96</td>
<td>0.002</td>
<td>0.024</td>
</tr>
<tr>
<td>EM2</td>
<td>5e</td>
<td>0.125</td>
<td>0.02</td>
<td>0.05</td>
<td>0.003</td>
<td>0.001</td>
<td>0.925</td>
</tr>
<tr>
<td>EM3</td>
<td>5d</td>
<td>0.082</td>
<td>0.014</td>
<td>0.002</td>
<td>0.032</td>
<td>0.951</td>
<td>0.001</td>
</tr>
<tr>
<td>EM4</td>
<td>5b</td>
<td>0.12</td>
<td>0.019</td>
<td>0.929</td>
<td>0.001</td>
<td>0.049</td>
<td>0.003</td>
</tr>
<tr>
<td>EX1</td>
<td>5e</td>
<td>0.01</td>
<td>0.002</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>0.994</td>
</tr>
<tr>
<td>EX2</td>
<td>5b</td>
<td>0.008</td>
<td>0.002</td>
<td>0.995</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>EX2</td>
<td>5c</td>
<td>0.029</td>
<td>0.005</td>
<td>0</td>
<td>0.983</td>
<td>0</td>
<td>0.012</td>
</tr>
<tr>
<td>EX4</td>
<td>5d</td>
<td>0.012</td>
<td>0.003</td>
<td>0</td>
<td>0.004</td>
<td>0.992</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 3. FKM applied to Figure 1 data. The movement of the centroids (represented by solid points) away from the clusters (crosses, triangles, circles, saltires, diamonds) that are supposed to envelop them demonstrates how extreme data can bias the result.

One solution to this problem has been to include an extragrade class, which includes data that are on the periphery of the data space. This reduces its capacity to affect other centroids. To demonstrate this effect, the example data have been analysed via the FKM with extragrades using Euclidean distances (Figure 4). Centroid positions and memberships are presented in Table 4 and Table 5.

Table 4. Positional data pertaining to centroids calculated using FKM with Extragrades. The table shows the positions of centroids recalculated from data in Figure 1 using Fuzzy K Means with Extragrades.

<table>
<thead>
<tr>
<th>id</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5a</td>
<td>75.8</td>
<td>48.89</td>
</tr>
<tr>
<td>5b</td>
<td>50.59</td>
<td>49.41</td>
</tr>
<tr>
<td>5c</td>
<td>49.32</td>
<td>47.70</td>
</tr>
<tr>
<td>5d</td>
<td>49.69</td>
<td>23.99</td>
</tr>
<tr>
<td>5e</td>
<td>50.59</td>
<td>74.87</td>
</tr>
</tbody>
</table>

43
Table 4. Memberships from FKM with extragrades applied to Figure 1 data. The Maximum allocation to Cluster (MaxCls) column now has point data for two extremely close centroids, and memberships are scattered.

<table>
<thead>
<tr>
<th>MaxCls</th>
<th>Cl</th>
<th>5a</th>
<th>5b</th>
<th>5c</th>
<th>5d</th>
<th>5e</th>
<th>5*</th>
</tr>
</thead>
<tbody>
<tr>
<td>5d</td>
<td>0.194</td>
<td>0.026</td>
<td>0.04</td>
<td>0.044</td>
<td>0.851</td>
<td>0.017</td>
<td>0.021</td>
</tr>
<tr>
<td>5d</td>
<td>0.089</td>
<td>0.012</td>
<td>0.02</td>
<td>0.023</td>
<td>0.934</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>5d</td>
<td>0.05</td>
<td>0.007</td>
<td>0.012</td>
<td>0.013</td>
<td>0.963</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>5d</td>
<td>0.084</td>
<td>0.012</td>
<td>0.019</td>
<td>0.021</td>
<td>0.937</td>
<td>0.007</td>
<td>0.003</td>
</tr>
<tr>
<td>5d</td>
<td>0.376</td>
<td>0.052</td>
<td>0.069</td>
<td>0.075</td>
<td>0.7</td>
<td>0.03</td>
<td>0.075</td>
</tr>
<tr>
<td>5*</td>
<td>0.576</td>
<td>0.049</td>
<td>0.118</td>
<td>0.125</td>
<td>0.08</td>
<td>0.078</td>
<td>0.549</td>
</tr>
<tr>
<td>5*</td>
<td>0.649</td>
<td>0.053</td>
<td>0.136</td>
<td>0.145</td>
<td>0.086</td>
<td>0.085</td>
<td>0.496</td>
</tr>
<tr>
<td>5*</td>
<td>0.71</td>
<td>0.056</td>
<td>0.149</td>
<td>0.162</td>
<td>0.098</td>
<td>0.082</td>
<td>0.452</td>
</tr>
<tr>
<td>5*</td>
<td>0.602</td>
<td>0.051</td>
<td>0.123</td>
<td>0.132</td>
<td>0.088</td>
<td>0.075</td>
<td>0.531</td>
</tr>
<tr>
<td>5*</td>
<td>0.67</td>
<td>0.055</td>
<td>0.142</td>
<td>0.15</td>
<td>0.082</td>
<td>0.092</td>
<td>0.48</td>
</tr>
<tr>
<td>5b</td>
<td>0.373</td>
<td>0.01</td>
<td>0.799</td>
<td>0.172</td>
<td>0.009</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>5b</td>
<td>0.182</td>
<td>0.003</td>
<td>0.905</td>
<td>0.087</td>
<td>0.003</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>5c</td>
<td>0.809</td>
<td>0.014</td>
<td>0.38</td>
<td>0.572</td>
<td>0.016</td>
<td>0.015</td>
<td>0.003</td>
</tr>
<tr>
<td>5b</td>
<td>0.396</td>
<td>0.006</td>
<td>0.792</td>
<td>0.188</td>
<td>0.007</td>
<td>0.007</td>
<td>0</td>
</tr>
<tr>
<td>5c</td>
<td>0.194</td>
<td>0.004</td>
<td>0.091</td>
<td>0.897</td>
<td>0.005</td>
<td>0.004</td>
<td>0</td>
</tr>
<tr>
<td>5a</td>
<td>0.558</td>
<td>0.553</td>
<td>0.11</td>
<td>0.105</td>
<td>0.072</td>
<td>0.053</td>
<td>0.107</td>
</tr>
<tr>
<td>5a</td>
<td>0.222</td>
<td>0.827</td>
<td>0.049</td>
<td>0.046</td>
<td>0.027</td>
<td>0.031</td>
<td>0.02</td>
</tr>
<tr>
<td>5a</td>
<td>0.154</td>
<td>0.879</td>
<td>0.033</td>
<td>0.031</td>
<td>0.022</td>
<td>0.021</td>
<td>0.013</td>
</tr>
<tr>
<td>5a</td>
<td>0.305</td>
<td>0.756</td>
<td>0.061</td>
<td>0.057</td>
<td>0.036</td>
<td>0.043</td>
<td>0.047</td>
</tr>
<tr>
<td>5a</td>
<td>0.281</td>
<td>0.778</td>
<td>0.058</td>
<td>0.055</td>
<td>0.04</td>
<td>0.033</td>
<td>0.036</td>
</tr>
<tr>
<td>5e</td>
<td>0.065</td>
<td>0.01</td>
<td>0.016</td>
<td>0.015</td>
<td>0.006</td>
<td>0.951</td>
<td>0.002</td>
</tr>
<tr>
<td>5e</td>
<td>0.128</td>
<td>0.019</td>
<td>0.03</td>
<td>0.027</td>
<td>0.012</td>
<td>0.901</td>
<td>0.01</td>
</tr>
<tr>
<td>5e</td>
<td>0.033</td>
<td>0.005</td>
<td>0.008</td>
<td>0.008</td>
<td>0.003</td>
<td>0.975</td>
<td>0.001</td>
</tr>
<tr>
<td>5e</td>
<td>0.065</td>
<td>0.01</td>
<td>0.016</td>
<td>0.015</td>
<td>0.006</td>
<td>0.95</td>
<td>0.003</td>
</tr>
<tr>
<td>5e</td>
<td>0.032</td>
<td>0.005</td>
<td>0.008</td>
<td>0.007</td>
<td>0.003</td>
<td>0.976</td>
<td>0.001</td>
</tr>
<tr>
<td>5*</td>
<td>0.164</td>
<td>0.018</td>
<td>0.026</td>
<td>0.027</td>
<td>0.037</td>
<td>0.018</td>
<td>0.873</td>
</tr>
<tr>
<td>5*</td>
<td>0.16</td>
<td>0.033</td>
<td>0.023</td>
<td>0.023</td>
<td>0.031</td>
<td>0.017</td>
<td>0.873</td>
</tr>
<tr>
<td>5*</td>
<td>0.177</td>
<td>0.019</td>
<td>0.029</td>
<td>0.029</td>
<td>0.02</td>
<td>0.04</td>
<td>0.863</td>
</tr>
<tr>
<td>5*</td>
<td>0.185</td>
<td>0.037</td>
<td>0.027</td>
<td>0.026</td>
<td>0.019</td>
<td>0.038</td>
<td>0.854</td>
</tr>
<tr>
<td>5*</td>
<td>0.409</td>
<td>0.088</td>
<td>0.057</td>
<td>0.057</td>
<td>0.086</td>
<td>0.034</td>
<td>0.679</td>
</tr>
<tr>
<td>5*</td>
<td>0.444</td>
<td>0.097</td>
<td>0.062</td>
<td>0.058</td>
<td>0.037</td>
<td>0.094</td>
<td>0.653</td>
</tr>
<tr>
<td>5*</td>
<td>0.385</td>
<td>0.035</td>
<td>0.057</td>
<td>0.061</td>
<td>0.099</td>
<td>0.034</td>
<td>0.714</td>
</tr>
<tr>
<td>5*</td>
<td>0.328</td>
<td>0.032</td>
<td>0.052</td>
<td>0.051</td>
<td>0.032</td>
<td>0.08</td>
<td>0.752</td>
</tr>
</tbody>
</table>
Figure 4. Fuzzy K Means with Extragrams applied to Figure 1 data. This plot demonstrates how exclusion of extragrams (circles) can help correctly position the centroids (black solid circles). However, there are too many extragrams and the centroids have partitioned themselves amongst the remaining clusters. Errors in proportions of extragrams have caused this to occur.

While effective, the extragram model is hampered by the factor of how many extragrams to recognize. As shown in Figure 4, triangles, crosses, diamonds and saltires indicate cluster groups, and circles indicate the extragram class. Centroids are represented by solid points. The results of running FKM with extragrams are not as accurate as expected. As is demonstrated by the plot, all of the points in one of the generated clusters are identified as extragrams. The data in the middle of the plot were split into two groups. The original FKM with extragrams algorithm has the default proportion of extragrams as 1/(k+1). In the instance of the trial dataset (5 clusters), resulting in 1/6 of the data being classified as extragrams, an excessive number. In order to accurately run FKM with extragrams, it is necessary to be able to correctly estimate the proportion of points that are extragrams in the dataset, which is difficult. Incorrectly estimating this value results in clusters becoming extragrams and vice versa. AM(4,5) with a w of 200 was performed on the constructed data set (Figure 5).
Figure 5. Dataset from Figure 1 corrected using the akromeson algorithm. The triangles represent data that has been assigned to endpoint clusters. The open circles represent data that has been allocated to centroids that have been allocated to intragrade clusters. Note that there are four additional centroids on the periphery of the data space (which need to be accounted for when performing the calculation). However, using this method the centroids (solid points) are correctly positioned.

The four fixed centroids on the periphery encapsulate the extreme data. These clusters form a simple convex hull around the data set, represented by a grey line. Centroids (solid points) are now positioned more accurately in the centre of the remaining clusters. This translates to a better estimate of distances to and from centroids and clusters. Should there be any clusters in the extreme regions of the data, then one of the end point clusters would have a large membership. Data from that end point class can be included in any potential classification.
2.4 Application in the Edgeroi area

The akromeson algorithm was compared to an earlier study which employed FKM with extragrades. The data originated from the results of an ongoing CSIRO survey of the Lower Namoi Valley in the north east of New South Wales, Australia, which began in 1989. This was compiled into a report in 2006 (Triantafilis & McBratney, 2006). The Edgeroi district is situated in the northern end of the Namoi valley. Previous work (Ward, 1999) has located three main landscapes which most likely influence soil development: an eroded shield volcano to the east in the Nandewar mountain range of which outcrops still exist, Pilliga sandstone that can be found in the foothills to the east, or more weathered sandstone remnants to the west. The 1989 data set consisted of 210 sampling sites. Each site had 3m cores extracted and measurements taken at 10, 20, 30, 70, 120 and 250 cm, all of which were used in the calculations. Since this time, additional data have been collected (Triantafilis et al., 2013). The modern data set comprises of approximately 341 pedons. Experimental data used previously were organic C, CaCO₃, Sand, Silt, Clay (recorded as percentages), Cl and P (recorded in mg/kg), Ca, Mg, K, and Na (in mmol/kg), EC (recorded in S/m) and pH. These will therefore be used for a comparison. Research in 1993, using a phi of 1.2, determined the area consisted of 10 main clusters plus an additional extragrade cluster. The data will be analysed using FKM with extragrades, then these results compared with akromeson, (AM 5,6) using weights (w) of 30, 60, 100, 200 and 400.
2.5 Results

The first two principal components of the resulting clusters of the FKM with extragrades on the Edgeroi data were converted into principal components, a method of orthogonalising and reducing the size of complex data, extracted and plotted individually via a pairs plot (Figure 6). This demonstrated that while the centroids were positioned in the middle of clusters, a large number of extragrades were in the centre of the data (Figure 6, plot 10*), indicating that the number of extragrades selected was too high. This means that some legitimate data has not been allocated to the correct cluster, affecting the position of the centroids. This was exactly what happened in the constructed dataset (Table 6).

The problem with FKM with extragrades is that the proportion of the data that is regarded as extragrades (set at 1/(k+1)) is determined by the number of clusters chosen and can therefore be arbitrary. In this dataset with 10 clusters and 2072 data points, there will be 188 extragrades, regardless of the distribution of the data. As a result, the extragrades “cluster”, called 10* in the top left of (Figure 6), is diffuse and does not necessarily represent true extragrades. The centroid table and pairs plot form the basis for comparison between FKM with extragrades and the akromeson method.

Table 5. Properties of Edgeroi centroids as portioned by Fuzzy K Means with extragrades. A total of 10 Fuzzy K Means clusters are represented in this table. Each centroid has the mean values for 12 properties. The accuracy of the nominated centroid properties is dependent on whether the number of extragrades in the data correspond to the equation 1/K+1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Silt (%)</th>
<th>Clay (%)</th>
<th>pH</th>
<th>EC (MS/CM)</th>
<th>Cl^- (mg/kg)</th>
<th>CaCO3 (mg/kg)</th>
<th>C (%)</th>
<th>P (mg/kg)</th>
<th>Ca (mmol/kg)</th>
<th>Mg (mmol/kg)</th>
<th>K (mmol/kg)</th>
<th>Na (mmol/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10a</td>
<td>15.34</td>
<td>47.57</td>
<td>8.49</td>
<td>18.43</td>
<td>42.01</td>
<td>0.89</td>
<td>0.96</td>
<td>10.32</td>
<td>25.524</td>
<td>124.83</td>
<td>8.04</td>
<td>24.75</td>
</tr>
<tr>
<td>10b</td>
<td>15.41</td>
<td>59.95</td>
<td>8.41</td>
<td>18.74</td>
<td>57.9</td>
<td>0.46</td>
<td>0.7</td>
<td>24.93</td>
<td>218.4</td>
<td>154.56</td>
<td>13.06</td>
<td>28.87</td>
</tr>
<tr>
<td>10c</td>
<td>8.4</td>
<td>30.52</td>
<td>8.67</td>
<td>23.74</td>
<td>127.41</td>
<td>1.25</td>
<td>0.25</td>
<td>8.91</td>
<td>69.02</td>
<td>94.63</td>
<td>5.48</td>
<td>27.92</td>
</tr>
<tr>
<td>10d</td>
<td>16.49</td>
<td>56.81</td>
<td>9.02</td>
<td>51.03</td>
<td>239.11</td>
<td>1.32</td>
<td>0.32</td>
<td>14.98</td>
<td>201.61</td>
<td>163.36</td>
<td>9.78</td>
<td>94.05</td>
</tr>
<tr>
<td>10e</td>
<td>5.84</td>
<td>14.43</td>
<td>5.78</td>
<td>7.17</td>
<td>52.29</td>
<td>0.02</td>
<td>0.37</td>
<td>9.05</td>
<td>13.52</td>
<td>21.4</td>
<td>2.96</td>
<td>6.46</td>
</tr>
<tr>
<td>10f</td>
<td>21.32</td>
<td>43.85</td>
<td>7.84</td>
<td>11.06</td>
<td>31.96</td>
<td>0.26</td>
<td>1.02</td>
<td>61.25</td>
<td>179.09</td>
<td>105.06</td>
<td>9.01</td>
<td>11.46</td>
</tr>
<tr>
<td>10g</td>
<td>16.21</td>
<td>55.29</td>
<td>8.71</td>
<td>90.48</td>
<td>738.8</td>
<td>1.51</td>
<td>0.18</td>
<td>8.35</td>
<td>215.63</td>
<td>133.91</td>
<td>8.14</td>
<td>103.83</td>
</tr>
<tr>
<td>10h</td>
<td>21.55</td>
<td>47.47</td>
<td>8.55</td>
<td>20.96</td>
<td>68.89</td>
<td>0.8</td>
<td>0.62</td>
<td>13.88</td>
<td>178.82</td>
<td>120.54</td>
<td>6.98</td>
<td>32.73</td>
</tr>
<tr>
<td>10i</td>
<td>11.92</td>
<td>29.43</td>
<td>6.72</td>
<td>12.26</td>
<td>22.44</td>
<td>0.08</td>
<td>1.75</td>
<td>20.2</td>
<td>104.53</td>
<td>63.12</td>
<td>10.36</td>
<td>4.82</td>
</tr>
<tr>
<td>10j</td>
<td>17.94</td>
<td>49.84</td>
<td>8.99</td>
<td>36.41</td>
<td>156.05</td>
<td>2.13</td>
<td>0.48</td>
<td>13.43</td>
<td>140.88</td>
<td>222.69</td>
<td>7.65</td>
<td>62.79</td>
</tr>
</tbody>
</table>
An AM (5,6) was run, based on a total of 11 clusters. The diameter of exclusion \((u)\) was set at 0.8 to accommodate the low number of clusters. The stopping criterion was set at 40% of the maximum length of \(u\) \((0.4r)\). The weighting factor \((w)\) was set at 40 and the \(\varphi\) was 1.25, determined by a comparison of the fuzziness performance index and number of clusters, as described in (Odeh et al., 1990). The end point positions are presented in Table 7. This resulted in 11 clusters where the first 5 of these clusters are derived from the end point analysis. The centroid table \((w=60)\) is as follows (Table 8).
Table 6. Initial centroids that were input into the akromeson algorithm. Soil ID relates to a physical location within the Edgeroi region. These locations are linked to soil properties displayed in the table below. These locations represent the initial end point clusters prior to the running of the akromeson algorithm.

<table>
<thead>
<tr>
<th>Soil ID</th>
<th>Silt (%)</th>
<th>Clay (%)</th>
<th>pH</th>
<th>EC (CS/M)</th>
<th>Cl⁻ (mg/kg)</th>
<th>CaCO₃ (mg/kg)</th>
<th>C (%)</th>
<th>Ca (mmol/ kg)</th>
<th>Mg (mmol/ kg)</th>
<th>K (mmol/ kg)</th>
<th>Na (mmol/ kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ed204(E1)</td>
<td>16</td>
<td>45</td>
<td>8.66</td>
<td>118.5</td>
<td>1059</td>
<td>8.5</td>
<td>0.9</td>
<td>6.1</td>
<td>76.7</td>
<td>471.4</td>
<td>2</td>
</tr>
<tr>
<td>ed337(E2)</td>
<td>6.5</td>
<td>19.9</td>
<td>6.67</td>
<td>26.3</td>
<td>29</td>
<td>0.05</td>
<td>8.27</td>
<td>91.5</td>
<td>101.4</td>
<td>66.9</td>
<td>13.8</td>
</tr>
<tr>
<td>ed601(E3)</td>
<td>7.8</td>
<td>17.1</td>
<td>9.18</td>
<td>14.3</td>
<td>12</td>
<td>77.6</td>
<td>0.04</td>
<td>0.05</td>
<td>32.2</td>
<td>102.8</td>
<td>0.05</td>
</tr>
<tr>
<td>ed601(E4)</td>
<td>18.1</td>
<td>27.5</td>
<td>4.34</td>
<td>64.3</td>
<td>603</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>145</td>
<td>1.9</td>
<td>27.6</td>
</tr>
<tr>
<td>na222(E5)</td>
<td>18.3</td>
<td>58</td>
<td>8.1</td>
<td>171.2</td>
<td>764</td>
<td>0.5</td>
<td>0.55</td>
<td>26.6</td>
<td>196</td>
<td>111.9</td>
<td>133.9</td>
</tr>
</tbody>
</table>

Table 7. Cluster centroids as determined by akromeson algorithm. This table indicates the final positions of all clusters, end points (E1-E5) will demonstrate some movement away from their initial position (Table 7).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Soil ID</th>
<th>Silt (%)</th>
<th>Clay (%)</th>
<th>pH</th>
<th>EC (CS/M)</th>
<th>Cl⁻ (mg/kg)</th>
<th>CaCO₃ (mg/kg)</th>
<th>C (%)</th>
<th>Ca (mmol/ kg)</th>
<th>Mg (mmol/ kg)</th>
<th>K (mmol/ kg)</th>
<th>Na (mmol/ kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>E1</td>
<td>15.94</td>
<td>45.36</td>
<td>8.70</td>
<td>111.70</td>
<td>104.2</td>
<td>8.70</td>
<td>0.85</td>
<td>6.04</td>
<td>77.97</td>
<td>455.40</td>
<td>2.22</td>
</tr>
<tr>
<td>b</td>
<td>E2</td>
<td>6.64</td>
<td>19.90</td>
<td>6.68</td>
<td>26.21</td>
<td>29.96</td>
<td>0.11</td>
<td>8.20</td>
<td>90.22</td>
<td>99.51</td>
<td>65.82</td>
<td>13.66</td>
</tr>
<tr>
<td>c</td>
<td>E3</td>
<td>7.59</td>
<td>17.64</td>
<td>9.14</td>
<td>14.47</td>
<td>14.75</td>
<td>74.9</td>
<td>0.07</td>
<td>0.76</td>
<td>35.68</td>
<td>97.82</td>
<td>0.40</td>
</tr>
<tr>
<td>d</td>
<td>E4</td>
<td>17.01</td>
<td>30.43</td>
<td>4.57</td>
<td>64.99</td>
<td>583.40</td>
<td>0.16</td>
<td>0.08</td>
<td>2.90</td>
<td>18.58</td>
<td>138.45</td>
<td>2.33</td>
</tr>
<tr>
<td>e</td>
<td>E5</td>
<td>18.28</td>
<td>57.92</td>
<td>8.11</td>
<td>168.94</td>
<td>752.50</td>
<td>0.50</td>
<td>0.55</td>
<td>29.08</td>
<td>195.00</td>
<td>112.41</td>
<td>133.78</td>
</tr>
<tr>
<td>f</td>
<td>C1</td>
<td>21.41</td>
<td>46.17</td>
<td>8.30</td>
<td>20.48</td>
<td>66.52</td>
<td>0.92</td>
<td>0.80</td>
<td>25.70</td>
<td>182.43</td>
<td>122.74</td>
<td>8.00</td>
</tr>
<tr>
<td>g</td>
<td>C2</td>
<td>7.20</td>
<td>18.78</td>
<td>6.04</td>
<td>11.20</td>
<td>52.51</td>
<td>0.28</td>
<td>0.66</td>
<td>13.61</td>
<td>34.00</td>
<td>31.81</td>
<td>4.71</td>
</tr>
<tr>
<td>h</td>
<td>C3</td>
<td>16.25</td>
<td>53.72</td>
<td>8.61</td>
<td>92.75</td>
<td>739.62</td>
<td>1.83</td>
<td>0.25</td>
<td>10.87</td>
<td>207.73</td>
<td>133.66</td>
<td>8.18</td>
</tr>
<tr>
<td>i</td>
<td>C4</td>
<td>15.40</td>
<td>52.46</td>
<td>8.31</td>
<td>19.39</td>
<td>44.55</td>
<td>0.96</td>
<td>1.02</td>
<td>17.68</td>
<td>249.75</td>
<td>144.41</td>
<td>11.04</td>
</tr>
<tr>
<td>j</td>
<td>C5</td>
<td>16.98</td>
<td>55.30</td>
<td>8.98</td>
<td>48.18</td>
<td>200.98</td>
<td>1.75</td>
<td>0.40</td>
<td>18.76</td>
<td>171.93</td>
<td>174.88</td>
<td>9.64</td>
</tr>
<tr>
<td>k</td>
<td>C6</td>
<td>9.82</td>
<td>33.05</td>
<td>8.61</td>
<td>26.58</td>
<td>129.61</td>
<td>1.67</td>
<td>0.37</td>
<td>12.37</td>
<td>89.61</td>
<td>101.91</td>
<td>6.09</td>
</tr>
</tbody>
</table>

2.5.1 Weighting factor (w)
Panel plots in Figures 7 and 8 indicated differences between FKM with extragrades and akromeson, but also between different weighting factors within akromeson. With a weighting (w) of 60. Panel plot i indicated a centroid that was in the periphery of the data rather than the centre.
Figure 7. Resolution of clusters in principal component space using akromeson, w=60. This figure displays Akromeson 5,6 with a w of 60, split into 11 separate panel plots, plotted against the scores of principal components 1 and 2, designated by letters at the top of each plot. Plots a to e represent end point clusters, plots f to k are central clusters. All of the data is plotted in dark grey to demonstrate where in the data cloud each particular cluster is located. The clusters themselves are heat mapped according to the scale on the top right. The circle and the triangle indicate the locations of each centroid. In the key on the bottom right hand side, E stands for extragrade centroid, C stands for central centroid.

When the weighting factor (w) was reduced to 30, the centroids moved to the centre of their respective clusters (Figure 8, Table 9).
Figure 8. Resolution of clusters in principal component space using akromeson, w = 30. Panel plot of Akromeson 5.6 with w of 30 split into 11 separate panel plots, plotted against the scores of principal components 1 and 2, designated by letters at the top of each plot. Plots a to e represent end point clusters, plots f to k are central clusters. All of the data is plotted in dark grey to demonstrate where in the data cloud each particular cluster is located. The clusters themselves are heat mapped according to the scale on the top right. The circle and the triangle indicate the locations of each centroid. In the key on the bottom right hand side, E stands for extragrade centroid, C stands for central centroid. Note the resolution of cluster and centroid for plot i.
Table 8. Centroid table for AM 5.6 with a \( w \) of 30. This table demonstrates more change in the property values of the fixed centroids, owing to the lower weighting factor. This numerical movement is towards the mathematical centre of any clusters of extreme data.

<table>
<thead>
<tr>
<th>Soil ID</th>
<th>Silt (%)</th>
<th>Clay (%)</th>
<th>pH</th>
<th>EC (CS/MM)</th>
<th>CaCO(_3) (mg/kg)</th>
<th>C (%)</th>
<th>P (mg/kg)</th>
<th>Ca (mmol/kg)</th>
<th>Mg (mmol/kg)</th>
<th>K (mmol/kg)</th>
<th>Na (mmol/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>15.9</td>
<td>45.7</td>
<td>8.7</td>
<td>107.3</td>
<td>969.4</td>
<td>8.7</td>
<td>0.8</td>
<td>5.8</td>
<td>79.5</td>
<td>440.6</td>
<td>2.4</td>
</tr>
<tr>
<td>E2</td>
<td>6.7</td>
<td>19.9</td>
<td>6.7</td>
<td>26.2</td>
<td>30.7</td>
<td>0.1</td>
<td>8.1</td>
<td>89.4</td>
<td>98.3</td>
<td>65.1</td>
<td>13.6</td>
</tr>
<tr>
<td>E3</td>
<td>7.5</td>
<td>18.4</td>
<td>9.1</td>
<td>14.8</td>
<td>14.8</td>
<td>72.0</td>
<td>0.1</td>
<td>1.4</td>
<td>40.3</td>
<td>94.5</td>
<td>0.8</td>
</tr>
<tr>
<td>E4</td>
<td>9.2</td>
<td>22.2</td>
<td>5.5</td>
<td>24.2</td>
<td>198.0</td>
<td>0.2</td>
<td>6.9</td>
<td>23.3</td>
<td>58.3</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>18.3</td>
<td>57.8</td>
<td>8.1</td>
<td>166.9</td>
<td>742.1</td>
<td>0.5</td>
<td>6.6</td>
<td>31.3</td>
<td>194.1</td>
<td>112.9</td>
<td>133.7</td>
</tr>
<tr>
<td>C1</td>
<td>15.5</td>
<td>54.3</td>
<td>8.4</td>
<td>19.8</td>
<td>48.4</td>
<td>1.0</td>
<td>0.9</td>
<td>17.2</td>
<td>250.0</td>
<td>151.7</td>
<td>10.6</td>
</tr>
<tr>
<td>C2</td>
<td>21.6</td>
<td>47.0</td>
<td>8.5</td>
<td>21.6</td>
<td>71.3</td>
<td>1.0</td>
<td>0.7</td>
<td>22.4</td>
<td>183.9</td>
<td>127.2</td>
<td>7.3</td>
</tr>
<tr>
<td>C3</td>
<td>16.9</td>
<td>55.4</td>
<td>9.0</td>
<td>50.7</td>
<td>219.4</td>
<td>1.8</td>
<td>0.4</td>
<td>17.9</td>
<td>172.6</td>
<td>173.9</td>
<td>9.8</td>
</tr>
<tr>
<td>C4</td>
<td>13.5</td>
<td>32.6</td>
<td>7.0</td>
<td>18.1</td>
<td>40.9</td>
<td>0.5</td>
<td>1.7</td>
<td>31.9</td>
<td>135.7</td>
<td>72.0</td>
<td>11.7</td>
</tr>
<tr>
<td>C5</td>
<td>9.0</td>
<td>30.1</td>
<td>8.4</td>
<td>24.1</td>
<td>117.3</td>
<td>1.4</td>
<td>0.3</td>
<td>12.6</td>
<td>80.7</td>
<td>91.9</td>
<td>5.7</td>
</tr>
<tr>
<td>C6</td>
<td>16.3</td>
<td>53.7</td>
<td>8.6</td>
<td>94.9</td>
<td>770.2</td>
<td>1.8</td>
<td>0.2</td>
<td>10.5</td>
<td>205.0</td>
<td>134.6</td>
<td>8.0</td>
</tr>
</tbody>
</table>

The weighting factor became a point of interest. As \( w \) is a parameter that determines the influence of end point clusters, increasing or decreasing this parameter affects the amount of data which is captured in the end point clusters. Several different weights were applied and the number of points per cluster were plotted against the weighting factor (Figure 9). There were three major observations from this;

1. As the weighting factor increased, so did the number of points in central clusters.
2. Between \( w \) of 60 and 200 there was a period of stability, where the distributions of clusters exhibited little change. Past 400, the points in each cluster redistributed.
3. It may be possible to determine the optimum \( w \) and as such, a correct estimate of the proportion of extragrades.

From these experiments, the weighting factor has been determined to be just before the plateau, which corresponds to a weighting of 30, equating to 1% of the data. In the case of the Edgeroi data set, this organizes one of the end points into a large cluster with known properties. If this cluster is accepted into the lexicon of soil groups, only 1% of the data is considered extreme, compared to extragrades in which the rule of thumb necessitated the writing off of 10% of the data, which were more randomly spread and none of which were partitioned into meaningful clusters. It was determined that
experiments with $w$ could optimise the number of extreme data identified. A small value of $w$ could be used to identify small extreme clusters and also place end point centroids in the middle of their corresponding data cloud. A large $w$ performs the primary function of akromeson, which is to neutralise the influence of extreme data without excluding an excessive number of data from the calculation. Extreme groups can still be detected using a large $w$ without excessive modification. Figure 9 demonstrates that $w$ is a relatively robust parameter, organizing clusters similarly from values of 30 to 200.

Figure 9. Comparison of weight factor distribution within clusters. This figure indicated the number of points in a cluster compared to the size of the weighting factor. The solid lines at the bottom of the plot are the numbers of memberships to extragrade clusters. The dotted lines are memberships to intragrade clusters. One extragrade cluster has large memberships with low weightings, which reduce to numbers in line with the rest of the end points as the weighting increases. The lower the value of $w$, the more inclusive the end point clusters. High values of $w$ stabilise the end points into groups of around 20, but the positions of centroids become less acquainted with the clustering of the data. Past 400, the combined leverage of the end points exceeds that of the data. Subsequent centroids are determined more by the positions of individual end points rather than any clustering effect.

2.6 Discussion

2.6.1 End point positions
End members may be a useful tool to partition extragrades into either meaningful clusters or extreme data. The final positions of the fixed end points in the data space are different to the positions that were initially set. As the centroids of the end points are derived from end point vectors ($E$) that are weighted by a factor of $w$, in effect, turned into a set number ($w$) of copies (Bensaid et al., 1996), they
behave just like ordinary clusters. This means that a cluster of nearby data has the capacity to move the centroids. This feature may be useful in outlying cluster detection.

2.6.2 Cluster types
The akromeson algorithm identified similar groups as the FKM with extragrades. There were minor differences in cluster centroids between extragrades. The primary cause was related to the distribution of classes. It was expected that all of the end points would make small clusters which were equivalent to the originally defined classes, but this was not the case. Only one end point resulted in a significant cluster class, reducing the overall number of clusters from eleven to seven. All the non-extragrade data attributes were coerced into these classes, resulting in different centroid positions. The solution to this problem is to account for the number of viable extragrade clusters when estimating the overall cluster numbers.

Akromeson resolved three distinct types of clusters, which have been tentatively defined as intragrade, extragrade and akrograde. Intragrades are akin to normal clusters that are derived by fuzzy-k algorithms. Extragrades in this context are defined by extreme data. The clusters identified can vary from a few points of data to a single data point. It is therefore conceivable that single point, or small cluster data are less extragrades and more genuine outliers.

Akrogrades as extragrade clusters, which have a significant number of data in their memberships were then defined. The significance is currently defined by deviation. If the centroid in the output has deviated from its defined position, this is evidence that a large amount of data points in the vicinity have applied some influence to the cluster. In order to determine whether this is important or not, some things need to be considered such as the strength of the weighting factor and the overall deviation of the extragrades.
2.6.2.1 The weighting factor in akromeson

If the weighting factor is equivalent to 1 data point, then another data point will have sufficient leverage to easily deviate it. If conversely, the weighting factor is many times the strength of the overall data set, then achieving a large deviation will be difficult. In the experiments performed so far, in order to completely immobilize the cluster centroids, the weighting factor would have to be larger than the data set which may not be advisable. The weighting factor can be arbitrarily set to:

\[ w = \frac{0.15 \times np}{ne} \]

A parameter that creates a weight equivalent to 15% of the data (which equates to a \( w \) of 30 in the Edgeroi case), distributed over the outside of the data cloud. This action can correctly position the central clusters and provide some insights into the numbers and behavior of extragrade data without spending too much time adjusting the \( w \) factor. This arbitrary number may be universal, but needs to be tested on more data sets. By examining the Edgeroi data through the lens of akromeson it can be seen that the actual extragrades are less than \( 1/(k+1) \). So akromeson is in this sense more powerful -it can find the actual proportion of extra grading individuals rather than finding them arbitrarily as in FKME.

2.6.2.2 The overall deviation of all the extragrades

It would be difficult to set the extragrade clusters in such a way that no deviation occurred. It is therefore reasonable to expect some deviation from all of the clusters. Deviation may be a diagnostic feature of an extragrade cluster. This makes it imperative to determine how much deviation represents a significant cluster. In future studies, the relationship between number of data points in the cluster and overall deviation will be explored, along with other methods of identifying these extreme clusters.

2.6.3 Significance

Clusters that have been derived more conventionally, especially ones that are the result of FKM, tend to ignore data that is on the periphery of the soil universe. This peripheral data is usually defined by extreme values in one or more area. Akromeson determines if any of this extreme data is well
represented and defines clusters accordingly. This is useful because it may be necessary to know if there is even a small amount of land that requires specific attention. In some cases it may be that it’s the extreme issues with land that are more important than the amount of area it represents. Identifying extreme clusters and knowing the proportion of land that is in such a cluster allows for specialized land management. The number of intragrades are determined by inputs into the algorithm, but the ratio of akrogrades to extragrades are determined by the distribution of the data. So far it would appear that the number of akrogrades created by the algorithm are low. When the Edgeroi data is examined through the lens of akromeson it can be seen that the actual extragrades are less than $1/(k+1)$. So akromeson is in a sense more powerful- it can find the actual proportion of extra grading individuals rather than finding them arbitrarily as in FKM with extragrades.

2.7 Conclusions
Powerful methods now exist to analyse soil data, by resolving multidimensional information into clusters. Fuzzy-k means with extragrades is a tool by which extreme data can be placed in a separate cluster, reducing its leverage on the rest of the data set. It may be possible to refine this concept to reduce the need to estimate extragrade numbers and to further partition extragrades into clusters of their own. The akromeson algorithm works by identifying extreme examples of data at the periphery of the dataset and partitions them into their own cluster groups. The centroids of these groups are end points to which soil properties and profiles grade. This allows clusters in the middle of the data space to resolve more naturally. The algorithm can identify clusters within the extragrades, which may ordinarily be difficult to differentiate from extragrade data. Extragrade clusters can be determined by comparing the positions of the end points from the akro algorithm with the positions of the centroids from the second part or meson section of the algorithm. Any significant deviation of end point centroids from the initial end points is evidence of a cluster. Also a large proportion of the amount of memberships to an end point can be seen as evidence of clustering. The accurate identification of extragrades and their clusters ensures that more legitimate data is included in a main cluster, while more extreme data is separated out into end point clusters. These extreme clusters may represent unusual soils, so it may be advantageous for landholders to be able to identify these small extreme
groups. This system may be a useful method of quickly identifying the true number of extragrades in the data set without requiring estimation or reliance on default settings. These extragrades once identified can be then quickly passed on to experts who ordinarily would be forced to sift through hundreds of profiles to find anomalies.

2.8 Acknowledgments
This is part of an ongoing project towards developing a Universal Soil Classification System. I would like thank Jon Hempel and the USDA NRCS for support. I would also like to thank Dr Jaap de Gruijter for insightful comments on an earlier draft of the manuscript.

2.9 References


Commonwealth Scientific and Industrial Research Organisation and the International Society of Soil Science, Glenside, S.A.


Chapter 3

Creation of soil surface horizon classes from the USDA soil characterization database

Summary
Creating classes for soil surface horizons is a stated goal of the IUSS universal soil-working group. It was decided that for the creation of these classes, a numerical approach would be desirable. Soil surface horizon data was therefore retrieved from the USDA NRCS soil database and compiled into two data sets; one containing 10 soil properties, the other containing 10 soil properties with their corresponding horizon thicknesses. A specially derived fuzzy clustering algorithm called akromeson was trialled on the two data sets. The data without the thickness information was partitioned into 29 assemblages of mean properties, designated horizon classes- 11 central and 18 end point classes, while the data with the thickness information was partitioned into 12 central classes and 23 end point classes. The classification demonstrated greater differentiation when the thickness data was included; as a result, the latter analysis was preferred. The size of the study area meant that innovative strategies were required to demonstrate where on the landscape these classes occurred. A probability occurrence of the soil horizon classes map was constructed based on localized data density. It shows that the landscape classes created conform to what is known about the soils of specific regions and also presents evidence of possible anthropogenic land degradation. When comparing the classes to the US Soil Taxonomy, the end point, or extreme classes mostly had properties such as an Ochric epipedon. These ochric properties were typically defined from one uncommon soil characteristic. The purpose of this analysis is to create better clarity with the definition of ochric horizon classes. These numerical horizon classes may also be useful for universal soil classification.
3.1 Introduction

3.1.1 Development of soil classification
Soil classification as we know it today is a combination of methodologies that have their origins in differing countries. Each different location has produced different regional soil description systems which are based on local conditions. As a result, from the 1930’s to today, many disparate soil classification systems have emerged (Eswaran et al., 2010). Such differences are an impediment to information transfer across the globe. Compounding this problem is the fact that as understanding of soil systems has grown; old systems have been updated or replaced with more comprehensive schemes. Not all land users adapt to change readily resulting in many continuing to use older systems or archaic terms. With increased communications, these differing systems have been required to interact more often on a management and scientific level, with mixed results. It is not uncommon to hear someone discussing Vertisols and cracking clays in the same breath in order to be understood. Most natural sciences such as botany, anthropology and astronomy have common systems (Hempel et al., 2013). It has therefore been recognised that a truly global, easily understood soil classification is needed. The universal soil classification scheme working group was commenced as a consequence and The Gödöllő Accord in 2009 determined the best way forward was to create a world soil map building on the collective experience of systems such as soil taxonomy and world reference base (WRB) in an effort to create a truly universal classification scheme, possibly by taking advantage of numerical classification which can be used to elucidate unfamiliar classifications by expressing soil individuals in terms of a mathematical distance from a calculated average or centroid. In the ongoing research over a global soil map, the concept of soil horizon classification and better soil surface horizon differentiation was highlighted.

3.1.2 The soil horizon
Soil is a complex interaction of organisms and mineral strata (Simonson, 1957). Adding into this complexity is the fact that soils are a three dimensional structure, so when considering the location of a soil, one must also consider the depth. Most classifications recognise that soils are actually
collections of soil layers known as horizons, which are usually defined as a layer within the solum that have roughly uniform properties and are bounded by other soil layers, the geosphere or the soil surface. Although soil horizons are associated by location, the proceeding and preceding layers may not be related. The lower layer could have formed in-situ; the layer above could be a collection of transported material, the next layer gradational with the previous layer, the next with a sharp boundary and so on. In this sense, attempting to make soil profiles based on what could be assemblages of seemingly disparate or random elements could lead to erroneous results. Indeed the concepts of the toposquence and the catena recognise how different soil forming processes can be (Odgers et al., 2008). Soils are understood to contain a mixture of diffuse, gradual, clear and abrupt horizons and predicting when and where an abrupt or contiguous boundary occurs with a view to creating a soil description system is not a simple matter. It may make sense to consider all soil horizons as the basic unit of soil classification and describe their properties accordingly (FitzPatrick, 1993). With horizon classes better understood, the creation of pedon classes may be a more tangible proposition. Two years after the Gödöllő Accord, it was determined that pedon and soil horizon classification schemes were required (Hempel et al., 2013). Also it was recognised that topsoil assessments would need to be a priority in any future classification scheme. There are too few surface horizon classes in current systems considering the large amount of available world soil data and the breadth of ST and WRB in other areas. To create a system of exclusion or inclusion criteria for each horizon class could result in marginal properties being assigned do definitive classes, more commonly referred to as taxonomic chop (Butler, 1980), whereas numerical classification sidesteps this by considering all variables equal and distance to the centroid as the most important factor.

3.1.3 Previous work

It has been recognised since the 1960’s that soil could be analysed using numerical methods and large data sets (FitzPatrick, 1967). The work from this period was hampered by absence of comprehensive data, but continued until promising results were realized in the 1980’s. From this, a need for more exhaustive horizon description systems was recognised and the most important terms plus possible horizon descriptors were outlined (Fitzpatrick, 1993). More recently, it was recognised that soils do
not have sharp boundaries. Maps have therefore been created with memberships to groups rather than hard map units (Verheyen et al., 2001). Fuzzy k means has been used to continuously classify soils in localized areas such as Edgeroi, Australia in the mid 1990’s (Triantifilis and McBratney, 1993). The clustered soils matched the geomorphology of the region, without necessarily having geology as a factor in the clustering process- providing some indication that the method was picking out landscape features. The same data, with some refinements was clustered again in 2013 using a new method known as akromeson (Hughes et al., 2014). The newer algorithm was similarly able to identify landscape features and some idea of the three dimensional structure was ascertained from the analysis. Algorithms such as akromeson are capable of analysing large diverse data sets and creating logical classes. It is proposed that soil data be taken from the (USDA) resource and horizon classes created from this. These surface horizon classes can either be used as a standalone system or as a guide to further refining existing classifications.

3.2 Method

3.2.1 Developing a soil horizon database
The object of this exercise is to create a series of horizon classes for the entire United States, so a database of relevant soil properties is required. The USDA NRCS soil characterization database is one such dataset, having over 80,000 pedons and 200,000 individual soil samples represented. This data ranges over a wide range of soil properties and it is continually updated. The dates and locations of samples are typically included providing up to date soil information as well as legacy data. The United States of America covers over 9 million square kilometers and despite the large store of information available; it is neither complete, nor comprehensive enough to provide detailed resolution in maps. Topographically, the northern American continent is diverse. There are three main mountain ranges, vast plains and large water bodies, each resulting in different pedogenic processes. It cannot be assumed therefore that the conditions in one part of the country are going to be similar to conditions in another. This means that when performing a numeric analysis, there will need to be some post processing to create a realistic geographic distribution. In this study, only horizons, not
pedons are examined. The data would form a matrix in which the rows would represent samples taken from a given horizon and there would be one column for each property sampled.

3.2.2 Soil properties
An important indicator of soil health and plant viability is the pH (Pankhurst et al., 1995). It is also one of the most common variables collected from a soil sample site. These two facts make it an ideal soil property for cluster analysis. Soil texture is likewise commonplace in the data set and is important because texture classes are well known to experts, stakeholders and farmers alike. Similarly, colour is included because it can be diagnostic of many soil properties such as mineralogy, drainage and oxidation state and is easily recognised and understood (Simonson and Boersma, 1972). These soil properties are also useful in traditional soil and horizon classification. The well represented data that was used to make horizon classes includes CEC which along with base saturation, is important for fertility, organic carbon which is recognised as a huge store of the planets carbon (Padarian et al., 2012) and is fast becoming an official indicator of soil and environmental health which drives policy makers (Brandão et al., 2011). Soil horizon thickness is an easily acquired attribute as it could be obtained from the horizon depths. As it is already well established as a defining criterion in Soil Taxonomy (Schaetzl and Anderson, 2005), this soil property could be useful in soil horizon identification and was therefore trialled as an inclusion. Two analysis were performed, one with horizon thickness and one without. The performance of both data sets was evaluated. All of the data for this analysis was extracted from the USDA data set. Soil data with thickness information has been referred to as “Horizon classes”, while data without horizon thickness information has been referred to as “Material classes”.

3.2.3 Filling in the gaps
For numerical clustering algorithms such as FKM, it is necessary to collect a complete record with no gaps and the data cannot be overly correlated. Soil data was downloaded from the USDA database (National Cooperative Soil Survey, 2017), pedon information below 10cm extracted and collated. This data included sand silt and clay. One component, either sand, silt or clay needed to be removed
as it is compositional. This can be easily reconstituted post analysis. Colour is a factor that can be easily collected in the field, but it is typically recorded in the Munsell system. The Munsell colour system ® is commonly used for describing soils. It is a coordinate system which is arranged in a three dimensional prolate spheroid. The non-standard shape the data forms makes its use in quantitative analysis limited. The colours were therefore converted into the CIELAB system. A colour coordinate system that is standardized and orthogonal, and therefore more suited to numeric analysis. Data such as pH was mostly complete but the methods in which the pH was collected differed. There were four main methods in which the pH was collected: pH in a 1:1 solution of water, pH in KCl, pH in saturated paste, and pH in CaCl₂. Differing pH collection methods have been standardized in the past (Minasny et al., 2011). The pH in water was equivalent enough to pH in saturated paste to merge these two data sets together. The combined strength of the remaining values was under 1% of the total data pool and as such their influence was negligible. CaCO₃ was also included as it is important in the maintenance of pH. As this property was usually left blank if there was no CaCO₃ in the soil (making it difficult to determine if it was an absence of carbonates or an omitted test), and because carbonate affected soils tend to have a high pH, a rule was applied: if the data returned an absent value for CaCO₃ and the pH was below 7.5, then it could be presumed the absent value was actually zero. This allowed otherwise incomplete data to be included in the analysis. Other soil attributes were present in sufficient quantity without requiring adjustment or estimation. There were many other soil properties that would have been useful in the identification of specific horizon and soil types, but as the number of soil properties increased, the need for a complete data set necessitated the deletion of incomplete rows, thus creating a large corresponding decrease in the amount of available data- to the point where large scale data analysis was not possible. One possible solution to this problem is FKM with estimations based on an idea in (Li et al., 2004) which calculates means for missing properties within the clustering process.

3.2.4 Algorithm
The soil data was analysed using a modified version of FKM called akromeson (11,18) for data without thickness information and akromeson (12,23) for the data with thickness added (Hughes et
This is a technique that creates a series of fixed clusters from the positions of extreme points (known as end point clusters), and then allows FKM to resolve normally in the centre of the data field (creating central clusters). By doing this, the algorithm takes all soil data into account including what would be normally disregarded as extragrade (outlier) information. The central clusters include the majority of the data and end point clusters contain extreme data which still could be useful for classification purposes but may be missed by conventional clustering techniques (Hughes et al., 2014). The number of central clusters was selected by analysing a combination of the Fuzziness Performance Index (FPI) and the Modified Partition Entropy (MPE), a method found in (Odeh et al., 1990), while end point cluster numbers were automatically determined from the number of end points in the data.

### 3.3 Results

#### 3.3.1 Horizon classes and material classes created

The algorithm partitioned the material data set into 29 surface classes, 11 being central classes and 18 being end-point clusters. The horizon data set was partitioned into 35 surface classes, 12 central and 23 end points. The results are presented in the form of four tables; central and end point centroids for material classes and horizon classes. There is a distinction between end point classes and central classes. Generally speaking central properties are more representative of the global mean and there is a high membership. Most of the data falls into these central cluster categories. End point classes tend to exhibit extreme soil properties and have lower memberships. Typically this means that only a handful of groups are well represented in the data, but these classes are important, however, as they tend to require special care in management and are based on a single data point and therefore exist-as opposed to classes based on a mean of properties which represent a group but may not necessarily have a specific horizon which exactly matches the calculated average. As such, end point data may be indicative of a group that could be ubiquitous but are unrecognised simply because they are, as yet, un-sampled. The columns represent the names of the classes and the average value of each property that was analysed.
Table 9. Central material classes organised by sand fraction, then clay fraction. This table demonstrates the partitioning of US surface horizons. It demonstrates the vast changes in texture of these surface classes in reference to the other material properties.

<table>
<thead>
<tr>
<th>Key</th>
<th>Type</th>
<th>pH</th>
<th>CaCO$_3$ (cmol/kg)</th>
<th>CEC$_{NH_4}$ (cmol/kg)</th>
<th>Clay (%)</th>
<th>Sand (%)</th>
<th>Silt (%)</th>
<th>OC (%)</th>
<th>Hue</th>
<th>Value</th>
<th>Chroma</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta</td>
<td>c</td>
<td>5.7</td>
<td>1.4</td>
<td>11</td>
<td>11.2</td>
<td>64.6</td>
<td>24.2</td>
<td>1.4</td>
<td>7.50YR</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>gamma</td>
<td>c</td>
<td>5.6</td>
<td>0.7</td>
<td>13.9</td>
<td>11.3</td>
<td>63.9</td>
<td>24.8</td>
<td>2.4</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>alpha</td>
<td>c</td>
<td>6</td>
<td>1.4</td>
<td>10.3</td>
<td>10.4</td>
<td>61.3</td>
<td>28.3</td>
<td>1.5</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>eta</td>
<td>c</td>
<td>6.1</td>
<td>2.4</td>
<td>15.1</td>
<td>19.3</td>
<td>48.9</td>
<td>31.8</td>
<td>1.7</td>
<td>5.00YR</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>theta</td>
<td>c</td>
<td>7.8</td>
<td>3.1</td>
<td>18.8</td>
<td>20.4</td>
<td>39.1</td>
<td>40.5</td>
<td>1.8</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>epsilon</td>
<td>c</td>
<td>5.7</td>
<td>0.9</td>
<td>42.3</td>
<td>15.8</td>
<td>39.1</td>
<td>45.1</td>
<td>9.9</td>
<td>7.50YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>zeta</td>
<td>c</td>
<td>5.8</td>
<td>1</td>
<td>15.7</td>
<td>18.6</td>
<td>25.5</td>
<td>55.9</td>
<td>2.2</td>
<td>7.50YR</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>iota</td>
<td>c</td>
<td>6.7</td>
<td>0.8</td>
<td>20.7</td>
<td>21.2</td>
<td>22.2</td>
<td>56.6</td>
<td>2.8</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>kappa</td>
<td>c</td>
<td>5.2</td>
<td>0.8</td>
<td>22.1</td>
<td>24</td>
<td>21.6</td>
<td>54.4</td>
<td>2.7</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>delta</td>
<td>c</td>
<td>5.6</td>
<td>0.6</td>
<td>12.9</td>
<td>15.4</td>
<td>16.7</td>
<td>67.9</td>
<td>1.8</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>lambda</td>
<td>c</td>
<td>6.1</td>
<td>1.7</td>
<td>38</td>
<td>49.3</td>
<td>12.5</td>
<td>38.2</td>
<td>2.7</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 10. End point material classes sorted by sand fraction, followed by clay fraction. These surface end point material classes are calculated, sorted and displayed the same way in table 10. These tend to have some more extreme characteristics and are typically less represented in the data.

<table>
<thead>
<tr>
<th>Key</th>
<th>Type</th>
<th>pH</th>
<th>CaCO₂</th>
<th>CEC NH₄⁺</th>
<th>Clay</th>
<th>Sand</th>
<th>Silt</th>
<th>OC</th>
<th>Hue</th>
<th>Value</th>
<th>Chroma</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>H₂O</td>
<td>(cmol/kg)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(cmol/kg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>E</td>
<td>4.5</td>
<td>0.1</td>
<td>17.6</td>
<td>1.2</td>
<td>91.9</td>
<td>6.9</td>
<td>6.7</td>
<td>00R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>m</td>
<td>E</td>
<td>8.4</td>
<td>0.1</td>
<td>10.3</td>
<td>6.1</td>
<td>85.2</td>
<td>8.7</td>
<td>2.3</td>
<td>10.00YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>i</td>
<td>E</td>
<td>4.9</td>
<td>0</td>
<td>2.3</td>
<td>3.2</td>
<td>83.3</td>
<td>13.5</td>
<td>0.6</td>
<td>7.50YR</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>j</td>
<td>E</td>
<td>6.9</td>
<td>0</td>
<td>12.2</td>
<td>3.6</td>
<td>78.9</td>
<td>17.5</td>
<td>40</td>
<td>7.50YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>o</td>
<td>E</td>
<td>5.1</td>
<td>0</td>
<td>74.4</td>
<td>14.3</td>
<td>65.7</td>
<td>20</td>
<td>1</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>h</td>
<td>E</td>
<td>11.8</td>
<td>4</td>
<td>4</td>
<td>2.6</td>
<td>64.3</td>
<td>33.1</td>
<td>0.2</td>
<td>5.00YR</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>r</td>
<td>E</td>
<td>10.2</td>
<td>5</td>
<td>61.2</td>
<td>27.9</td>
<td>45</td>
<td>27.1</td>
<td>0.1</td>
<td>10.00YR</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>t</td>
<td>E</td>
<td>6.9</td>
<td>3</td>
<td>1.3</td>
<td>51.2</td>
<td>42.2</td>
<td>6.6</td>
<td>0.6</td>
<td>10.00YR</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>l</td>
<td>E</td>
<td>7.4</td>
<td>2.1</td>
<td>0.7</td>
<td>5.7</td>
<td>33.7</td>
<td>60.6</td>
<td>0.3</td>
<td>10.00GY</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>p</td>
<td>E</td>
<td>6</td>
<td>103</td>
<td>79.2</td>
<td>19.6</td>
<td>30.8</td>
<td>49.6</td>
<td>22.8</td>
<td>7.50YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>n</td>
<td>E</td>
<td>3.9</td>
<td>0</td>
<td>11.7</td>
<td>11.4</td>
<td>29.4</td>
<td>59.2</td>
<td>0.9</td>
<td>10.00R</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>s</td>
<td>E</td>
<td>5.9</td>
<td>0</td>
<td>169.3</td>
<td>29.2</td>
<td>23.9</td>
<td>46.9</td>
<td>44.6</td>
<td>10.00R</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>q</td>
<td>E</td>
<td>7.2</td>
<td>57</td>
<td>5.7</td>
<td>22</td>
<td>18.6</td>
<td>59.4</td>
<td>1</td>
<td>5.00YR</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>k</td>
<td>E</td>
<td>5.7</td>
<td>0</td>
<td>53.7</td>
<td>4.7</td>
<td>18.6</td>
<td>76.7</td>
<td>26.8</td>
<td>5.00YR</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>u</td>
<td>E</td>
<td>4.5</td>
<td>0</td>
<td>7.9</td>
<td>74.2</td>
<td>14.6</td>
<td>11.2</td>
<td>2.5</td>
<td>10.00R</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>w</td>
<td>E</td>
<td>7.1</td>
<td>0</td>
<td>74.6</td>
<td>90.5</td>
<td>0.8</td>
<td>8.7</td>
<td>16.2</td>
<td>7.50YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>v</td>
<td>E</td>
<td>4.9</td>
<td>53</td>
<td>18.8</td>
<td>80.7</td>
<td>0.8</td>
<td>18.5</td>
<td>1.1</td>
<td>10.00YR</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>f</td>
<td>E</td>
<td>4.7</td>
<td>0</td>
<td>22.7</td>
<td>0.1</td>
<td>0.1</td>
<td>99.8</td>
<td>10.6</td>
<td>00R</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 11. Central horizon classes organized by sand and clay fraction. Horizon classes are calculated and displayed in Table 12 using the same method as in Table 10, only with thickness added as a horizon property. As with material classes, they are sorted by texture.

<table>
<thead>
<tr>
<th>Key</th>
<th>Type</th>
<th>pH (H₂O)</th>
<th>CaCO₃ (Cmol/kg)</th>
<th>CEC_NH₄ (Cmol/kg)</th>
<th>Clay (%)</th>
<th>Sand (%)</th>
<th>OC (%)</th>
<th>Thickness</th>
<th>Hue</th>
<th>Value</th>
<th>Chroma</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>C</td>
<td>6</td>
<td>0.8</td>
<td>6.7</td>
<td>5.3</td>
<td>81.7</td>
<td>1.1</td>
<td>11.4</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>gamma</td>
<td>C</td>
<td>6</td>
<td>0.7</td>
<td>14.6</td>
<td>10.2</td>
<td>63.5</td>
<td>2.6</td>
<td>10.5</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>delta</td>
<td>C</td>
<td>6.2</td>
<td>3.2</td>
<td>10.4</td>
<td>10.1</td>
<td>59.9</td>
<td>1.2</td>
<td>8.8</td>
<td>5.00YR</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>eta</td>
<td>C</td>
<td>6</td>
<td>1</td>
<td>18.1</td>
<td>17.2</td>
<td>45.7</td>
<td>2.6</td>
<td>12.2</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>alpha</td>
<td>C</td>
<td>5.7</td>
<td>0.7</td>
<td>40.7</td>
<td>14.4</td>
<td>42.4</td>
<td>8.9</td>
<td>9.8</td>
<td>7.50YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>zeta</td>
<td>C</td>
<td>6</td>
<td>1.9</td>
<td>13.1</td>
<td>15</td>
<td>36.3</td>
<td>1.8</td>
<td>14.1</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>iota</td>
<td>C</td>
<td>6.1</td>
<td>1.2</td>
<td>27.5</td>
<td>26.6</td>
<td>29.7</td>
<td>3.8</td>
<td>13.6</td>
<td>7.50YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>lambda</td>
<td>C</td>
<td>6</td>
<td>0.9</td>
<td>17.8</td>
<td>17.6</td>
<td>23.3</td>
<td>2.6</td>
<td>12.3</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>epsilon</td>
<td>C</td>
<td>6.4</td>
<td>2.9</td>
<td>28.1</td>
<td>40.2</td>
<td>13.6</td>
<td>2.6</td>
<td>13.1</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>beta</td>
<td>C</td>
<td>5.7</td>
<td>0.5</td>
<td>13.5</td>
<td>17.3</td>
<td>9.6</td>
<td>1.6</td>
<td>17</td>
<td>10.00YR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>kappa</td>
<td>C</td>
<td>6</td>
<td>0.7</td>
<td>23.7</td>
<td>27.3</td>
<td>8.3</td>
<td>2.6</td>
<td>16.7</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>theta</td>
<td>C</td>
<td>6.3</td>
<td>2.1</td>
<td>43</td>
<td>53.2</td>
<td>7.4</td>
<td>3.5</td>
<td>14.1</td>
<td>10.00YR</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 12. End point horizon classes organized by sand and clay fraction. This table displays the end point horizon classes are sorted by texture as in the previous Tables 10-12. As they are end point classes they show some property deviation and tend to have lower memberships.

<table>
<thead>
<tr>
<th>Key</th>
<th>Type</th>
<th>pH (H₂O)</th>
<th>CaCO₃ (Cmol/kg)</th>
<th>CEC_NH₄ (Cmol/kg)</th>
<th>Clay (%)</th>
<th>Sand (%)</th>
<th>OC (%)</th>
<th>Thickness</th>
<th>Hue</th>
<th>Value</th>
<th>Chroma</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>E</td>
<td>4.7</td>
<td>0.1</td>
<td>2</td>
<td>1.1</td>
<td>96.8</td>
<td>0.1</td>
<td>1.3</td>
<td>10.00YR</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>j</td>
<td>E</td>
<td>5.3</td>
<td>0</td>
<td>4.5</td>
<td>1.1</td>
<td>95.4</td>
<td>2.2</td>
<td>6.1</td>
<td>00R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>q</td>
<td>E</td>
<td>7.4</td>
<td>0</td>
<td>3.8</td>
<td>0.9</td>
<td>92.8</td>
<td>0.3</td>
<td>14</td>
<td>10YR</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>m</td>
<td>E</td>
<td>5.1</td>
<td>0</td>
<td>8.8</td>
<td>3.7</td>
<td>83.9</td>
<td>1.6</td>
<td>47</td>
<td>00R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>n</td>
<td>E</td>
<td>6.9</td>
<td>0</td>
<td>19.9</td>
<td>0.5</td>
<td>80.3</td>
<td>4.8</td>
<td>35.9</td>
<td>2.50YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>t</td>
<td>E</td>
<td>4.9</td>
<td>0</td>
<td>39.9</td>
<td>1</td>
<td>77.9</td>
<td>18.5</td>
<td>5</td>
<td>00R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>h</td>
<td>E</td>
<td>8.9</td>
<td>1</td>
<td>5.4</td>
<td>8.3</td>
<td>67.8</td>
<td>0.3</td>
<td>22.9</td>
<td>7.50YR</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>ff</td>
<td>E</td>
<td>7.0</td>
<td>0</td>
<td>148.9</td>
<td>14.7</td>
<td>48.2</td>
<td>49.8</td>
<td>8</td>
<td>5.00YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>x</td>
<td>E</td>
<td>8.3</td>
<td>16</td>
<td>21.5</td>
<td>16.4</td>
<td>44.7</td>
<td>0</td>
<td>3</td>
<td>2.50YR</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>f</td>
<td>E</td>
<td>7.4</td>
<td>2</td>
<td>0.7</td>
<td>5.7</td>
<td>33.7</td>
<td>0.3</td>
<td>12</td>
<td>7.50Y</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>z</td>
<td>E</td>
<td>6.6</td>
<td>103</td>
<td>79.3</td>
<td>19.6</td>
<td>30.8</td>
<td>22.8</td>
<td>18</td>
<td>7.50YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>y</td>
<td>E</td>
<td>8.5</td>
<td>74.9</td>
<td>7.6</td>
<td>18.2</td>
<td>24.6</td>
<td>0.8</td>
<td>8.1</td>
<td>10.00YR</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>fg</td>
<td>E</td>
<td>6.1</td>
<td>0</td>
<td>14.4</td>
<td>26.4</td>
<td>24.5</td>
<td>0.1</td>
<td>160</td>
<td>7.50YR</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>l</td>
<td>E</td>
<td>7.0</td>
<td>0</td>
<td>7.5</td>
<td>6</td>
<td>23.4</td>
<td>2.7</td>
<td>5</td>
<td>5.00YR</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>v</td>
<td>E</td>
<td>9.8</td>
<td>4</td>
<td>43.6</td>
<td>55.8</td>
<td>23</td>
<td>0.9</td>
<td>61</td>
<td>00R</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>o</td>
<td>E</td>
<td>7.3</td>
<td>0</td>
<td>44.8</td>
<td>6.1</td>
<td>20</td>
<td>4.5</td>
<td>2</td>
<td>10.00YR</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>g</td>
<td>E</td>
<td>3.5</td>
<td>0.1</td>
<td>7.4</td>
<td>14.4</td>
<td>14.6</td>
<td>1.2</td>
<td>5.3</td>
<td>10.00YR</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>k</td>
<td>E</td>
<td>6.7</td>
<td>0</td>
<td>15.8</td>
<td>68.5</td>
<td>13.2</td>
<td>2.3</td>
<td>18</td>
<td>2.50Y</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>u</td>
<td>E</td>
<td>4.4</td>
<td>0</td>
<td>85.9</td>
<td>7.6</td>
<td>12.4</td>
<td>13.1</td>
<td>19</td>
<td>5.00YR</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>s</td>
<td>E</td>
<td>5.9</td>
<td>0</td>
<td>84.7</td>
<td>64.5</td>
<td>6.7</td>
<td>0.4</td>
<td>5</td>
<td>10.00YR</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>r</td>
<td>E</td>
<td>5.2</td>
<td>0</td>
<td>20.1</td>
<td>64.7</td>
<td>2.2</td>
<td>2.9</td>
<td>15</td>
<td>2.50Y</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>p</td>
<td>E</td>
<td>4.5</td>
<td>0</td>
<td>17.6</td>
<td>84.9</td>
<td>1.9</td>
<td>2.3</td>
<td>20</td>
<td>7.50YR</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>w</td>
<td>E</td>
<td>7.1</td>
<td>0</td>
<td>74.9</td>
<td>90.7</td>
<td>0.7</td>
<td>16.3</td>
<td>10</td>
<td>7.50YR</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
3.3.2 Comparison between material classes and horizon classes
As a simple measure of the amount of contrast between factors in the centroid tables, the ranges were calculated by subtracting the minimum values from the maximum values in each factor column excepting colour. The analysis of horizon classes demonstrated that for the central clusters there was less differentiation between pH, which ranged between 5.7 and 6.4, and CaCO₃ which ranged between 0.8 and 3.2, while there was greater differentiation of Cation Exchange Capacity (CEC) and texture. In the end point clusters there was little difference between horizon classes and material classes with the notable exception of pH, CEC and silt for which in each case the material classes demonstrated greater differentiation. As an example, the material end point pH ranged from 3.9 to 11.8 compared to horizon end point classes of pH of 3.5-9.8. In the case of colour, individual Munsell colours were counted. The horizon classes had more uniform central colours than the material classes but a wider range of end point colours. Most of the central horizon classes have a standard deviation of around 7cm for horizon thickness. This means that the maximum thickness of these centroids can be larger than thee means shown in the centroid table. Beta, for example has an average thickness of 16.9cm but a standard deviation of 6.7cm giving it a maximum average thickness of 23.6cm. This can imply that the fuzziness between each class is also large. One would expect as a result, that there is a correspondingly large amount of soil data at the boundaries between classes, where they will exhibit properties of two or three horizon classes in nearly equal measure. This demonstrates that many soil types regardless of the scheme used can be seen as a continuum (McBratney and DeGruijter, 1992). The practical result of this is a large amount of latitude when describing soil horizons.

3.3.3 Using dendrograms
It is possible to determine the differentiating soil properties of the cluster analysis using dendrograms. A dendrogram was therefore created using the Wards minimum variance method (Murtagh, F. & Legendre, P. 2014). Each branch in the dendrogram represents a soil property by which the data has been split. It is then possible to compare the split data by measuring the difference in means between all factors compared to the standard deviation, a procedure which is similar to the original Ward method or a t-test. The dendrogram of horizon classes is presented in Figure 10. In the case of both
horizon classes and material classes, texture is the primary differentiating constituent. The horizon classes have more differentiation at the texture level indicating that the horizon classes may be more suitable. This data was therefore plotted on the soil texture triangle (Figure 11).

Figure 10. A dendrogram demonstrating the priority set in the cluster analysis. In this figure, it can be seen that the primary properties are texture. The dendrogram demonstrates how each horizon class groups together. Also, critical properties that define each group can be ascertained from this analysis. The properties that most affect the definition of groups are identified at each node of the dendrogram. These nodes were identified by comparing mean values of each property in the centroid table.
Figure 11. Soil Classes plotted on a texture triangle. All the horizon classes plotted on the soil texture triangle. The blue Greek symbols represent the central clusters, the red alphabet symbols represent the end point clusters. Plotting the classes on the texture triangle (figure 11) shows how central and end point clusters tend to group. Appropriately, they tend to group more centrally, being less likely to have a disproportionate amount of sand, silt or clay.

3.3.4 Selection of data set
The material classes have increased spatial variability. This combined with the increased number of classes and the differentiation afforded to properties which are associated with ochric horizons make them most suitable. As this work is partially intended to assist in the differentiation of Ochric properties, material classes will be further studied.

3.3.5 Geographic representation
Each point from the analysis could be assigned a horizon class based on its mathematical distance from the centroids in the table (Table 12) and each position determined by using the x-y coordinates of the pedon the horizon came from. The results plotted over the United States were too sparse to make any conclusions so a method of creating a probability of occurrence map with the data available
was required. This requires several stages. The first is to first create a data density map, then a probability map for each class.

This is the procedure:

- Count the total number of observations \( (n_t) \) in a 10km by 10km moving window. Smooth this with a moving circular filter of diameter 600km. An example of this calculation can be found on the map in Figure 12.

- Perform the same procedure for each class to gain an estimate of classes \( (n_c) \), an example of this map can be found in Figure 14.

- The proportion of points \( (p) \) that are in each class is determined therefore by the equation:

\[
p = \frac{n_c}{n_t}
\]

In order to create a map indicating the maximum probability of a class in a given region of the United States, the raster of probabilities for each class can be converted into a table and the maximum likelihood of a class for each pixel extracted. These maximums can then be assigned a colour and plotted over the United States. The resulting map is presented in Figure 13.
Figure 12. Density plot of soil data based on a 600km window. The map is demonstrating a density profile of the soil horizon data in the United States. It is useful for determining where the most complete data is and also for determining gaps.

3.4 Discussion

3.4.1 What are the soil classes?
The soil surface horizon classes are essentially local averages of soil properties that would be found in a given region. From the dendrogram (Figure 10) it can be seen that the soil classes organise themselves into a few major groups. Gamma, mu and delta are associated by a high sand fraction. They are therefore represented on the main map (Figure 13) with different shades of the same colour. It can be seen that these surface horizon classes group together in the south west of the United States with some coastal areas included. The end point horizon class “I” also has a large average sand fraction. It can be seen that this class also has a representation in the south west of the United States (Figure 15). Alpha zeta eta lambda and iota are characterized by clay abundances that are on average under 20%. They have been coloured yellow/brown and tend to group in the North West of the United States. Beta is a class that is distinguished by extremely high silt content. It can, however be merged
with Kappa, epsilon and theta because, even though by comparison their silt content is much lower, they still have on average over 30% in their composition. These soil classes can also be grouped with H and G end point classes and can be most probably located in the middle to south east of the United States where a large proportion of loess soils are (Figure 15). These soil horizon classes have been assigned different shades of red in the central class map in Figure 13.

3.4.2 Central classes
The soil data density map displayed in Figure 12 demonstrates that data for soil surface horizons are abundant on the south west coast and in regions of the centre but are sparse towards the east coast and especially in Alabama, Georgia, South Carolina, Maine and Florida. Some of this data rarity is because of large water bodies such as Lake Michigan and Lake Superior to the north east, other reasons include the fact that this particular study collected specific horizons at the 10cm mark, therefore excluding any data sets that did not include the top 10 cm. The horizon class probabilities from that region would be subject to false positives to groups if an improbable soil happened to be sampled from there. The areas which had a point density of less than 200 samples per 500km² were therefore masked from the results. These are represented by white unlabeled areas. It can be seen that Florida and Maine are not well represented, but inclusion of these data in the smoothing window increases the likelihood of an accurate prediction in other regions of the United States. Low data density regions were masked out of the horizon class map, and are represented in white.
Figure 13. Probable distributions of surface horizon classes. This plot indicates areas in the United States where each soil horizon class is most likely to be found. The central horizon class Kappa is circled.

From Figure 13 it can be seen that there is a large amount of diversity towards the west of the continent, as would be expected in regions of mountains and faults, which contributes to soil variability along with other factors such as climatic conditions. Centrally there are larger, less diverse regions, as would be expected with less mountainous terrain. Towards the east, the surface landscape classes dissociate, these regions of diversity are associated with the Appalachian range and the lake systems around Michigan. One of the central soil classes, kappa, (circled) has been compared with what is known of the United States soils.

3.4.3 Kappa
The horizon class known as kappa (Figure 14) is a central class. Approximately 8% of all the surface horizons sampled fall into this category, making it one of the largest groups in the data set. The majority of these points are in an area commonly associated with loess soils. This horizon class is a silty clay loam, being composed of over 60% silt which is the composition of most loess surface
horizons. It is moderately acid with very low CaCO3 and a moderate CEC. It has high organic carbon content, its colour is 10YR 3/2 and at 16cm it is above average thickness relative to other surface horizons in the database. This is roughly comparable to the ST mollic epipedon, which also happens to be one of the most common horizon types in the USA.

![Map of the United States with a probability plot for the soil horizon class Kappa](image)

*Figure 14. Probability plot for the soil horizon class Kappa. This demonstrates an increased probability of encountering a kappa class in the upper mid west of the United States.*

### 3.4.4 End point classes

Creating a spatial representation of the calculated end points was a much more difficult task. The amount of data used to create the centroids was greatly reduced and much of it did not have geographical coordinates associated. As a result, although most data could be included in the calculations, they could not be subsequently plotted on a map of the United States, and therefore their distribution density could not be calculated. For relatively data rich central classes a few missing coordinates was not a problem, but it was sufficient to attenuate the spatial representation of the end point classes. The purpose of the end point classes is to reduce the effect of extreme data and pick up any residual clusters that may be written off as extragrades (Hughes *et al.*, 2014). This means that
some of the end points may not have large clusters associated. This is not to say that the other classes
don’t exist, it is simply a case of not enough data to plot a spatial distribution. The data was therefore
culled using two rules; 1) the classes needed a large enough proliferation of data to make a reasonable
spatial distribution, and 2) the classes needed a probability of occurrence greater than 10%. This
reduced the number of end point classes that could be represented spatially to four. The resulting plot
is presented in Figure 15.

Figure 15. Probable locations of soil surface end point classes. Only four were numerous enough for a reasonable
probability plot, the rest were excluded. End point class “G” is circled. This is a class that is high in silt with a
relatively low CEC but is notable for its extremely low pH (3.5).
3.4.5 End point class “G”

This end point horizon class occurs predominantly around Louisiana, Mississippi, Arkansas, Tennessee and Missouri. It is a silt loam and this class is distinctive because of its low pH. Strongly acid according to Murphey (2007). The horizon is also distinguished by very low CaCO₃ and low CEC. The organic carbon content is moderate and the colour is 10YR 5/4. It is unusually thin at just over 5cm, but numerically it is well represented for an end point class, bearing in mind that end point classes are always much lower in frequency than central classes. This class may be a product of erosion.

This peripheral soil class has some similarities and is geographically located in the proximity of the central class beta. Using Mahalanobis distances it was confirmed that this end class is mathematically the closest to central class beta, which is a neutral pH, moderately fertile soil. End point class G is much thinner and has a much lower pH than its central counterpart. Both erosion and acidification are known complications of agriculture, and the locations of this horizon class are either in farmland (on or around the Mississippi river) or East of the Mississippi delta. This could be evidence of land degradation from agriculture, or acidification that results from poor water management as can be seen in the Murray Darling river system (Glover et al., 2011). The low CEC can be associated with pH because in variable charge soils, the net CEC tends to decrease with a corresponding decrease in pH (Khawmee, et al., 2013).

3.4.6 Comparison with ochric horizons from Soil Taxonomy

One of the reasons for this analysis was to create some differentiation within the ochric and other surface horizon classes of Soil Taxonomy. Soil Taxonomy is a system which is concerned with entire pedons, so individual horizon descriptions are not equivalent, but there are several different horizon classes that are used in this scheme. These are divided into epipedon classes and diagnostic subsurface classes. Some of the horizon classes that were created numerically were therefore compared with the ochric horizon classes from ST.
3.4.7 Ochric horizons

Part of the classification of ochric epipedons requires knowledge of the overlying and underlying horizons, and some other information that is not present in this data set, but there are three indicators that can be extracted from the surface horizon clusters.

The surface horizon cluster data can identify horizons that are:
1. Less than 24 cm thick (dependent on solum thickness) and/or
2. Have an organic carbon content of less than 0.6% and/or
3. Have a colour chroma and value of greater than 3.5.

There are other, stricter defining criteria for ochric horizons (Soil survey staff, 2014), so without more ancillary data, the soil horizon classes identified here can only be described as ochric like.

3.4.8 Differences between systems

Using fuzzy classification has its own set of idiosyncrasies. The most prominent of these is the fact that the category that data falls in is determined by a distance ratio, rather than a hard or binomial cut off, sometimes referred to as taxonomic chop (Butler, 1980). Soil taxonomy can further subdivide classes to reflect other traits, but usually within a strict order of priority. There is therefore some extra leeway when describing a soil horizon numerically.

3.4.9 Identifying “ochric like” horizons

It appears that there is more diversity within the ochric category than has been recognized until now. In the range of central end point classes, the alpha, beta and mu horizons fit narrowly into the three ochric like criteria, but the majority of the end point classes (f,g,h,l,j,k,l,o,p,q,r,s,t,x,y,ff and fg) make a much better fit. Of the central classes that could be considered ochric as per the previously stated three conditions, mu was differentiated by its extreme sand content, beta by its extreme silt content and alpha by high organic carbon. It can be seen the colour hues for end point ochric horizons range over most of the Munsell spectrum, the colour values range from 2 to 7, the colour chromas range from 2 to 8. The thicknesses range from 2cm to 160cm and the organic carbon ranges from very high to extremely low. CEC values have similar tendencies. Textures can have high sand, silt or clay. The
colours are relatively uniform, all three having a similar Munsell colour of either 10YR 4/2 or 7.5YR 3/2. It is possible that these particular groups would have more affinity with other surface horizons in ST. The end point horizons have much wilder fluctuations in one particular area- colour. Of the end point classes determined to be ochric by the criteria described previously, 14 are differentiated primarily by colour, and within these there are 12 colours represented. It is possible that, if all the properties associated with an ochric epipedon were determined and attached, the diversity of colour would remain, and this diversity in the ochric horizon class could be exploited. Of the factors that could be used to further differentiate ochric horizons, the colours extend along a large range, and are easy to distinguish in the field. These therefore, may be a suitable metric to use, if the definitions of ochric epipedons were to be annotated.

3.5 Conclusions

3.5.1 Numerical analysis
There are advantages to be gained by analysing soil surface horizons in the USDA data set numerically. Using fuzzy distances rather than hard cut off values reduces the exclusion of soils which are otherwise similar to one class but are excluded by one moderately errant variable.

3.5.2 Class selection
Horizon classes are more suitable than material classes for two main reasons. The first is that the extra factor creates more variability to study and that variability is predominantly in the soil texture, which is the primary differentiating constituent according to a dendrogram using Wards minimum variance. The second being that soil thickness is an important characteristic in ST, making the classification much more compatible. This makes identification of horizons with ochric properties much simpler, and therefore easier to use for researchers seeking to improve this category. The horizon classes are therefore recommended. The 35 classes identified here should be tested for their taxonomic and management utility.
3.5.3 Probability map created
Soil surface horizons showed considerable differentiation. Some horizons were scattered over a large geographical area, so it was necessary to compare the density of horizon classes to the overall density of soil data if a meaningful probability map was to be created. This map demonstrated that there were soil horizon categories, which matched known soil properties for these regions.

3.5.4 Classes used for assessment of environmental damage
A few end point classes identified were similar to central classes but were different because of extreme properties. Some of these different properties could be associated with anthropogenic activities, something that could be studied in future work. End point analysis could be used in this way to identify areas that require special care.

3.5.5 Colour as a defining characteristic
Of all the factors that differentiated classes that had similar characteristics to ochric horizons in soil taxonomy, one of the most diverse separating features was colour. This large range of colours in the ochric class may be used as a guide for further compartmentalisation in the USDA.

3.6 Further work

3.6.1 Applying weightings
Numeric methods split data without consideration to factors that are important to stake holders. Also, soil properties that are known to have influence in pedogenesis can be obscured by less important properties. It is possible to semi-supervise clustering, identifying important factors and applying a weighting to them. By this the issue of taxonomic chop is not breached, yet issues that affect the productivity of horizons and pedologically important features can be addressed.
3.6.2 Analysing sub surface horizons

Studying surface horizons is important, but there is much detail to be gleaned from the underlying horizons as well. The subsurface B horizon is considered a master horizon (Buol et al., 2011), being used for classification and also its properties are important for agriculture, as typically the B horizon resides within the rooting zone of agricultural crops. It also accumulates minerals in the instance of more arid climate regimes, which is often diagnostic of certain soil taxa in ST. The same kind of analysis that was performed for the surface horizons should be performed for the subsurface horizons as well. It is intended that these horizons be analysed together and the centroids examined for commonalities that may or may not be associated with depth. ST has within its system the diagnostic criteria for many subsurface horizon types. These existing horizon classifications can be compared to numeric horizon classifications to assess empirically the rationale for each designation in ST.

3.6.3 Creating numerical pedons

It can be said that the most important unit of the soil is the horizon, but it is equally important to know how the soil horizons have developed with reference to the overlying and underlying layers. Most of the world’s taxonomies are based less on individual genetic horizons and more on diagnostic horizons, characteristics, and properties of the control section and entire pedons. If numeric horizons mirror the horizons created by more traditional taxonomies, numeric horizons could naturally coalesce in defined patterns that reflect pedogenesis. If so, important reference horizon assemblages could be identified numerically and used as a standard by which a soil profile description could be compared. World soil groups could be established on the basis of similarity to these reference taxa based on a critical set of properties. If soils can be identified this way it is an empirical confirmation of what experts in the field have known all along– and while a new taxonomy could emerge from this, it gives credence to the underlying principles of most of the world’s soil classification systems. This current analysis has created 35 soil horizon classes and the next analysis will no doubt create additional classes, so documenting each combination of soil horizons could result in millions of numerical pedons. There needs to be a method to efficiently organize numerically created horizon classes. Similar methods have been organized using computational systems such as OSAKA (Odgers et al.,
2011) but it may be possible to use a variation of FKM such as a modified version of the method that was used in the Hunter Valley, Australia (Malone et al., 2014), circumventing the need for additional software.

3.7 Acknowledgments

Many thanks to Jose Padarian who provided some of the programming, which enabled the creation of a probability map of soil classes.

3.8 References


National Cooperative Soil Survey Characterization Database

http://ncsslabdatamart.sc.egov.usda.gov/ Accessed Thursday, July 13, 2017


87


Chapter 4

Comparisons between USDA Soil Taxonomy and the Australian Soil Classification System

I: Data harmonization, calculation of taxonomic distance and inter-taxa variation

Summary

Soil classification is typically nationally based and as such consists of predominantly individual organizations, such as landcare NZ, the CSIRO in Australia and the USDA in the USA, creating locally meaningful categories for nationally important soils. This process has inevitably created a recognized disconnect between classification systems, and a push for a more universal classification has been proposed. To that end, numerical methods are explored. As a way of standardization between systems, soil taxa at the Great Group level from two separate regions and soil classification systems (Australia and the USA) are compared. Australia and the USA are represented by separate databases of soil profile descriptions (SPDs) comprising the same 23 properties at 18 depth intervals (no individual SPD is allocated to both systems). Taxa centroids from Soil Taxonomy (ST) and the Australian Soil Classification system (ASC) are calculated via principal component analysis and projected into the same property space and compared. The taxonomic relationships within and between ST taxa and ASC taxa are obtained by distance calculations. Convex hulls (a method of visualizing data in a plot) of each soil order of both systems are created and the associations each taxon has with other individuals in the same taxon discussed, as well as the variance. We determined that ASC orders have smaller overall dispersion compared with Soil Taxonomy and there is a greater probability of an ASC Great Group associating with another ASC Great Group of the same order, compared to the same occurring in ST. The influence of each property to the overall taxonomic distances is explored. This analysis opens the way for the possibility of comparing convex hulls of similar groups within differing taxonomies. A combined dataset has been created, and using
taxonomic distances between SPD’s allows objective comparisons of soil profiles and could pave the way for a more comprehensive classification method.

4.1. Introduction

4.1.1 Comparison of differing soil classification schemes

4.1.1.1 The soil profile description
The object of soil classification is to create simple concepts from what in reality are complex biogeophysicochemical systems. These concepts can then be transferred according to need and utility. The basic unit of soil classification in most cases is the soil profile description (SPD). Often referred to as a pedon, this is one of the fundamental sampling units in a soil scheme. It is an analysis of characteristics and properties of the soil, usually to describe a taxonomic feature or diagnostic element, organized by layers or depth increments from the soil surface to lithic contact or the maximum depth that the soil can be considered useful. Layers are normally referred to as horizons, and there is a semi standard method that describes horizons around the world, such as the FAO guidelines for soil description (Jahn et al., 2006), and ST with its own field guide (Schoeneberger, P.J., 2002), using alphabetical and numeric designations to delineate the sequence in which the horizons appear (alphabetical order is not strictly adhered to e.g. in the case of an E or O horizon, which is defined according to properties, rather than position, can overlie A or B horizons). SPD’s are married by properties to reference groups and often these groups have a descriptive key to which the user can refer to identify the group in which the soil profile description belongs. This allocation of SPD’s to groups with a known relationship to each other comprises a taxonomic system.

4.1.1.2 Soil Profile Description’s around the world
SPD’s around the world are defined with differing taxonomic systems of differing levels of complexity. According to the US Soil Taxonomy (Soil Survey Staff, 1960; 2010), SPD’s are organized into Orders, Suborders, Great Groups, Sub Groups and Family determined by a hierarchical
classification key. The internationally developed world reference base for soil resources (WRB) (IUSS Working Group, 2006) on the other hand, consists of one categorical level which can be split in two, the 32 Reference Soil Groups that are determined by the classification key, and second level of qualifiers. Both WRB and ST are comprehensive and internationally acknowledged, and the majority of research in soils or soil related matters requires SPD’s to be recognizable to one of these two taxonomies. Most countries however have their own systems, sometimes cloned from ST or WRB, sometimes derived from other systems, or developed independently. The currently applied national systems have similar or complementary goals but the format and priorities of each differing system make comparing individual taxa between systems difficult (Krasilnikov and Arnold, 2002). While most systems are internally comprehensive and logical, a question arises when an unknown SPD is introduced from another taxonomic system. Considering the needs of reproducibility in research or management, it is imperative to know how the taxonomic system of another country compares with either of the main two systems. Typically, researchers from countries that do not subscribe to either the WRB or ST have to “translate” taxa, or allocate their SPD’s into one of the more globally accepted taxonomies. Each of the lesser-known systems may be unique and possibly confusing to the uninitiated, but most have a logical structure and usually employ a hierarchical approach. A good example of a lesser-known taxonomy is the Australian system. The Australian soil classification system, or ASC (Isbell, 2002) is organized into Order, Sub Order, Great Group and Sub Groups and associated Family criteria. Comparing this with, for example, ST demonstrates the inherent difficulties researchers from different parts of the world have in making their concepts understood by a broader audience, and can also provide methods by which these difficulties can be bridged.

4.1.1.3 Soil Taxonomy and the Australian Soil Classification

Soil Taxonomy has a specific structure in each name. An “Endoaqualf”, for example, is an Alfisol by the three letter soil Order designation at the end, aqu, from the “aqua” and from the Great Group at the beginning. The ASC operates in a level of detail up to five tiers, but typically the first three tiers-Order, Suborder and Great Groups, provide sufficient detail for comparison with taxa at the Great Group level within ST. The definitions for each sub unit can be readily retrieved from the website:
Which has been a reliable source of taxonomic identification in Australia at the time of writing (2017). The major difference between the Australian system and any other classification systems is ASC’s focus on the use of field estimated properties such as texture to complete the classification. Australian sub soils are typically high in clay, so the system has evolved with greater attention paid to clay content and textural changes with depth (Harms, 2014). A good example of this is the three Australian soil orders Chromosol, Kurosol and Dermosol. These three orders have a roughly equal chance of being classified as Alfisol in Soil Taxonomy. Chromosols and Kurosols have texture contrast, Dermosols do not, Kandosols have low contrast and structure. Comparing the ASC to Soil Taxonomy is an interesting test bed for comparative taxonomy. The ASC is functionally quite different from Soil Taxonomy, which is a comprehensive soil descriptive scheme that is especially advantageous when used within the context of large amounts of laboratory or field description data. A typical allocation in ST is a one or two word designation, consisting of sets of common character sequences that elucidate certain properties. The ST Great Group name “Glossudalfs”, for example, contains a formative element for soil Order, Suborder and Great group, combined to make one word. It is a name which although laborious reveals a wealth of information such as moisture status, horizon thickness, depth and form at which illuviation has occurred all in the one word. The ASC on the other hand, uses a much more basic approach. There is a designation based on soil properties, but most of the information required for classification can be determined from the profile with a few simple tests. Soils can be differentiated sometimes to the Great Group level with a minimum of available laboratory information. This ease of classification is probably offset by loss of detail. A soil name such as a “Calcic Red Dermosol” can be determined by colour, presence of calcium and lack of texture contrast (which is not prioritized the same way in Soil Taxonomy), much less than the former designation, but can all be determined in the field. Structure is implied in its definition, but not necessarily used. The extreme difference in these two approaches raises the legitimate question of whether it is possible to compare soils identified with these two systems, and if it were possible, are there any additional groups in either system that would make useful additions to the other?
4.1.2 Development of a universal soil classification system

Most national soil classification systems were developed for local needs and use of the time of elaboration based on different observation methods, concepts and structures. Modern agriculture, and the understanding of global environmental, food security issues, knowledge exchange, however require comparable, harmonized soil information. Soil Taxonomy is the best documented and most widely applied system with a large accumulation of data and experience. It is necessary however to test whether it fulfils the needs of the wide range of current soil data users, and if it can accommodate all soils of the world to serve as universal system, and if not, which improvements can be made (Hempel, 2014). Soil Taxonomy must be assessed for completeness, and improved if necessary to a more comprehensive, simpler system that can bridge the gap between soils around the world. Addressing this issue courts the very real danger of adding another layer of complexity to world soil classification. In order to avoid this, any system that is adopted around the world needs to be quantitative, simple, repeatable, transferrable and complementary. An ideal approach therefore, is the investigation of numerical comparison and taxonomic distances. Knowing the taxonomic distance between two taxa gives an indication of its similarity without necessarily having to understand the method by which it was named. By way of example, Albeluvisols in the WRB are comparable to Alfisols in Soil Taxonomy. How does a Red Chromosol from Australia compare to these two? Chromosols are one of the soil orders that rely heavily on profile texture contrast, a feature that does not emerge readily in either WRB or ST. Using common properties derived from experimental data, it may be possible to determine the taxonomic distance between all of these groups. The shortest distance is the closest taxon.

4.1.2.1 On the application of taxonomic distance calculations in soil classification

The idea of using calculated taxonomic distances to express the level of similarity and dissimilarity between different soil taxonomic units was first applied in the 1960's (Hole and Hironaka, 1960; Bidwell and Hole, 1964a, 1964b; Sarkar et al., 1966) but only with local data and limited scope. Taxonomic distance calculations were revisited by Minasny and McBratney (2007) who incorporated
taxonomic distances into spatial prediction and digital mapping of soil classes. Minasny et al., (2009) derived taxonomic distances for the WRB Reference Soil Groups (RSGs) based on the presence and absence of key properties. Fuchs et al., (2011) studied the taxonomic relationship of soils of Hungary based on their dominant soil forming processes to provide numerical support to the improvement of the national system. Soil taxonomic distance calculations were also applied to study the correlation possibilities of different national soil classification systems to the WRB. The calculated distances supported the expert knowledge based correlations with objective measures of taxonomic relatedness of compared classification units of different systems (Lang, et al., 2016). Michéli et al. (2016) tested the distance calculation for comparing Great Group of Soil Taxonomy and concluded that the method is useful to in determining objective differences between soil taxa and improving classification definitions and criteria.

4.1.3 Objectives

Based on the promising previous results, the objectives of the first part of this two part paper are to extend the comparison of ST Great Groups with ASC relevant level units in the following steps:

1. To establish a method of standardization of soil properties in a SPD to facilitate data based comparisons between any two pedons or taxa.

2. To initiate a database of soil taxa and associated properties that can be used as a reference for world data

3. To determine if any soil properties that are not used to define soil taxa in ASC or ST are properties that are useful for differentiation purposes.

4. To determine the importance of individual properties in the distance calculations between individual taxa.

5. To determine if there are clusters of properties that can be used as a guide for classifications.
4.2 Methods

4.2.1 The Soil Taxonomy Great Group centroids

4.2.1.1 USDA data
The USDA Natural Resource Conservation Service and its predecessor entities have been compiling soil data from around the United States for more than a century. In recent times, the scope of data collection has increased to include some data from other countries. This large amount of information, although collected at different times by different people for different purposes, has a degree of standardization. Organized into pedons, split into horizon depths and individual properties, the data set is highly complex. Some of these data are missing in places, but it is typically analysed by similar methods and where appropriate recorded in the same units and each pedon is typically linked to a ST classification (National Cooperative Soil Survey, 2012).

4.2.1.2 Properties
These data have been the basis for previous analysis (Michéli et al., 2014) in which 19 distinct soil properties are used; CaCO₃, pH, CEC, Ca, Mg, Na, K, Acidity, base saturation, exchangeable sodium percentage (ESP), electrical conductivity, gypsum, rock fragment percentage (e.g. gravel), texture, colour, evidence of water, evidence of ice, bulk density and organic carbon. Colour was broken into the red, green and blue section of the visible spectrum, and texture was likewise diverged into sand, silt and clay content, creating a total of 23 individual variables.

4.2.1.3 Database structure
The majority of data was collected initially, and conventionally, by soil horizon. This approach allows for the identification of different horizon types, and contrasts in properties between horizons, and the associated depths and thicknesses which are the cornerstone of most soil classifications, but can prove problematic when analyzing big data. Comparing SPDs simply by horizon designation becomes unnecessarily complicated because the depths of respective A, B and C horizons may not match, and in certain circumstances the A horizon of one profile may have properties more in common with a B
horizon of another. Similarly the depth at which each horizon occurs adds complexity to the analysis. Comparisons between thin A horizons and thick surface B horizons, such as can happen in Austrailia for example, may only serve to confound numerical analysis. There have been a variety of methods attempted to solve this, such as dissimilarity cluster analysis (Leblanc et al., 2016) and probabilistic representation (Beaudette et al., 2016). This approach yielded good results in Canada for potato fields.

4.2.1.4 Depth function using splines

The solution applied to horizon boundaries and depths in this circumstance is the removal of horizon designations and application of depth functions using equal-area splines (Malone et al., 2009) to all the soil properties at standardized depth intervals. This method also allows for comparisons of depth differences between complementary horizon characteristics, and it allows for gradational changes of all properties within a single horizon (Pinheiro et al., 2016), advantageous as soil horizons are typically part of a continuum (McBratney and Gruijter, 1992; Hempel et al., 2016). Using spline functions (Bishop et al., 1999), data representing 403,954 SPD’s, collected at varying depths was transformed into set depth increments of 0-5cm, 10-15cm, 15-20cm, 20-25cm, 30-35cm, 35-40cm, 45-50cm, 50-60cm, 60-70cm, 70-80cm, 80-90cm, 90-100cm, 100-110cm, 110-130cm and 130-150cm. Each pedon therefore, has a value for each property at each depth increment; a total of 415 individual properties. The data is then organized according to the taxon designation given to the SPD. There were 218 individual taxa represented at the Great Group level for Soil Taxonomy. The mean values of each property at each depth were then calculated for every individual taxon, resulting in a table of 218 taxa with 23 properties at 18 depth increments- a comparison table with 109,145 individual entries. When orders and suborder centroids are included, there are a total of 299 centroids representing a large proportion of the great groups that are expected in the United States. The orders and suborder centroids, although interesting are not at a level of detail which produces meaningful comparisons so typically they are left out of distance calculations, which does not impact the results of the study, but can be used as reference points as the situation demands.
4.2.2 Australian Soil Classification System Great Group centroids

4.2.2.1 The database
The existing US database is organized in such a way that provided the data is in the correct format, a comparison between a SPD and the compilation of centroids is possible. In Australia, the terrestrial ecosystem research network (TERN), an organization combining the research from the University of Sydney, CSIRO, Geoscience Australia and several other federal agencies has accrued a large range of soil data. These have been matched with soil designations, often up to the Great Group level, in a similar manner to the National Cooperative Soil Survey (NCSS) Soil Characterization Database. This data set is large, comprising of well over two million individual samples, from 285,572 SPD’s, from all over Australia. This similarity to the NCSS database means the data can be treated in a similar fashion and the properties used to numerically compare, for the first time soils from the Australian system with soils from the USA. Although the data has similar properties, a combination of geographical factors, time and localized issues has resulted in some differing methods for data collection. Compounding these issues, the International Union of Soil Sciences (IUSS) working group has not been able to resolve differences in laboratory procedures (Hempel et al., 2013). Great care needs to be taken identifying the methods and translating them in a manner that is meaningful for the ST reference centroids.

4.2.2.2 Extracting properties
Of the soil data, properties such as rock fragments, organic carbon, CaCO₃, pH, CEC, Ca, Mg, Na, K, Acidity, base saturation, ESP, EC, gypsum and organic carbon were extracted from the TERN database directly. The exact tests and the circumstances in which they are required are provided in Peverill et al., (1999). Wherever possible, the same methods were extracted, but this was not always feasible. Some data needs to be collected by special means if certain conditions e.g. high pH apply. Such changes to methodology increase accuracy, but are not necessarily consistent with the ST database. The fragments data was recorded as an ordinal value between 0 and 6. These were converted to their corresponding percentages by weight as found in (McDonald et al., 1998). CEC
data is dependent on factors such as clays dominated by variable charge, which requires e.g. compulsive exchange method, one of the many obscure chemical tests found in Peverill et al. (1999). Soil pH is typically recorded as an extract of a 1:1 soil water in the USA. The typical Australian method is with a 1:5 soil water extract. This mismatch was addressed using a regression from Libohova et al. (2014). Some properties such as texture, colour, evidence of water and ice and bulk density required a more innovative approach.

4.2.3 Properties that require processing

4.2.3.1 Processing texture
Extracting Australian texture data is obfuscated by two issues: The first is the fact that exact sand silt and clay fractions are not always recorded in the data. Texture class is therefore extracted in lieu of percentages. The mean values for each Australian texture class has been determined in previous research (Minasny et al., 2007). These means were used rather than the centre of values in the texture triangle as it represents a more true representation of particle size distribution. There is also unconformity between standard particle sizes for sand and silt (Marshall, 2003). The size range of silt in the US is 0.002mm - 0.05mm, compared to Australia which has a silt size range of 0.002 - 0.02mm. The Australian particle size was harmonized with the US data using the formula (Padarian et al., 2012):

\[
P_{2-50} = 2.26P_{2-20} + \frac{5.55P_{2-20} + 2.26(P_{2-20})^2}{0.996 - 1.236P_{2-20} - 1.349P_{2-2000}}
\]

Where:

- \( P_{2-50} \) is the adjusted silt size according to the USDA
- \( P_{2-20} \) is the original silt size according to Australia
- \( P_{2-2000} \) is the clay size
4.2.3.2 Processing colour
Munsell values were converted to three orthogonal properties. The individual colours are referenced and extracted from conversion tables in RGB and these values are then divided by 256. This is a procedure that harmonizes that data with the US centroid data set. The final result can be studied in either Munsell, red green or blue (RGB) or the orthogonal CIE LAB system (Torrent and Barron, 1993). CIE LAB is superior for numeric analysis.

4.2.3.3 Processing evidence of water and ice
The best possible match with the USDA centroid data corresponds with the NZ guidelines for soil classification. This standard defines gley (wet) soils using the assumption that water is usually present if the colour of the soil has a chroma of less than two (Hewitt, 2010). Colour data is therefore mined from the TERN database and all colours in the USDA data that matched this standard were considered to be under the influence of water. The TERN database also contains samples with redoximorphic features such as mottling and presence of manganese. These data can be used to provide a much more rigorous picture of the presence of water in Australian horizons. If however, this updated method is applied to Australian data without a parallel update to the existing US data, the increased proportionally of water evidence in Australian soils over the US can skew the analysis. The goal in this instance is to have as close a methodology as possible to the USDA data set provided, so the NZ method alone is used on the Australian data. Ice is a rare event in Australian soils, this figure was set to zero. There is data present from colder regions of the US, such as Alaska. In the case of this study, there is no need to establish a matching methodology. In future work, in profiles containing evidence of ice or where ice is suspected, other harmonizing methodologies will be required.

4.2.3.4 Processing bulk density
There is a small amount of bulk density data extant in the TERN database. Extending this data would increase the accuracy of calculations. In order to make the data as similar as possible to the ST data set, 18 pedotransfer functions were calculated from the Soil Taxonomy centroids at each depth increment using regression equations comparing BD with clay fraction and carbon:
US/AUS pedotransfer functions

\[
\begin{align*}
BD_{0-5} &= 1.576464564 + 0.005284245C - 0.211117792 OC \\
BD_{5-10} &= 1.556413898 + 0.005625197C - 0.208098212 OC \\
BD_{10-15} &= 1.551962823 + 0.005542375C - 0.211732860 OC \\
BD_{15-20} &= 1.532097988 + 0.005890699C - 0.208184639 OC \\
BD_{20-25} &= 1.518313062 + 0.005930441C - 0.205081067 OC \\
BD_{25-30} &= 1.51823700 + 0.00576151C - 0.20557758 OC \\
BD_{30-35} &= 1.504141067 + 0.005686418C - 0.200424225 OC \\
BD_{35-40} &= 1.499302067 + 0.005388116C - 0.189648473 OC \\
BD_{40-45} &= 1.483359891 + 0.005555584C - 0.182995526 OC \\
BD_{45-50} &= 1.478273215 + 0.005479181C - 0.175017836 OC \\
BD_{50-60} &= 1.489494186 + 0.005143243C - 0.177244460 OC \\
BD_{60-70} &= 1.486105256 + 0.004901262C - 0.163698298 OC \\
BD_{70-80} &= 1.504697296 + 0.004456061C - 0.166083714 OC \\
BD_{80-90} &= 1.514973571 + 0.004466581C - 0.182197499 OC \\
BD_{90-100} &= 1.53307909 + 0.00402969C - 0.19932879 OC \\
BD_{100-110} &= 1.535289163 + 0.003839931C - 0.200966371 OC \\
BD_{110-130} &= 1.505680119 + 0.004242295C - 0.208614447 OC \\
BD_{130-150} &= 1.475724040 + 0.004687231C - 0.210462076 OC
\end{align*}
\]

where:

BD = bulk density,
C = clay, and
OC = organic carbon

This pedotransfer function recognizes that all other factors being equal, BD increases with depth due to overburden pressure and decrease in OC. Missing Australian bulk density data was then estimated according to these functions. As these pedotransfer functions were calculated from the US data, this ensured that the bulk densities were as analogous to the US data as possible. The consequence of this was that US soil and Australian soils being very different results in some bulk densities below 0.4.
This would not be expected in Australian soil profiles. In these instances, values that were close to zero were recalculated using a pedotransfer function (Adams, 1973) evaluated by (Tranter et al., 2007):

$$\rho_p = \frac{100}{(\text{OM}\% \rho_{OM}) + (100 - \frac{\text{OM}}{\rho_m})}$$

Where:

OM% is the organic matter percentage (OC*1.72),

$\rho_{OM}$ is the organic matter bulk density and

$\rho_m$ is the mineral matter bulk density.

This uses organic carbon and texture to estimate bulk density for any problematic taxa but it presumes however, that the conditions over Australia are homogenous and the model functions optimally when the organic carbon is below 5%. This can affect Australian organic soils (Organosols). The presence of peat and histic materials are a very rare event in Australia, so procedures typical in the United States, such as organic matter stratification are not necessary. The combination of the two pedotransfer functions resulted in a data set of bulk densities within a range of 0.4 to 2.0.

4.2.3.5 Depth-function harmonization

In order to harmonize with the ST centroids, soil profile information optimally should be collected and set at standard depth intervals to 150cm. As the majority of the data is collected by horizon, which have no fixed depth, equal-area splines as performed by Lang et al. (2013) as described in section 2.2 are used. For further detail, the procedure is documented in previous literature (Bishop et al., 1999; Malone et al., 2009; Odgers et al., 2012). Soil reference taxonomy (SRT) data is splined by horizon and averaged at 5cm increments from 0-50cm, 10cm increments from 50 to 110cm and 20cm increments from 110cm to 150cm. Standardization of the data for comparison can be hampered by the depth to which samples are collected (Figure 16). There are two main reasons why depths differ - differences in equipment/experimentation and lithic contact. Much of the data was collected by an automated coring machine which extracts samples to a depth of 1m. Other experiments and
instruments can collect cores or samples to differing depths, and manually collected samples are variable. For the examples where samples were simply not collected at depth, the splines can be extended to obtain this information. Lithic contact is harder to account for. These kinds of issues are handled via principal component analysis (section 4.2.3.8). Many soils contain extremely hard clays so Australian coring equipment can sample paralicthic layers as well as penetrate densic layers and fragipans. Splines can miss-represent data at either end of the function. This will be addressed in the next few chapters. 4.2.3.1- Density analysis

**Depth function harmonisation**

Data comes as a horizon-based description

Spline function applied

Data converted to match the depth sequence of the SRTs

---

![Diagram of depth function harmonisation](image)

**Figure 16.** Depth function harmonisation. The conversion of horizon based data into common depth increments with the use of a spline function.

### 4.2.3.6 Calculation of centroids

Soil properties are extracted from the TERN data according to Great Groups and the mean value for each property calculated per taxon. This procedure was performed for each tier of the classification system; Order, Suborder and Great Group (doing this creates a level of redundancy within the dataset). If the mean for a particular Great Group is not available due to lack of data, it is then possible to substitute this with the mean from the categorical level above, replacing missing data with the mean of the next highest tier in the classification makes it possible to include over 470 complete
Great Group centroids to the data. In future revisions, it would be possible to append these means with more data. Once calculated, the data is organized by properties and depth (Table 14).

Table 13. Example of the layout of one of the properties—rock fragment percentage for vertosols. This layout continues at regular intervals to a depth of 150 cm. The final values in this table have been abridged so it can be easily viewed.

<table>
<thead>
<tr>
<th>VERTOSOLS</th>
<th>1.0-5cm</th>
<th>1.5-10cm</th>
<th>1.10-15cm</th>
<th>1.15-20cm</th>
<th>1.20-25cm</th>
<th>1.25-30cm</th>
<th>to 150cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravel vol%</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

These properties are aligned side by side for each centroid, creating a vector 415 elements long (Table 14), with the great group name included, a labelled 414 properties.

Table 14. Example of how all the properties are concatenated into a single vector for each taxa profile description. This table is representative of a much larger table, which in its totality consists of 299 soil profile allocations with 23 properties at 18 depth intervals. A section representing a summary of the first 6 properties and the first 5 rows is displayed here. This table continues, however for a further 17 properties, making 23 in total. Each property is further partitioned into 18 depth increments. This is demonstrated in Table 14. The gray section here is where Table 14 fits. This is repeated for every property and every soil taxonomic allocation.

<table>
<thead>
<tr>
<th>Gravel</th>
<th>Clay</th>
<th>Silt</th>
<th>Sand</th>
<th>Red</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertisols</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
</tr>
<tr>
<td>Ultisols</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
</tr>
<tr>
<td>Spodosols</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
</tr>
<tr>
<td>Oxisols</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
</tr>
<tr>
<td>Mollisols</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
<td>1-150cm</td>
</tr>
</tbody>
</table>

When combined this way, each property at each depth interval can be compared with the properties and depth intervals of other taxa or SPD’s. A summary of the exact procedure is presented via a flowchart (Figure 17). The result is the total reference taxonomic database (TRex).
4.2.3.7 Direct profile comparison

With the data collected and harmonised by depth, the individual soil descriptions can be observed and compared as a series of splines (Figure 18, Figure 19, Figure 20, Figure 21, Figure 22, Figure 23).
Figure 18. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for gravel, silt, sand, and clay. Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset.
Figure 19. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for the colours red, green and blue. Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset.
Figure 20. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for pH, calcium carbonate (labelled “carb”), CEC and Ca. Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset.
Figure 21. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for water, ice, bulk density and Organic Carbon (OC). Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset. Typically ASC soils do not have any evidence of ice, but it is a useful property for soils in other regions of the world.
Figure 22. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for sodium, magnesium, potassium and acid saturation (acidity, labelled “acid”). Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset.
Figure 23. Properties of the ASC taxa compared to the properties of the ST taxa, presented as a series of splines. This is a typical comparison of two taxa within the master dataset for base saturation, exchangeable sodium percentage, electrical conductivity and gypsum. Any of the soil profile allocations or SPDs can have any properties compared graphically in this way. This is the simplest method of visually assessing any of the 414 properties embedded in the dataset.

### 4.2.3.8 Principal component analysis and Euclidean distances

Soil data, once splined, estimated and calculated can be compared via an appropriate distance metric. Much of the data does not cover the same data ranges. Principal component analysis based on the correlation matrix between properties can be used to both reduce data dimensionality and standardize each variable. Euclidean distances in principal component space represent approximate Mahalanobis distances. Inversion of the 414 by 414 element correlation matrix is difficult to perform on this data set. The principal component analysis package in the statistical programming language “R” calculates the components by eigenvalues, and PCA itself can also be used to objectively determine the
properties that are contributing to the differences between soil types. Additionally, when making comparisons between soils, the 414 attributes produced from 23 distinct properties broken down into 18 depth increments makes the data unwieldy. PCA identifies within these multiple variables, a smaller number of key dimensions that explain the influences of the larger data set. In this instance, 28 principal components explain 90% of the variation in the data; the first two explain 30% A full breakdown of the influence of properties to PC’s is available (Table 17).

4.2.3.9 Missing data
Not all soil taxa described have all of the soil properties recorded, some of the SPD information is unavailable due to experimental constraints, other times by physical barriers such as lithic contact. For distance calculations, horizon properties need to be compared at each depth increment, which means some properties at lower depths need to be compared to horizon properties that do not exist. In order to solve this, several solutions were applied. The properties and the depths were extracted and standardized. They were sorted by Order, Suborder and Great Group. The mean of each property for each group was calculated. This mean could then be used in lieu of a missing property. For other occasions such as where lithic contact indicates there should not be a property to compare, the issue can be solved mathematically in scaled PCA space by replacing the missing variable with the mean of that variable at the closest taxonomic level.

4.2.3.10 Density analysis
The centroids being compared come from two databases; TERN and USDA. Although there has been much effort to keep the methodologies consistent in terms of data collection, the sheer amount of data, the number of personnel who collected it, the various processes involved in the cataloguing and storage of the data and the time scales involved can increase the chances of differing results. The major systematic issues are associated with the spline function. Although splines are representative of a data population, they can result in a multi modal distribution, which appear as high values at the maximum and minimum of the data. Other systematic issues such as a particular data series using the wrong units, or insufficient smoothing, create artificial data densities at particular depths. These can
have an impact on the PCA and can be identified with density analysis. Although the data for each continent has different attributes, the distributions remain relatively consistent. It is therefore possible to use density plots (Figure 24, Figure 25) to compare taxa from differing countries or data sets and employ them to find and correct some inconsistencies with the data. Density analysis is useful because it can be standardized—i.e. all of the values, regardless of the disproportionate number of taxa per group, sum up to 1 so abnormal density can be identified regardless of the proportionality of representation, advantageous, as there is roughly twice the amount of Australian to US centroids.

**Figure 24.** Density plot for carbon comparing Soil Taxonomy and Australian Soil Classification. Although there are vastly different carbon percentages in each group, the density distribution is consistent between groups. The mean of organic carbon in Australia is around 1%.
Figure 25. Density plot for Soil Taxonomy and Australian Soil Classification demonstrating a systematic error in the clay fraction of the Australian dataset. This particular error has artificially high numbers of zero, which produce humps at the beginning of the plot.

4.3 Results of Comparative Taxonomy

4.3.1 Centroid comparison

4.3.1.1 Distance between and mapping of centroids in the two systems
A total of 475 centroids representing ASC Great Groups were calculated from the TERN data and then combined with the 218 ST Great Group centroids. Principal component analysis performed on this combined data set. Each principal component is scaled by mean and standard deviation and explains a percentage of the variance. In this data set, 60% of the variance is explained by six components, and 95% explained by 36, which is more manageable than the original 414 variables. In order to visualise the data, the two largest components, PC1 and PC2 (which represent 30% of the variability) are typically plotted against each other. These explain the largest proportion of the
variability that exists in all 414 soil properties, which works for visual assessment. Calculating
distances properly requires at least 36 components. Past this number, the probability that the
components are simply representing noise in the data set increases.

4.3.1.2 Soil Taxonomy in Principle Component space
The principal components of each soil taxon are assigned to an ST order, plotted against all the other
soil taxa, in principal component space, the convex hull and variance calculated. This is displayed on
each of the following plots. The soil taxa are represented in red and the hull is the shaded area. As can
be seen in the diagrams (Figure 26), points from other taxa exist within the convex hulls. The hulls
themselves represent the boundaries for where taxa at the group level exist in taxonomic space. The
axis of the plots are standardized and scaled to accommodate the ranges of both ST and ASC. In the
plots, note Histosols, the soils that demonstrate the most variance, and also are the only group that
have nothing but their own taxa within the convex hull.
Figure 26. Convex Hulls of every Order in Soil Taxonomy. This figure shows all of the soil taxonomy orders converted into principal components and plotted using PC1 and PC2. Convex hulls demonstrate the relative positions in which each of these groups reside.

For these plots:

NTS = Entisols, NDS = Andisols, LLS = Mollisols, LFS = Alfisols
LTS = Ultisols, STS = Histosols, ELS = Gelisols, ODS = Spodosols
IDS = Aridisols, PTS = Inceptisols, RTS = Vertisols, OX = Oxisols.

4.3.1.3 Australian soil classification Principle Component Analysis plot
The same procedure was performed for soils in the ASC (Figure 27). The ASC taxa are more numerous, but also more compact. The variance is also recorded above each order to give an
indication of the dispersion of the groups. As with the US data, taxa from other groups reside within the hulls of each order, but it can be seen that these groups occupy a zone in PC space. Organosols are one of the only groups that demonstrate any exclusivity, only having one or two taxa within the hull that do not belong to the Organosol group. For the purposes of the next illustration (Figure 27), the two letter codes have been translated into the full ASC names:

AN = Anthroposols, CA = Calcarosols, CH = Chromosols, DE = Dermosols
FE = Ferrosols, HY = Hydrosols, KA = Kandosols, KU= Kurosols,
OR = Organosols, PO = Podosols, RU = Rudosols, SO =Sodosols,
TE = Tenosols, VE = Vertosols
Figure 27. Convex Hulls of every Order in the Australian Classification System. All of the Australian soil classification Great Groups plotted according to their orders with their corresponding convex hulls as with Figure 16 demonstrating where they reside in PC space. Each Great Group is plotted in grey, Australian Great Groups are represented by circles, USA Great Groups represented by crosses.
4.3.2 Taxonomic precision and accuracy of Soil Taxonomy and the Australian Classification System

4.3.2.1 Inter taxa variation
The Euclidean distance matrix for all the original ST taxa in PC space was calculated. From these data the individual minimum distances between taxa from ST and ASC were obtained. These minimum distances identify taxa, which have the sum of the most common properties at each depth interval. This calculation treats every depth interval as a separate property. It does not make associations with neighbouring intervals or similar properties. Each distance is made from principal components, and the number of components is typically cut down to decrease complexity. The disparity in variances between the two systems are partially due to the fact that the Australian taxa are more numerous, but numbers alone do not explain the large difference. This requires further investigation.

Table 15. Soil Taxonomy Orders and their variance. The average variance is 50.21.

<table>
<thead>
<tr>
<th>Order</th>
<th>Variance</th>
<th>Order</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertisols</td>
<td>22.74</td>
<td>Alfisols</td>
<td>17.37</td>
</tr>
<tr>
<td>Histosols</td>
<td>222.21</td>
<td>Entisols</td>
<td>66.79</td>
</tr>
<tr>
<td>Gelisols</td>
<td>86.89</td>
<td>Inceptisols</td>
<td>33.07</td>
</tr>
<tr>
<td>Ultisols</td>
<td>7.29</td>
<td>Spodosols</td>
<td>24.52</td>
</tr>
<tr>
<td>Aridisols</td>
<td>33.67</td>
<td>Oxisols</td>
<td>6.96</td>
</tr>
<tr>
<td>Andisols</td>
<td>68.82</td>
<td>Mollisols</td>
<td>12.17</td>
</tr>
</tbody>
</table>
Table 16. Australian Soil Classification Orders and their variance. Average variance is 10.82

<table>
<thead>
<tr>
<th>Order</th>
<th>Variance</th>
<th>Order</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthroposols</td>
<td>19.23</td>
<td>Organosols</td>
<td>10.11</td>
</tr>
<tr>
<td>Chromosols</td>
<td>15.78</td>
<td>Podosols</td>
<td>4.26</td>
</tr>
<tr>
<td>Dermosols</td>
<td>12.12</td>
<td>Kurosol</td>
<td>3.99</td>
</tr>
<tr>
<td>Calcarosols</td>
<td>11.44</td>
<td>Sodosols</td>
<td>8.36</td>
</tr>
<tr>
<td>Ferrosols</td>
<td>7.21</td>
<td>Tenosols</td>
<td>6.92</td>
</tr>
<tr>
<td>Kadosols</td>
<td>14.72</td>
<td>Rudosols</td>
<td>10.53</td>
</tr>
<tr>
<td>Hydrosols</td>
<td>17.67</td>
<td>Vertosols</td>
<td>9.18</td>
</tr>
</tbody>
</table>

The conversion of soil information into comparable components makes it possible to visualise the groupings of each Order in PC space. The plots are the result of centroid calculation from the two distinct data sets which are combined. The data is scaled and both ASC and ST are plotted over the same range to demonstrate their relative spread. It can be seen that the ASC taxa are more compact in their association. The variance of each soil taxa at the Order level has been calculated (Table 16, Table 17). The ST Order variances are much larger than those of the Orders in the ASC. Using Euclidean distances, it is possible to objectively measure how well each Order groups together. This is the taxonomic accuracy, defined by the proportion of soil Great Groups that have the next nearest neighbour in the same taxonomic Order. In order to do this, the following method was utilised:

- Create a distance matrix of all the soil centroids in PC space \((D)\)
- Pair all of the taxa with the shortest Euclidean distance

Taxonomic accuracy can then be calculated using the equation:

\[
Ta = \frac{n(sp)}{n(di_{min})}
\]

Where:

- \(Ta\) is the taxonomic accuracy,
- \(n(sp)\) = The number of taxa pairs that consist of the same Order,
- \(n(di_{min})\) = The total number of taxa pairs.
When calculated, Soil Taxonomy has a taxonomic accuracy of 0.75 and the Australian classification system has a taxonomic accuracy of 0.94, which means:

- For each Great Group in The Australian Soil Classification system belonging to order x the nearest Great Group is in the same soil order x, 94% of the time,
- For Great Groups in Soil Taxonomy belonging to order y, the nearest Great Group is in the same soil order y, 75% of the time

Although the ASC has a high accuracy compared to ST, this number still indicates a large number of Great Groups in either system do not have another taxon of the same order as their nearest neighbour. The precision of this method also needs to be contemplated. In this context, precision can be considered to be a measure of the overall distance between soil taxa in a given group. The data for Soil Taxonomy ranges over a larger volume of PC space (Figure 28). This means that although individual taxa in Soil Taxonomy associate well with their own kind, the level of dispersion in PC space is higher, suggesting that ST orders are more general concepts. The implications of how this affects global soil analysis can only be understood as data from more countries and classification systems are compared with Soil Taxonomy centroids.

Data dispersion of ASC vs ST
4.3.2.2 Adding Australian Soil Classification System Great Group centroids to Soil Taxonomy
By positioning the ST and ASC centroids together in PC space, it is possible to visualize the relationships of the two systems for the first time. Great Group centroids for the two taxonomies, having been combined and principal components extracted, are plotted in PC space using the first two principal components, coloured according to the taxonomic system to which they belong (Figure 29).

Figure 28. Data dispersion of Australian Soil Classification versus Soil Taxonomy. This is a plot demonstrating the range of points in ST as opposed to the ASC.

Figure 29. Soil Taxonomy Soil Taxonomic Allocations plotted against Australian Soil Classification Soil Taxonomic Allocations in 3d PCA space. Australian ASC classification GG centroids (black) plotted against US ST GG centroids (red) in principal component space. The labels “allscores” refers to each principal component i.e. allscores1 is pc1.
There is a great amount of diversity in both the US data set and the Australian one, but there is a zone in which the ASC Great Groups tend to occupy, which is within the range of the ST centroids (Figure 29). Soil Taxonomy is capable of covering a wide range of soils, but it can be seen, however that the area in which the Australian data sits is relatively unpopulated by ST taxa, which indicates there is scope for improvement and void reduction of Soil Taxonomy. It could be argued that creating a world classification system would be possible if the starting framework was a system such as Soil Taxonomy, with SPD’s and taxa from other countries helping to flesh out any gaps and expand the universe. The area in which the majority of ASC centroids fall is one occupied mostly by Entisols and Alfisols (this makes sense as there is a high probability that most of the ASC texture contrast soils will key out as Alfisols in ST).

4.3.3 Principal components and soil properties
Each principal component is a linear combination of soil properties, and with 414 soil properties, the contribution each property provides at each depth increment is small. The largest contributor is still only a fraction of the contribution at any given time, nevertheless, it is important to understand soil profiles in terms of pedology and attributes. To that end, a list of mathematical influences to the individual principal components has been constructed from their rotation values (proportion of variance explained) in PC space. These in turn have been combined at all depth intervals for each property to give a single number for each property. This way, the list of potential properties is reduced from 414 to 23, but the importance of depth is removed from the equation. When this technique is applied, the most influential three properties for the most influential component, PC1 are Ca, sand (texture), and CEC. All the properties and their associated correlations to PC1 and PC2, which are the components, in which the convex hulls were created, are presented (Table 18).

The absolute values of both components are summed to demonstrate which properties would create the most movement in PC space. From this it can be seen that the most influential properties in this particular model are ion saturation (acid or base), organic carbon, calcium and magnesium. In a way,
these two principal components are determining if the soil has mollic or mollic like characteristics, and the proportions of the ions in the CEC.

Table 17. Cumulative variance calculated from rotational values of principal components. This table shows each soil property and their cumulative variance for PC1 and PC2.

<table>
<thead>
<tr>
<th>Property</th>
<th>PC1</th>
<th>PC2</th>
<th>Sum PC's</th>
<th>Property</th>
<th>PC1</th>
<th>PC2</th>
<th>Sum PC's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidity</td>
<td>-0.97</td>
<td>-1.41</td>
<td>2.38</td>
<td>Sodium</td>
<td>-0.97</td>
<td>0.61</td>
<td>1.58</td>
</tr>
<tr>
<td>OC</td>
<td>-1.18</td>
<td>-1.12</td>
<td>2.30</td>
<td>Blue</td>
<td>0.66</td>
<td>0.82</td>
<td>1.48</td>
</tr>
<tr>
<td>Ca</td>
<td>-1.70</td>
<td>-0.60</td>
<td>2.30</td>
<td>Clay</td>
<td>-0.46</td>
<td>1.01</td>
<td>1.47</td>
</tr>
<tr>
<td>CEC</td>
<td>-1.29</td>
<td>0.81</td>
<td>2.09</td>
<td>Esp</td>
<td>-0.20</td>
<td>1.12</td>
<td>1.32</td>
</tr>
<tr>
<td>Mg</td>
<td>-1.19</td>
<td>0.84</td>
<td>2.03</td>
<td>silt</td>
<td>-0.89</td>
<td>0.36</td>
<td>1.25</td>
</tr>
<tr>
<td>Bsat</td>
<td>-0.47</td>
<td>1.52</td>
<td>1.99</td>
<td>EC</td>
<td>-0.50</td>
<td>0.64</td>
<td>1.15</td>
</tr>
<tr>
<td>K</td>
<td>-1.01</td>
<td>0.93</td>
<td>1.94</td>
<td>Frags</td>
<td>0.80</td>
<td>-0.16</td>
<td>0.96</td>
</tr>
<tr>
<td>pH</td>
<td>-0.36</td>
<td>1.56</td>
<td>1.93</td>
<td>Ice</td>
<td>-0.44</td>
<td>-0.40</td>
<td>0.84</td>
</tr>
<tr>
<td>Red</td>
<td>1.11</td>
<td>0.73</td>
<td>1.83</td>
<td>Gypsum</td>
<td>-0.13</td>
<td>0.42</td>
<td>0.55</td>
</tr>
<tr>
<td>Sand</td>
<td>1.34</td>
<td>-0.49</td>
<td>1.83</td>
<td>Water</td>
<td>-0.24</td>
<td>-0.13</td>
<td>0.37</td>
</tr>
<tr>
<td>Green</td>
<td>0.98</td>
<td>0.83</td>
<td>1.81</td>
<td>CaCO₃</td>
<td>0.11</td>
<td>0.12</td>
<td>0.23</td>
</tr>
<tr>
<td>BD</td>
<td>0.67</td>
<td>1.12</td>
<td>1.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using this broad assessment of principal components to soil properties over all depths can gain some basic understanding on how soils of the ASC liken or differ from soils from ST, but determining how each component responds to each property at each depth increment is desirable. In order to create some clarity on how depth functions in this model, the influence each property has of PC1 has been plotted over depth (Figure 30, Figure 31, Figure 32, Figure 33). This provides context without being overwhelmed by extraordinary dimensionality of the data. Taxa that require further investigation can then be cross-referenced with individual components to depth, visually, as necessary. If further analysis is required the specific numbers can be mined and converted into properties.
Figure 30. Cumulative variance of gravel, clay, silt, sand, red and green on Principal Component one calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component one at depth increments of 0cm to 150cm.
Figure 31. Cumulative variance of blue, water, ice, bulk density, organic carbon and carbonates on Principal Component one calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component one at depth increments of 0cm to 150cm.
Figure 32. Cumulative variance of pH, CEC, Ca, Mg, Na and K on Principal Component one calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component one at depth increments of 0cm to 150cm.
Figure 33. Cumulative variance of acid saturation, base saturation, Exchangeable Sodium Percentage, Electrical Conductivity and gypsum on Principal Component one calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component one at depth increments of 0cm to 150cm.

The same procedure has been employed for the second component (Figure 34, Figure 35, Figure 36, Figure 37). This second component explains 14% of the variability, while component 1 explains 15%. These two are similar in importance. In both PC 1 and 2, the influence of gypsum and silt to the model increases with depth, while the majority of the other soil properties tend to behave in opposite directions. This is in concert with the orthogonal nature of principal components.
Figure 34. Cumulative variance of gravel, clay, silt, sand, red and green on Principal Component two calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component two at depth increments of 0cm to 150cm.
Figure 35. Cumulative variance of blue, water, ice, bulk density, organic carbon and carbonates on Principal Component two calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component two at depth increments of 0cm to 150cm.
Figure 36. Cumulative variance of pH, CEC, Ca, Mg, Na and K on Principal Component two calculated from rotational values. These plots demonstrate the influence of the properties on Principal Component two at depth increments of 0cm to 150cm.
4.3.3.1 Soil properties in relation to depth

The influence of soil properties in PC space to depth is worth considering. As seen in the diagrams, these important variables contribute to the soils relationship to other soils in different ways at different depth increments. This has been alluded to when Minasny et al. (2007) were calculating bulk densities, where depth was considered to be an important factor. For most properties, the influence decreases with depth, but for some, such as colour and Ca, which can be considered to in part represent carbonates, the influence peaks half way down the profile. For others, such as water and
gypsum, the influence increases with depth. This relationship can be attributed to things such as uniformity of the surface horizon, especially under agricultural conditions making variations of features such as sand and ion availability stand out in mathematical modeling. Conversely, features at depth such as water table and parent material make presence of gypsum and sodium prominent at the lower depth increments. This is borne out in real life when a soil is named according to the position of certain horizons at depths, whether there is an underlying concentration of minerals etc. (Fajardo et al., 2016) indicates that when mathematical techniques are used to cluster horizons and horizon boundaries, there is some harmony between human observations and the clusters created. So in many cases, what has been advocated by expert opinion has a level of empirical truth.

4.4 Discussion

4.4.1 Data size
When assessing the soil data from two large continents, analysis is difficult due to data overload. Principal component analysis has proved itself as a useful method of isolating important variables from the pool of information. Data dimensionality is also improved, as 414 variables are reduced to the 36 most important components. Basing the analysis on existing taxonomies also removes some of the burden as masses of data can be condensed into the means of known taxa, reducing millions of rows into a few hundred. The result of these two processes is the reduction of a data set millions of rows deep and 415 columns long to a smaller data frame of 36 columns and around 700 rows. This makes analysis less time consuming.

4.4.2 Comparing taxonomies
PCA demonstrates that coalescing large amounts of data from two differing pools can produce results that are in the same data space. The ASC and ST centroids can be seen in relation to one another, and owing to the standardization, accurate Euclidean distances between all taxa can be calculated. Comparing profiles with truncated depths has been demonstrated to be possible via PCA. Australia
has been shown to be different, the most likely explanation is that it being a much older soil/continent, it is more textually differentiated.

4.4.3 Taxonomic accuracy
The association between Orders in each taxonomy have been determined. In the case of ST, using 36 principal components, Orders associated with a taxa of the same Order 75% of the time while in ASC this occurred 94% of the time. This phenomenon has been tentatively referred to as taxonomic accuracy, and may be a useful measure on how well a particular taxonomic system categorizes its soils. There are slightly differing results if different numbers of principal components are used. Ensuring that the effect of the properties is captured without harnessing some random variability can be done by removing the last n components. Accuracy can also be changed by adding or removing combinations of components. Such optimisation may be worth exploring in the future.

4.4.4 Important variables
In isolating variables, some detail has to be lost. It is noted that according to (Minasny et al., 2007), depth is an important factor. The analysis performed took into account depth, but teasing this out of the stream of data is difficult. Broad principles however, can be extracted. Some of the most influential factors can now be ascertained from a pure numerical perspective. These properties, derived from soil types as delineated by experts in the field, are organic carbon, CEC, base saturation, the presence of Mg and the presence of calcium. These are properties commonly associated with the mollic epipedon in ST. If these properties emerge from a relatively unbiased mathematical process, then the importance of the ST mollic horizon should be considered for other taxonomical systems. In the case of Australian soils, a mollic epipedon would be uncommon. This raises the question of the creation of a mollic-like horizon, or are there horizons in the ASC that perform the same purpose e.g. ASC melanic horizon.
4.5 Conclusions

With the techniques used it has been possible to establish a method of standardization, by using a variety of regression, composition and data harmonization techniques. This can enable two data sets to be compared. It was also possible to establish a database of soil taxa and associated properties, using the pre-existing soil taxonomy centroids and adding Great Groups from the ASC. The full range of properties in each SPD was also established for the ASC and the existing averages extracted from the USDA NRCS database, and this was used to make comparisons over 23 soil properties- a more extensive process than the more narrow criteria used to differentiate taxa. The most influential soil properties were able to be determined by examining the PCA rotation values. These influential properties represent a cluster that can be used as a guide for navigating the ASC and ST. Furthermore, using Euclidean distances in PC space, it is possible to test these data sets and determine how well the individual taxa coalesce with each other at the primary level (in this case both primary levels are Order). Secondly, by using convex hulls it is possible to mathematically and visually demonstrate the differences between Orders within and between taxonomic systems. Thirdly, with PCA it is possible to project into multi-dimensional space each taxon and demonstrate the relationship each taxon has with every other taxon that has been projected. Finally, by using basic concepts such as which taxonomic system the data belongs to, this projection can demonstrate broad differences.

4.6 Further work

In chapter five, standardization methods and convex hulls are used to determine how groups within a taxonomy of one nation compare with the taxonomic groups of another. Considering that comparing one group with another is no simple matter, there are 14 ASC orders and 12 ST orders which create a comparison matrix of groups that is immense. The complete analysis must be complimented with a rigorous but simple way of understanding it.
4.6.1 Visualization
Convex hull analysis can be made between the two taxonomic systems. Comparing all the components is mathematically complex, so using the two most influential components is both quicker and easier to visualize. The analysis of the distance between means of Orders within each taxonomy, plus the measurement of overlap between individual Orders will be presented and discussed.

4.6.2 Properties in relation to differing taxonomies
The issue of whether this larger set of soil properties is more robust than the properties used in current taxonomies will be addressed. There is an increasing body of work that calls into question the set of properties that are commonly used in taxonomic assessment (Nagy et al., 2016) If so, then work should begin on the creation of a global taxonomy based on a wider, standardized array of soil properties. PCA has already demonstrated its ability to sum a large complicated system to the influence of a few elements, this will be done at the Order level, further elucidating the ways in which the individual parts of each taxonomy function and interact.

4.6.3 Creating comparable groups
Hulls and distances were used to determine which of the 14 Australian orders are comparable to the ST orders. In doing so, the continental differences are addressed. It is common knowledge that silt is typically lower in an Australian SPD, this assumption need to be carefully compared with the data to ensure its veracity. Addressing this issue also assists in compensating for the difference between countries, and determining if there should be a differing order, or identifying if there is a different set of national properties within what is essentially the same soil type.
4.7 References


Isbell, R., 2002. The Australian Soil Classification, CSIRO Australia, Collingwood, VIC, Australia.


Chapter 5

Comparative taxonomy between the USDA Soil Taxonomy and the Australian Soil Classification System

II: Comparison of Order, suborder and Great Group taxa for Soil Taxonomy and Australian Soil Classification

Summary

Soil classification systems over the world are incongruent as they are based on different tiers and different sets of properties. This chapter is concerned with understanding the relationships of each tier (Order, Suborder and Great Group) in both Soil Taxonomy (ST) and the Australian Soil Classification system (ASC) using mean nearest neighbour distances and convex hull areas in two principal component dimensions. It is determined that in most instances, convex hull comparisons using only two principal components, representing 30% of the variation in the data describe much of the variability between and within the orders, suborders and Great Groups of each classification system. These are useful for visual comparisons of taxa at various levels, however, mean nearest neighbour distances can include all 415 variables if necessary, which is more rigorous but complex. Both distance calculations and convex hulls highlight the same associations between taxa from ASC and ST. Both these methods demonstrate the robust nature of Soil Taxonomy and the Australian Soil Classification, but with convex hull sizes that are smaller, and nearest neighbour distances that are lower, the ASC proves to be slightly more coherent. The two systems occupy somewhat different areas in PC space, and ST covers a larger overall area, demonstrating that ASC is a purpose built classification for Australian conditions while ST is a more general system that can cover a wider variety of soils and management issues. This study also demonstrates that Great Groups in ST are at about the same level of taxonomic generalisation as suborders of the ASC. Combining the best elements and taxa of both these systems would be a positive step in the creation of a comprehensive system.
5.1. Introduction

The previous chapter is a roadmap, which leads the way to data harmonization between two disparate classification systems, which are typically defined by a range of different factors. The significant achievement from this is twofold: 1) It results in the creation of TRex (a combined database consisting of data from both the ASC and ST), 2) This enables the calculation of taxonomic precision and accuracy, which is a measure of how well each taxonomic system defines its individual sub-units. This had never been attempted before, partially because data collection is not necessarily consistent between nations, partially because differing data sets from differing classification systems emphasize different aspects of soil, typically pertaining to important local issues, and partially because each different classification scheme is constructed differently, making a harmonious definition between SPDs and taxa difficult. The solution, which allows the initiation of TRex, involves an ST centroid database as a starting point. This database has mean values of soil properties of ST taxa, described to the Great Group level, arranged according to depth along single vectors. This creates a dataset that is 299 rows (one row for each taxa) and 415 columns (one column for each of 23 properties at 18 standard depth increments). This data can be compared with data from other taxonomies, so long as the properties and depths are consistent. Data which could be used for each of the 23 properties at a variety of non-standard depths was taken from Australian soil profile descriptions (SPDs) in the Terrestrial Ecosystem Research Network (TERN) database. With this, each of the 23 required properties at the 18 standard depth increments could be estimated and splined. The mean values for this data at each tier of the ASC were calculated. TRex is the amalgamation of this data and the original ST centroid set. With TRex, side by side comparisons of ASC and ST taxa at the Order, Suborder or Great Group level could be made via a metric such as Euclidean or Mahalanobis distances. Standardization was achieved by the use of scaled principal components. When converted, the 415 soil properties at depth could be represented by a lower number of components. The first two components represent 30% of the variability, 36 components represent 95% of the total variability. Accurate Euclidean distances can be rapidly achieved on 36 components as opposed to the original 415 columns as the overall size of the data is reduced by 90% without a corresponding loss to data quality. With these distances, the accuracy of each taxonomy could be obtained. The way in which
this was done was by the creation of a distance matrix of ST and ASC, pairing the taxa with the shortest Euclidean distances and obtaining the number of Great Groups which had Great Groups of the same Order as their nearest neighbour. This was expressed as a ratio of the total number of Great Groups within the Order- taxonomic accuracy. It was determined that ASC taxa have a higher taxonomic accuracy.

5.1.1 Grouping taxa numerically

5.1.1.1 Intra Distance analysis
Distance calculations are recognized as a method of understanding taxonomic features in soil. Distances are a key component of cladistics analysis (Miltenyi et al., 2015) but can also be used for a variety of other analysis. The distances between objects have already been used to calculate taxonomic precision and accuracy. Other aspects of distance can be used to ascertain details on the construction and integrity of taxonomic systems. The nearest neighbour distance between principal components of properties of soil taxa can be used to determine the closest taxon to another. Useful for comparisons on a case-by-case basis, nearest neighbour distances lose meaning as the number of individual comparisons increase. The mean value however, can be used as a measure of the coalescive ability of a given group. It can be inferred that a suborder from a given taxonomy fits into a defined space if the nearest neighbour distance is small. There may be a limited amount of predictive power as well. The taxonomic analysis has been performed to the Great Group level, because to elucidate further would be impractical owing to data availability. Both taxonomic systems studied, however, have further tiers to which SPDs can be classified. ST classifies to Order, Suborder, Great Group, Sub Group, family and series (Service, 1975). ASC classifies to Order, Suborder, Great Group and Sub Group (Isbell, 2002). Beyond this the level of family technically exists but it is rarely used or defined. In most cases this level would have more in common with the individual soil profile. With sufficient data on the upper three tiers, some predictions on the average nearest neighbour distances for these lower levels may be possible.
5.1.1.2 Inter Distance analysis

Intra distance calculations can provide details on the construction of the taxonomy, inter distance calculations can provide details on similarity. A distance matrix between means of groups such as soil taxonomy Orders against ASC Orders can be calculated. The smallest Order to Order distance for a given order is the taxonomically closest taxa. Comparing Orders alone (12 for ST and 14 for ASC) provides 168 distances. These can be ranked numerically from smallest to greatest, creating a snapshot of which ST at the Order level has similarities to ASC Orders. While being quick and comprehensive, these calculations being possible on up to the full 415 components, a single number for a distance can be a blunt instrument. Other methods can be used to reveal the nature of connections between soil groups and how all the properties interact.

5.1.1.3 Intra hull analysis

With TRex operational, inter and intra taxonomic distances can be used to create a picture of the performance of each taxonomy on a global scale. Convex hull analysis is used for a variety of applications from image correction to mapping (Barber et al., 1996). This simple technique can also be used in assessment of taxonomic systems. The simplest method of visualizing convex hulls is by imagining what a particular group of data would look like if it were plotted on an x-y axis and a rubber band were stretched around the data points. The points that touch the rubber band, and the shape which the band takes is the convex hull (Fadel et al., 2001). The creation of TRex has provided a universe in which soil profile descriptions (SPDs) can occupy. A convex hull can define the areas in this universe in which a SPD of a particular taxa tend to reside. A logical soil taxonomy system should be hierarchical. If this is the case, then all data pertaining to a particular taxonomy should occupy a large hull, with smaller hulls corresponding to the first tier of the classification being completely encapsulated within the primary hull. Hull taxa from the next tier down should be completely contained within the hull of the first tier, and the third tiers contained by the second. Graphically the setup should resemble Babushka dolls (Matroschka dolls in WRB), a series of hulls each within the other (Figure 45). The Matroschka structure should reflect the priorities of the hierarchy; one hull should be within another and the size of the hulls should decrease with tier
number. If a Suborder can exist independently of the Order, then the convex hulls would overlap two or more Orders. Such a scenario can be realized with simple suborder definitions such as colour in the ASC and climate in ST. Colour for example, should not cross boundaries as it can be influenced by a variety of different mechanisms (Simonson and Boasma, 1972; White, 2013). A Chromosol is typically higher in clay than a Podosol. The colour present in Podosols is typically from mineral accumulation down a profile, while the colour in the Chromosol more likely from weathering of the in situ minerals. Rationales such as this can demonstrate why colour should only be considered within the bounds of the Order level in ASC. Further research is required to determine if a link between simple properties and soil types can be made in the case of climate regime.

![Graph showing All Australian SPDs (black) with US GGs (red)](image)

**Figure 38.** All Australian SPDs overlaid with US centroid data in 2D (left) and 3D (right). This figure demonstrates both Soil Taxonomy allocations and all Soil Profile Descriptions in the TERN dataset overlaid together in 3D PCA space. This demonstrates the total area in which both taxonomies inhabit.

### 5.1.1.4 Inter hull analysis

TRex is a database, which places taxa from two separate systems into the same property space. Convex hulls can therefore be used to determine the similarity of groups and tiers between taxonomies. The rationale for such an analysis is simplicity and visuality. There are 299 taxa from ST in TRex if Orders and suborders are included, 475 ASC taxa are similarly included. A distance matrix
of all taxa is immense. Simplification by using only the Orders (12 in ST and 14 in ASC) creates 168 potential comparisons. The overall distance can be interrogated for pertinent properties and depths, but interactions of groups of Suborders between Orders and Suborders becomes time consuming and complex. Convex hulls provide a more simple comparison. Overlap area can be determined numerically, and large overlaps can be visually assessed. Conspicuous overlaps can be attributed to the proportion of hull 1 or the proportion of hull 2, and hull plots are useful even if no overlap exists. The proximity of hulls can sometimes indicate if a group has any similarities with another, and such commonalities can be picked up visually. The method is however, constrained by dimensionality, such visualizations being difficult in anything greater than three dimensions, which would capture less of the total variability in PC space.

5.2 Aims
This chapter aims to:

- Calculate mean nearest neighbour distance for Order, suborder Great Groups in Soil Taxonomy and Australian Soil Classification System
- Estimate hull sizes for Order, suborder and Great Group between taxa in Soil Taxonomy and Australian Soil Classification.
- Estimate overlap of Orders between ST and ASC, and determine any inferences from this.

5.3. Method

5.3.1 Soil Profile Description centroids
The centroids in TRex were calculated using TERN data, which was extracted and splined into the same format as the USDA centroids, which were already supplied. These were then merged together. In order to complete some of the calculations, SPD data was required. All soil properties were extracted, splined and merged with their unique pedon identification number along with their classification to the Great Group level. Any missing data was then extracted from the TRex database
for that particular Great Group. This provides 7366 SPDs, which can be traced to their Great Group. These were added to TReX and PCA performed.

5.3.2 Nearest neighbour distances
A calculated SPD or a taxon in TReX is represented by a single row. Organized into 23 properties at 18 depth increments, each row has a total length of 414. Any SPD or taxon can therefore be compared numerically with another, provided the data is standardized. The nearest neighbour distances between all SPDs and centroids were calculated using Euclidean distances of principal components. Typically 36 components capture 95% of the variability, but in this instance it is relatively inexpensive computationally to make calculations on all components. These are organized into a distance matrix (TReX distance matrix or TDM). Overall nearest neighbour distances were achieved by extracting the smallest distance for each SPD from TDM (lowest number of each row that is not zero). The mean of these was calculated. For Great Group and Suborder, there was SPD data available for ASC calculations. No equivalent for ST was available. TDM was sorted by Great Group and Suborder respectively, then any distances pertaining to SPDs of that classification level excluded. The minimum distance for each SPD was then extracted, and the mean for all these distances calculated. For Order calculations, the distances between the means of the Orders were calculated and the shortest distance for each Order extracted. The mean of these distances were calculated.

5.3.3 Convex hulls
The calculation of convex hulls is commonly done in a variety of settings. Convex hull area can be calculated using a simple function and the Splanks package in R. The hull identification procedure is performed on the first two principal components. Calculations of hulls in more than three components are difficult to visualize, convex hulls become mathematically expensive to calculate as the number of dimensions increase and overlap size and occurrence reduce as the number of components increase, rarely occurring beyond five. The optimal configuration therefore is two dimensions. This represents 30% of the total variability.
5.3.3.1 Hull sizes to taxonomic level
Soil centroid data was subset by Order, the hulls extracted and area calculated. This procedure was repeated for Suborder and Great Group. The ASC data has been calculated from TERN data so there are 7366 SPDs at the Great Group level, which enables calculation of ASC Great Group hulls. For ST the 299 centroids have one example for each Great Group, with no supporting SPD data. Convex hull calculations for ST can therefore only be performed at the Order and Suborder level.

5.3.3.2 Estimation
The area of hulls and the mean nearest neighbour distances can be calculated for Order and Suborder in most cases, and in some cases can be calculated for Great Group. TRex is a composite database, which offers extra flexibility for these calculations as there are two extra levels that can be added; hulls and distances of each nation and hulls and distances of the composite data set. Using these larger levels, trends can be plotted and regressions calculated. From this, the levels that are not represented by the data can be estimated.

5.3.3.3 Overlap
With convex hulls calculated, their overlap can be established. The procedure involves overlaying the two hulls of interest and identifying the area of intersection. The area can be quantified by a standard computer polygon calculation, which involves using the vertices as the basis for a series of triangles, for which the area of each individual triangle is calculated and summed. Hull sizes of suborders are small, and overlaps scarce, so only Order overlaps were calculated.

5.4. Results

5.4.1 All Soil Profile Descriptions
Using TERN and TRex data, it is possible to fully populate 8000 SPDs. When combined with TRex, and plotted in principal component space, the relationship between SPDs from Australia and taxa from ST are revealed. In this instance all taxa and SPDs from both taxonomic systems are plotted
together. The taxonomy of origin is identified by colour. Three principal components can be represented in three dimensions (Figure 38).

![Mean nearest neighbour distances according to classification system](image)

**Figure 39.** Mean nearest neighbour distances. This plot demonstrates the mean nearest neighbor distances between STAs at each tier of each classification. Instead of considering Great Groups, for example, to be on the same taxonomic level for both classification systems, the taxonomic tier of similar distance is considered to be the equivalent operational taxonomic level. The closest intersections are represented with black circles and demonstrate that ASC Suborders are equivalent in nearest neighbor distance to Soil Taxonomy Great Groups.

### 5.4.2 Nearest neighbour distances

The nearest neighbour distances decrease on a linear scale for both taxonomies. Australian distances are considerably different from US nearest neighbour distances. It can be seen, for example, that nearest neighbour distances on the Order level for Australia correspond to distances in ST that are analogous to Suborders (Figure 39). ASC Suborders match ST Great Groups. It can be said therefore, that based on nearest neighbour distances, the ASC Suborders are of a similar taxonomic
generalization to that of Great Groups in ST. Using regression, the next tiers can be estimated. Extending the line to the next tier is a point beyond all samples collected, the theoretical soil population. The nearest neighbour distance for this is 3.8 taxonomic distance units. The variability of taxa at the most basic level in ST could be at around 3%. This implies that the average sampling of all data in the ASC corresponds to the family tier in ST. This is a reasonable assumption as the usable SPD data for the ASC at this level is relatively low, corresponding to about 10 samples per Great Group. 97% similarity is what pedologists suggest is the taxonomic species level. This corresponds to a soil taxonomic distance of 3.0 or less.

![Convex hull areas by classification system](image)

![Logged convex hull areas by classification system](image)

Figure 40. Hull areas according to taxonomy natural (left) and logged (right). This figure demonstrates that the overall area in which each taxonomic Order occupies for each classification system are approximately the same.

### 5.4.2.1 Nearest Neighbour Distance (NND) tables

Next nearest neighbours (Table 19) can be used for calculating the similarity of two data entities empirically. They produce a single number to represent proximity in data space, which is good for rapid assessment, but are disadvantageous in the computational effort required to convert back to properties, and the single dimensionality of the number, considering the number of dimensions the figure represents makes interpretation difficult. It is, however, a good starting point in an investigation of soil Orders for both taxonomies.
Table 18. Average nearest neighbour distances between orders of the ASC and ST. This demonstrates the Orders in ST which have the highest level of equivalence to Orders in ASC.

<table>
<thead>
<tr>
<th>Order</th>
<th>Closest</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PODOSOLS</td>
<td>SPODISOLS</td>
<td>9.83</td>
</tr>
<tr>
<td>HYDROSOLS</td>
<td>ENTISOLS</td>
<td>10.92</td>
</tr>
<tr>
<td>KANDOSOLS</td>
<td>INCEPTISOLS</td>
<td>11.31</td>
</tr>
<tr>
<td>FERROSOLS</td>
<td>OXISOLS</td>
<td>11.45</td>
</tr>
<tr>
<td>ANTHROPOSOLS</td>
<td>ENTISOLS</td>
<td>12.33</td>
</tr>
<tr>
<td>KUROSOLS</td>
<td>ULTISOLS</td>
<td>13.43</td>
</tr>
<tr>
<td>DERMOSOLS</td>
<td>ENTISOLS</td>
<td>13.5</td>
</tr>
<tr>
<td>CHROMOSOLS</td>
<td>ENTISOLS</td>
<td>13.67</td>
</tr>
<tr>
<td>VERTOSOLS</td>
<td>VERTISOLS</td>
<td>15.98</td>
</tr>
<tr>
<td>SODOSOLS</td>
<td>ENTISOLS</td>
<td>16.32</td>
</tr>
<tr>
<td>TENOSOLS</td>
<td>SPODISOLS</td>
<td>16.53</td>
</tr>
<tr>
<td>CALCAROSOLS</td>
<td>ENTISOLS</td>
<td>18.78</td>
</tr>
<tr>
<td>RUDOSOLS</td>
<td>INCEPTISOLS</td>
<td>18.89</td>
</tr>
<tr>
<td>ORGANOSOLS</td>
<td>ANDISOLS</td>
<td>34.22</td>
</tr>
</tbody>
</table>

5.4.2.2 Nearest neighbour distances Australian Soil Classification to Soil Taxonomy
The nearest neighbour distances to ASC groups show that Ferrosols and Oxisols have a close affinity, even though there is no overlap. Similarly, Hydrosols do not demonstrate as close an affiliation with Aridisols (Table 20).
Table 19. Average next nearest neighbour distances between Orders of ST and the ASC. This demonstrates the Orders in the ASC that have the highest level of equivalence to Orders in ST

<table>
<thead>
<tr>
<th>Order</th>
<th>Closest</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPODISOLS</td>
<td>PODOSOLS</td>
<td>9.83</td>
</tr>
<tr>
<td>ENTISOLS</td>
<td>HYDROSOLS</td>
<td>10.92</td>
</tr>
<tr>
<td>INCEPTISOLS</td>
<td>KANDOSOLS</td>
<td>11.31</td>
</tr>
<tr>
<td>OXISOLS</td>
<td>FERROSOLS</td>
<td>11.45</td>
</tr>
<tr>
<td>ULTISOLS</td>
<td>KANDOSOLS</td>
<td>12.75</td>
</tr>
<tr>
<td>ALFISOLS</td>
<td>KANDOSOLS</td>
<td>13.96</td>
</tr>
<tr>
<td>MOLLISOLS</td>
<td>DERMOSOLS</td>
<td>15.66</td>
</tr>
<tr>
<td>VERTISOLS</td>
<td>VERTOSOLS</td>
<td>15.98</td>
</tr>
<tr>
<td>ANDISOLS</td>
<td>FERROSOLS</td>
<td>16.89</td>
</tr>
<tr>
<td>ARIDISOLS</td>
<td>SODOSOLS</td>
<td>20</td>
</tr>
<tr>
<td>GELISOLS</td>
<td>HYDROSOLS</td>
<td>31.95</td>
</tr>
<tr>
<td>HISTOSOLS</td>
<td>ORGANOSOLS</td>
<td>48.44</td>
</tr>
</tbody>
</table>

5.4.2.3 Nearest neighbour distances Soil Taxonomy to Australian Soil Classification

From the perspective of ST to ASC, the nearest neighbour distances generally conform to what is expected of each respective Order. Andisols link with Ferrosols, but the distance is large. The closest ASC Order to ST Gelisols is Hydrosols, but the distance is over three times the average, indicating a lack of similarity between Gelisols and anything in ASC. Gelisols are cold climate soils of which there are very few in Australia. Histosols link with Organosols, but demonstrate a huge disparity of distance as well. This can be attributed to the vast differences in organic matter that are considered typical for the two continents. Organosols are almost equally distant from Andisols and Mollisols, yet it is the ASC soil Order that demonstrates the most organic carbon.
5.4.3 Hulls
The areas of convex hulls diminish with each tier of the classification. Nationally, the size difference is considerable between hulls of the two taxonomies, but there is a level of convergence that occurs at the Order and Suborder level. Great Groups were not calculated for ST as there was less data availability. Hull size changes are linear on the log scale (Figure 40).

5.4.3.1 Hull overlay analysis

![Figure 41. Overlap of convex hulls. These two plots demonstrate the overlap of the taxonomically similar orders; Vertisols and Vertosols (left) and Spodosols and Podosols (right).](image)

5.4.3.2 Hull Tables
Total area of hull convergence is represented on the hull overlay table (Table 21). The intersecting hull overlap area has been calculated in two dimensions, based on the list of centroids from TRes. The variability represented by these overlays is small but represent the properties that are most significant in the distances between centroids. Large convergences can be considered to be points of investigation, rather than empirical measures of similarity as the NND tables tend to be.
Table 20. Hull overlap areas between Orders of ST and Orders of ASC. This table demonstrates the proportional areas of occupation in Principal Component space. This is calculated in two dimensions, representing approximately 30% of total variability. “Orders” are the soil Orders, “Olap” is the overlap of the convex hulls. “ST” is the amount of overlap attributed to ST hulls, “ASC” is the amount of overlap attributed to ASC hulls, “%ST” is the same overlap for ST expressed as a percentage, “%ASC” is the same overlap for ASC expressed as a percentage.

<table>
<thead>
<tr>
<th>Orders</th>
<th>Olap</th>
<th>ST</th>
<th>ASC</th>
<th>%ST</th>
<th>%ASC</th>
<th>Orders</th>
<th>Olap</th>
<th>ST</th>
<th>ASC</th>
<th>%ST</th>
<th>%ASC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTS KA</td>
<td>65.1</td>
<td>213.4</td>
<td>177.9</td>
<td>30.5</td>
<td>36.6</td>
<td>ELS HY</td>
<td>10.4</td>
<td>116.4</td>
<td>138.9</td>
<td>9.0</td>
<td>7.5</td>
</tr>
<tr>
<td>NTS KA</td>
<td>50.5</td>
<td>125.6</td>
<td>177.9</td>
<td>40.3</td>
<td>28.4</td>
<td>NTS DE</td>
<td>8.2</td>
<td>125.6</td>
<td>116.1</td>
<td>6.5</td>
<td>7.0</td>
</tr>
<tr>
<td>IDS CA</td>
<td>50.1</td>
<td>130.8</td>
<td>59.0</td>
<td>38.3</td>
<td>84.9</td>
<td>LLS CH</td>
<td>7.5</td>
<td>110.0</td>
<td>185.1</td>
<td>6.8</td>
<td>4.0</td>
</tr>
<tr>
<td>IDS HY</td>
<td>45.8</td>
<td>130.8</td>
<td>138.9</td>
<td>35.0</td>
<td>33.0</td>
<td>LFS DE</td>
<td>6.6</td>
<td>29.6</td>
<td>116.1</td>
<td>22.3</td>
<td>5.7</td>
</tr>
<tr>
<td>PTS VE</td>
<td>45.0</td>
<td>213.4</td>
<td>97.5</td>
<td>21.1</td>
<td>46.1</td>
<td>PTS CH</td>
<td>6.5</td>
<td>213.4</td>
<td>185.1</td>
<td>3.1</td>
<td>3.5</td>
</tr>
<tr>
<td>PTS HY</td>
<td>42.6</td>
<td>213.4</td>
<td>138.9</td>
<td>19.9</td>
<td>30.6</td>
<td>PTS OR</td>
<td>5.8</td>
<td>213.4</td>
<td>21.0</td>
<td>2.7</td>
<td>27.7</td>
</tr>
<tr>
<td>IDS SO</td>
<td>40.2</td>
<td>130.8</td>
<td>125.5</td>
<td>30.7</td>
<td>32.0</td>
<td>ELS KA</td>
<td>5.3</td>
<td>116.4</td>
<td>177.9</td>
<td>4.6</td>
<td>3.0</td>
</tr>
<tr>
<td>LLS VE</td>
<td>39.2</td>
<td>110.0</td>
<td>97.5</td>
<td>35.6</td>
<td>40.2</td>
<td>ELS DE</td>
<td>5.1</td>
<td>116.4</td>
<td>116.1</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>IDS CH</td>
<td>38.6</td>
<td>130.8</td>
<td>185.1</td>
<td>29.5</td>
<td>20.9</td>
<td>LTS DE</td>
<td>4.6</td>
<td>15.8</td>
<td>116.1</td>
<td>29.0</td>
<td>3.9</td>
</tr>
<tr>
<td>RTS VE</td>
<td>36.4</td>
<td>55.4</td>
<td>97.5</td>
<td>65.6</td>
<td>37.3</td>
<td>PTS CA</td>
<td>4.1</td>
<td>213.4</td>
<td>59.0</td>
<td>1.9</td>
<td>6.9</td>
</tr>
<tr>
<td>NTS VE</td>
<td>32.1</td>
<td>125.6</td>
<td>97.5</td>
<td>25.6</td>
<td>33.0</td>
<td>LLS CA</td>
<td>3.3</td>
<td>110.0</td>
<td>59.0</td>
<td>3.0</td>
<td>5.5</td>
</tr>
<tr>
<td>NDS FE</td>
<td>31.0</td>
<td>408.3</td>
<td>31.0</td>
<td>7.6</td>
<td>100.0</td>
<td>NTS PO</td>
<td>2.3</td>
<td>125.6</td>
<td>37.1</td>
<td>1.9</td>
<td>6.3</td>
</tr>
<tr>
<td>PTS FE</td>
<td>30.1</td>
<td>213.4</td>
<td>31.0</td>
<td>14.1</td>
<td>97.0</td>
<td>LTS HY</td>
<td>2.1</td>
<td>15.8</td>
<td>138.9</td>
<td>13.5</td>
<td>1.5</td>
</tr>
<tr>
<td>NDS FE</td>
<td>31.0</td>
<td>408.3</td>
<td>31.0</td>
<td>7.6</td>
<td>100.0</td>
<td>NTS PO</td>
<td>2.3</td>
<td>125.6</td>
<td>37.1</td>
<td>1.9</td>
<td>6.3</td>
</tr>
<tr>
<td>LLS VE</td>
<td>30.1</td>
<td>213.4</td>
<td>31.0</td>
<td>14.1</td>
<td>97.0</td>
<td>LTS HY</td>
<td>2.1</td>
<td>15.8</td>
<td>138.9</td>
<td>13.5</td>
<td>1.5</td>
</tr>
<tr>
<td>NTS VE</td>
<td>32.1</td>
<td>125.6</td>
<td>97.5</td>
<td>25.6</td>
<td>33.0</td>
<td>LLS CA</td>
<td>3.3</td>
<td>110.0</td>
<td>59.0</td>
<td>3.0</td>
<td>5.5</td>
</tr>
<tr>
<td>ODS PO</td>
<td>28.0</td>
<td>83.6</td>
<td>37.1</td>
<td>33.4</td>
<td>75.3</td>
<td>OX HY</td>
<td>0.8</td>
<td>27.1</td>
<td>138.9</td>
<td>3.0</td>
<td>0.6</td>
</tr>
<tr>
<td>PTS DE</td>
<td>27.4</td>
<td>213.4</td>
<td>116.1</td>
<td>12.8</td>
<td>23.6</td>
<td>ELS OR</td>
<td>0.8</td>
<td>116.4</td>
<td>21.0</td>
<td>0.7</td>
<td>3.6</td>
</tr>
<tr>
<td>NDS KA</td>
<td>26.9</td>
<td>27.1</td>
<td>177.9</td>
<td>99.5</td>
<td>15.1</td>
<td>ODS KA</td>
<td>0.5</td>
<td>83.6</td>
<td>177.9</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>LFS KA</td>
<td>23.3</td>
<td>29.6</td>
<td>177.9</td>
<td>78.5</td>
<td>13.1</td>
<td>NTS FE</td>
<td>0.2</td>
<td>125.6</td>
<td>31.0</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>ELS KA</td>
<td>22.8</td>
<td>110.0</td>
<td>116.1</td>
<td>18.8</td>
<td>17.8</td>
<td>LTS FE</td>
<td>0.1</td>
<td>15.8</td>
<td>31.0</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>NDS OR</td>
<td>20.4</td>
<td>408.3</td>
<td>21.0</td>
<td>5.0</td>
<td>97.2</td>
<td>ELS CH</td>
<td>0.1</td>
<td>116.4</td>
<td>185.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>LFS KA</td>
<td>14.8</td>
<td>15.8</td>
<td>177.9</td>
<td>93.8</td>
<td>8.3</td>
<td>NDS DE</td>
<td>0.0</td>
<td>408.3</td>
<td>116.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>NTS HY</td>
<td>12.4</td>
<td>125.6</td>
<td>138.9</td>
<td>9.9</td>
<td>8.9</td>
<td>ODS OR</td>
<td>0.0</td>
<td>83.6</td>
<td>21.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>
This list demonstrates some plausible similarities. Kandosols and Inceptisols, for example, are well overlapped and both soil taxa have a degree of descriptive similarity. Similarly Calcarosols and Aridisols emerge. Being soils that would typically exist in dry locales, these two would have commonality. Orders of extreme similarity, such as Vertisols and Vertosols, Podosols and Spodosols, have been plotted and can be used to show the efficacy of hull analysis. There are some remarkable overlaps as well. In this list is the envelopment of Ferrosols within the hull of Andisols. Aridisols also display a level of association with Hydrosols. These surprising results can be scrutinized with hull plots and nearest neighbour distances.

5.4.3.3 Similar Orders

Soil Orders are the most consistent of groups and the most useful for comparison between taxonomies (Figure 41). Vertisols and Vertosols have similar descriptions, with there being some discrepancy on the particle size, exacerbated by differences in national particle size definitions (Hazleton and Murphey, 2007), and different Suborder and Great Group categories. Spodosols and Podosols are similar as well. This similarity in description is matched by hull overlap. There is a significant overlap between hulls of the two taxonomies using either the Spodosol/Podosol or Vertisol/Vertosol comparisons.

![Figure 42. Overlaid convex hulls. These plots demonstrate the overlap between two Orders that are not typically regarded as similar yet share some characteristics; Ferrosols and Andisols (left) and Aridisols and Hydrosols (right).](image-url)
5.4.3.4 Different Orders
Some Orders that are superficially different but share some characteristics overlap in the first two components (Figure 42), for example Ferrosols and Andisols. These are soils that would not necessarily associate. Aridisols and Hydrosols similarly are an unexpected combination. The overlay plots demonstrate how they interact in a limited variability space.

Figure 43. Overlaid convex hulls. These plots demonstrate the relative positions of taxonomically similar Orders; Oxisols and Ferrosols (left) and Histosols and Organosols (right).

5.4.3.5 Similar but different Orders
There are soils that have limited or no overlap, but are considered similar (Figure 43). Oxisols have low overlap with Ferrosols, and then only if SPDs are plotted in lieu of Great Group centroids. Histosols and Organosols are different when plotted in the same property space relative to each other, but when considered relative to the taxonomy from which they were derived, the hull positions are similar (Figure 44). This suggests that these two groups function the same way with reference to their own specific taxonomic system.
Figure 44. Positions of convex hulls of Orders over their native taxonomic systems in PC space. These plots demonstrate the relative positions of Organosols plotted using only data from the ASC (left) and Histosols using only data from ST (right).

§4.4 Concentricity

The Orders, Suborders and Great Groups display concentric or Matroshka behaviour such as demonstrated in (Figure 45). Much of this is because the names of the taxa themselves constrain them within certain groups. ST Great Groups for example, have names such as Udipsamments, which include the word “Udi”, referring to a udic or humid climate, at the beginning. Climate regimes do not readily translate numerically in this data set. While this is not mathematically optimal, the practical necessity of including features such as climate regime in taxonomic descriptions need to be acknowledged. For many cropping situations, if an SPD has the required agronomic properties but the fact it resides in a xeric regime is not expressed, then that would constitute a crucial omission.
Figure 45. Matroschka plot. This plot demonstrates the organization of convex hulls between and within a taxonomy such as ASC, one hull fitting inside the other as happens with the dolls.

5.5 Discussion

5.5.1 Hulls
Convex hull analysis was able to confirm that taxa that have the same pedological description in both taxonomies occupy a similar space in the PC universe of TReX. This implies that although the properties that define Orders are different and reduced from the full 23 properties that are used for numerical classification, there is still a level of harmony between the two methodologies/philosophies. When using distances alone, the accuracy is high, but overlaps of hulls are able to bring out subtle intermediate associations and characteristics that may have been missed by straight numeric analysis. Five dimension hulls have little or no overlap, as dimensionality increases so does the complexity of the object represented and the likelihood of it emerging on a different plain in reference to another
object. Five components represent about 60% of the total variability. Two dimension hulls have the most overlap, but only represent 30% variability. Hull analysis involves finding the most association between groups with the best amount of variability represented. In finding these associations it is inevitable that some detail will be lost. The first two components, however, represent the most and the most important variability and have been able to highlight associations that require an explanation or possibly an update to taxonomic systems.

5.5.2 Hydrosols and Aridisols
When plotted, the overlap between Hydrosols and Aridisols seems incongruous. This issue with the description of this Australian wet soil class is known however (Fitzpatrick et al., 2003). If the ASC Hydrosol definition is completely understood however, the overlap between this wet Australian Order and what is essentially USA desert soils is understandable. The ASC specifically caters for arid areas in its classification of Hydrosols. According to Isbell (1975) the Order is:

"...designed to accommodate a range of seasonally or permanently wet soils and thus there is some diversity within the order. The key criterion is saturation of the greater part of the profile for prolonged periods (2-3 months) in most years. The soils may or may not experience reducing conditions for all or part of the period of saturation, and thus manifestations of reduction and oxidation such as 'gley' colours and ochrous mottles may or may not be present."

This Order also defers to Organosols, Podosols and Vertosols, which have a higher taxonomic priority, thus making Hydrosols a suborder like Order. Many of the locations plotted out in the official designations have not been tested, being based primarily on elevation data, so it transpires that some Hydrosols exist in some of the most arid regions on Earth, like a desert playa, having not experienced rainfall for a significant period of time. The association with Aridisols is therefore understandable.
5.5.3 Andisols in Australia

The association between Ferrosols and Andisols is based on convex hulls (which account for 30% of total variability), and distances between ST Orders and ASC Orders (which does not include associations with other Orders in ST). In a comparison of ASC Orders to ST orders, the nearest association to Ferrosols are Oxisols. This is empirically a closer match. The overlap between Andisols and Ferrosols in the more diffuse convex hull analysis can be explored however.

Despite the lack of a definitive Andisol class in the ASC, there are documented Andisols in Australia (Lowe and Palmer, 2005). Their relative rarity means that they are not routinely tested for and the ASC does not have much scope for their characteristics. Considering that some data is collected by technicians and students, often for a specific purpose other than taxonomy and the designation of the soil is sometimes an afterthought, it can be inferred that SPDs that would be considered Andisols in any other part of the world may well end up in a variety of Orders in the ASC. Similarities in certain characteristics add to the overlap. Phosphate retention, for example predominates in both Ferrosols and Andisols, albeit in a more extreme incarnation for the latter. A quick check of the literature picks up a cluster of Australian soils, especially in Tasmania, which have low bulk densities (Sparrow et al., 1999; McKenzie et al., 2004), are acidic (Lacey and Wilson, 2001) and have massive phosphate retention (Moody, 1994). A large number of these soils are classified as Ferrosols, or are sometimes classified with the older designation, Kraznozems (Stace et al., 1968). Typically from basalt, ash derived Kraznozems would be a close analogue of Andisols. It is possible, therefore for certain types of Ferrosols to have association with Andisols in data space and this may need to be recognized as some of the management issues are the same. Tests for allophane and other diagnostic features of Andisols are unlikely to have been performed, making a definitive answer on how alike some Ferrosol classes are to Andisols difficult. Fortunately, the rarity of this soil Order and dissonant descriptions in Australia makes this a relatively insignificant problem. The TERN data set has a reasonable number of Ferrosols for which the bulk densities are extremely low. TERN was mined for 23 specific properties that match with the USDA centroid data set. There are many other properties recorded in the TERN dataset and further revelations and associations could be determined with a more targeted
search. Appending ASC soil descriptions to include horizon materials that are spongy and sparkly may be a simple work around. Spongy refers to the capacity to retain moisture, Ferrosols would have a tendency to be friable, while Andisols would be squishy. Similarly, volcanic glasses tend to shine, so presence of sparkles, and the way the soil compacts in the hand could be used as a rapid low cost diagnostic feature which is in step with the ASC concept of soil descriptions with a minimum of laboratory data. Alternately, simply borrowing Andisols from ST or combining the two systems would be an acceptable solution.

5.5.4 Distances
Nearest neighbour distances are empirical and can also provide details on level of similarity between Orders. Distances are arranged in two groups; the distances of ASC Orders to ST Orders and the distance of ST orders to ASC orders. When calculating these distances, the taxonomy of origin is excluded, i.e. when determining the distance ASC Ferrosols to ST Orders, all other ASC Orders are excluded from the calculation. Likewise from ST to ASC. The nearest ASC Order to a given ST Order is not necessarily the same in the reverse direction. This is why Ferrosols link simultaneously to Andisols and Oxisols in the ANND tables. In this case, the lowest ANND is the most important. In this instance, Oxisols and Ferrosols have the best association. The link to Andisols has been explored, but is not the true match. It can also be seen that Kandosols are close to a variety of ST Orders. The reason is because the defining criteria for the Kandosol Order in ASC would be considered criteria for a sub Order in ST. This has the effect of spreading Kandosols more or less evenly over several ST Orders. The Inceptisol-Kandosol connection has the shortest distance (11.31), therefore, this is the best match.

5.5.5 Order priority
Soil Orders in ST operate with different priorities to the ASC. This is a result of differing climatic and mineralogical circumstances requiring different approaches. A table of differing priorities to taxonomic systems has been created (Table 22). This table is only a general concept, as there is some disconformity in definitions of properties and diagnostic criteria. Note that there is a specific Order in
the ASC related to the presence of sodium, the Sodosols, although there are references to sodium in some of the ST Order criteria.

Table 21. Priorities of properties in reference to tier level of ST and ASC. This is a rough guide on how various properties guide the allocation of Orders, Suborders and Great Groups in both the ASC and ST.

<table>
<thead>
<tr>
<th>Property</th>
<th>ASC level</th>
<th>ST level</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaCO₃</td>
<td>Order</td>
<td>Suborder</td>
</tr>
<tr>
<td>H₂O</td>
<td>Order</td>
<td>Suborder</td>
</tr>
<tr>
<td>Na</td>
<td>Order</td>
<td>Suborder</td>
</tr>
<tr>
<td>Human activity</td>
<td>Order</td>
<td>Subgroup</td>
</tr>
<tr>
<td>Fe</td>
<td>Order</td>
<td>Order</td>
</tr>
<tr>
<td>Texture</td>
<td>Order</td>
<td>Order</td>
</tr>
<tr>
<td>pH</td>
<td>Order</td>
<td>Order</td>
</tr>
<tr>
<td>Structure/ped organization</td>
<td>Order</td>
<td>Order</td>
</tr>
<tr>
<td>Permafrost</td>
<td>NA</td>
<td>Order</td>
</tr>
<tr>
<td>Organic Matter</td>
<td>Order</td>
<td>Order</td>
</tr>
<tr>
<td>Base saturation</td>
<td>Suborder</td>
<td>Order</td>
</tr>
<tr>
<td>Kandic properties</td>
<td>Order</td>
<td>Suborder</td>
</tr>
<tr>
<td>Aridic temperature regime</td>
<td>NA</td>
<td>Order</td>
</tr>
</tbody>
</table>

The differences in priority of various properties should manifest itself as convex hulls of one ASC Order overlapping convex hulls of several ST Orders or vice versa. This can be seen in the hull overlap and NNDs of Kandosols. The level of cohesion is associated with data availability, but the ASC tends to have classifications to the Suborder and Great Group level less frequently than in ST, in effect making ASC Orders appear in data space with similar characteristics to suborders in ST. This level of resolution that the Orders in ASC and the suborders in ST seem to share, combined with level of priority means that investigations into a combined taxonomy should involve adding ASC suborders to ST at the Great Group level.
5.6 Conclusions

TRex was successfully combined with TERN SPDs. Missing data was appended from the centroid list in TRex. The resulting data set will provide the basis for any optimization of the ASC section of TRex. Adding to the database would necessitate adding SPD’s in the correct format. Existing algorithms could then create centroids.

The data was then used to create a distance matrix. From this, a series of mean nearest neighbour distances in and between taxa from the ASC and ST was calculated. Average NNDs for each Order, suborder and Great Group were created or estimated for both ST and ASC. These estimates demonstrate the coherence of both taxa and point to the operational levels at which both function. It therefore can be seen that the ST suborders and ASC Great Groups are on a similar taxonomic level.

Convex hulls were created from the PCA of the combined dataset. Overlap areas were calculated and these diffuse overlaps provided areas of investigation. The hull analysis and the NNDs of Orders highlights issues in classifications which push Ferrosols close to Andisols, and the bimodal nature of ASC hydrosols was demonstrated. Organic soils from both taxa had hulls plotted and overlaid, their structure in PCA space was shown to be similar when plotted in reference to the taxonomy of origin. Vertosols and vertisols were overlaid as were Spodosols and Podosols. These hull overlays demonstrate the coherence of similarly classified taxa.

5.6.7 Future work

The data suggests that as a general concept, ASC Suborders have a level of affinity with ST Great Groups. Trials, therefore, thinning ST to the suborder level, and adding ASC Great Groups may assist in the optimization of a combined taxonomy. The disparity between Order priorities will need to be negotiated, and a threshold of similarity at which a given taxa would be considered a duplicate of another, needs to be set. A procedure therefore needs to be set to thin the ranks of ST, priming the data to be a consistent, international set of reference taxa.
5.7 References


Chapter 6

Sequential creation of a generic soil classification system using taxa from existing systems

Summary
A classification or taxonomic system based on a common set of principles can be created. In order achieve this a universal database is required. This database needs to be comprehensive, non-duplicated, have the capacity for other suitable taxon to be added to it, and at the same time to be small and simple. As an initial step, soil profile description data and taxa from two differing taxonomies have been successfully combined into a single database, referred to as the Total Reference taxonomic database, or TRex. This has the data range to suit most soil properties in both Australia and the United States, but there is extensive duplication and issues of equivalence, whether a taxonomic tier such as Order, Suborder or Great Group of one taxonomic system is equivalent to the same tier of another taxonomic system. An additive model by which the impediments of redundancy and equivalence can be countered is proposed and the creation of a combined data set that is just as effective for the purpose of creation of a world taxonomy, yet has less unnecessary overlap, is demonstrated. These ideas are incorporated, creating a hybridized data set that covers that same range as TRex in principal component space, yet is significantly smaller, making distance calculations and optimizations of various kinds much faster. The possibility merging and simplifying of such different systems demonstrates the possibility of similar mergers with taxonomies of other nations, creating a global data pool for a world taxonomic system.

6.1 Introduction
There are many differing soil taxonomic systems around the world. Soil profile descriptions (SPDs) have always been a difficult to successfully translate between taxonomic systems (De La Rosa et al., 2002; Shi et al., 2006). This is concurrent with the real time modern challenges of agricultural production for food, fuel and fibre which is becoming an increasingly global exercise, yet relies on
taxonomic methods that are sometimes siloed (Shi et al., 2006), sometimes anachronistic and
sometimes based on traditional knowledge (Barrera-Bassols, 2003). There has been much work in the
field of data generation for a host of nations, and the idea of numeric analysis for systematic soil
descriptions, proposed in the 1950’s (Hole and Hironaka, 1960) has begun to gain momentum. It has
always been desirable to create a universal classification that can somehow transcend the gaps
between taxonomies and systems. The creation of a global numeric taxonomic system is confounded
by the same distance, anachrony and siloing that has hampered translation. These difficulties, using
modern methods can be overcome. It is possible to create a classification system based on common
world soil principles, identifying a set of common properties and a common way of dealing with them
via depth functions. Additionally, a series of classes based on Soil Profile Descriptions (SPDs) can be
created. For each class we can create a centroid for each taxon. While the identification of properties
and synchronization can be achieved, a problem exists in the fact limited sets of profile descriptions
are available and these are typically bound to a specific taxonomic system and confined to the country
from which the taxonomic system originated. A few systems can be drawn from which are purported
to be global. Soil Taxonomy (ST) and World Reference Base (WRB) are typical example of these
“global” systems. It is proposed that it is possible to use global and regional systems to assess the
feasibility of creating a world taxonomic system. There is not enough data available for WRB, but the
USDA has recorded enough soil profile and soil property information for a large scale analysis of ST.
The USDA data can be used to create centroids for ST, which represents the mean values of the
common properties, previously alluded to. The previous chapters recognize the Great Group is a
useful level with which to work. There enough data available to calculate centroids for the majority of
ST Great Groups (Lang et al., 2013). When this is done and the resulting properties projected into
PCA space it is possible, using Euclidean distances, to recognize that some taxa are close and some
are far apart. Therefore it can be said that within this data is a level of redundancy, but it is also
possible to have missing taxa represented by large gaps between taxa in soil property space. ST can be
compared with another system at the same level of generalization (which can be determined by
average taxonomic distance). Doing this reveals that ST Great Groups are at the same level of
generalization as ASC suborders, these tiers of the taxonomies can be successfully compared. Using
these new comparisons, it can be seen that the ASC has taxa that fill in some of the gaps. Not all overlap, new taxa from the ASC can be brought into this increasingly more general combined classification. For this purpose, a method of selecting the correct taxa for insertion needs to be created. This in turn, leads to the creation of a global taxonomy, as the methods used to harmonize and combine Australian STAs can easily be applied to data from other classification systems.

The overall methodology is then simple; a global/universal taxonomy is created sequentially, starting with ST Great Groups and systematically adding groups from other systems which are of the same level in a given hierarchy. This chapter begins with Soil Taxonomy, which is a well-regarded taxonomy. Distances between its individual taxonomic units at the Great Group level are standardized and augmented by adding taxa from the ASC, in preparation for calculations to create a global taxonomy.

6.2 Aims

The aims of this study are to:

- Create criteria by which an unnecessarily large database can be culled.
- Organize data from ASC so it is on the same level of generalization as ST.
- Add ASC taxa to ST centroids sequentially.
- Reduce combined data set to a simpler state without losing information or regional diversity.

6.3 Method

Methodologically, ST should be the starting point thus removing redundancy by devising a thinning algorithm to eliminate taxa too close together in taxonomic space. Once that is achieved, taxa from another system, which are at the same level of generalization provided they are not within a certain distance from classes in the system, can be added. This chapter demonstrates how this can be accomplished. Not surprisingly, taxa from other world systems (Brazil, Canada, China, New Zealand,
Russia, WRB, etc.) could be added sequentially the same way to build up a fully populated taxonomic space. This is another stepping stone in the creation of a global taxonomy.

6.3.1 Removing redundancy
Taxa from soil classification systems, especially from systems such as ST can be classified to an extreme level of detail. In creating a more general classification for the world, the level of detail required needs to first be determined. Several hundred mollic soil taxonomic allocations (STAs) would be of limited value for reference in largely arid countries such as Australia. Similarly multiple ASC classifications that describe fragile dispersing natric soils (Sodosols in ASC) would be of less use in the United States, but having a small number of examples of each of these soil types would be of value as a point of reference when confronted with taxa from an unknown classification, so long as the range of soils these taxa represent is adequately covered. Redundant data from within a taxonomic system must be removed and good reference points need to be retained. On the broader scale, each world classification system has been designed to address issues endemic to the region from which the taxonomy was derived. It is possible therefore, for the same problems to occur the world over and sometimes similar environmental conditions. It then stands to reason that there may be taxa within each classification system, which are similar to taxa from other classification systems. By way of example, the ASC and ST are known to have two nearly identical Orders, and at least two further Orders that have many similar features. These groups are differentiated with different, diverse Suborder and Great Group categories, but there may be taxa at these reduced levels that are the same. It is necessary to find within a combined or merged parent taxonomy, the groups within each daughter taxonomy that are so numerically similar there is no purpose in giving them a separate designation. In the case of ST and ASC, the Orders Vertosols and Vertisols are but for a few minor details identical, and the same can be said for Spodosols and Podzols. For the similar soils that are not necessarily exactly the same e.g. Entisols and Rudosols, Oxisols and Ferrosols, the question of whether these taxa are different enough to warrant exclusion needs to be addressed. In answering this, it needs to be determined how close taxa should be to another before they are considered duplicates. This answer may not necessarily be consistent over the entire database as for example in the case of Histosols, the data ranges over a wide area as opposed to Oxisols, which occupy a much smaller area (Figure 46). It
cannot be assumed, therefore, that a distance between taxa within the Histosol Order is comparable to distances between taxa in the Oxisol Order. A method is required that is able to distinguish similarity without extinguishing small groups or tiers. As each Order is defined by Suborder, Great Groups and SPDs, which have greatly divergent definitions between ST and ASC, although the Order is identical, the individual taxa at the Great Group level may have no equivalent in the other taxonomy.

Figure 46. Convex hull plots of two Orders overlaid. This plot compares the magnitude of data dispersion of Histosols and Oxisols in ST over 2 principal components.

Figure 47. Mean nearest neighbor distances according to classification system. Plot demonstrating mean nearest neighbour distances in principal component space between individual taxa and SPDs in ST and ASC. This can be used to determine the level to which each tier of each classification is equivalent.
6.3.2 Combining at the correct level of generalization

It was determined in previous research that the ST Suborder level is the functional equivalent in ASC of Great Group. ASC Great Groups were therefore added as Suborders of ST. Designations and nomenclature can be determined using a variety of methods, the major criteria being functionality, logicality and simplicity. The ASC Great Group designations have been retained; these designations may need some revision in order to seamlessly merge with the ST names, but their nomenclature is such that they are easily identified within the combined data which is advantageous for the time being.

6.3.3 Adding sequentially

The majority of the SPDs that exist in ST and the majority of SPDs that exist in the ASC have been splined to a standard depth and the average of properties at these standard depth intervals calculated. Mean properties that are important numerically and agronomically have been extracted and have been incorporated into a large database of centroids for each taxa (Soil Taxonomic Allocations or STAs). ASC data, owing to its level of generalization, has had all of the Great Groups merged into their parent suborders. These are the groups to combine with the ST taxa. Once the STAs from ST are thinned, gaps in soil property space can be added to by the introduction of ASC taxa. These should be added in a manner similar to the thinning procedure, as the ultimate goal is to have a combined set of soil taxa that cover the maximum range of soil property space while still being simple. The STAs from both taxa can be combined and the same thinning procedure performed.

6.3.4 Thinning algorithm

In order to minimize duplication, the classes that are taxonomically close to each other can be removed. It is inevitable; however that there will always be a pair of points with the same minimum distance, the distance to each other. A decision therefore, needs to be made on which class within the pair should be retained in the data set and which class needs to be eliminated. In order to do this it needs to be determined how remote each of the classes in the pair is, and this is done by finding the taxonomic distance to the next closest class. If the next closest class is far away, it is a taxonomically
remote class and should stay. If the next closest point is close, then it is not taxonomically remote and can therefore be removed. The algorithm can be described as the following:

- Start with classification system (e.g. ST centroid set).
- Decompose classification system via principal component analysis.
- Calculate distance matrix of components \((D)\).
- Determine an acceptable minimum cut-off distance \((cd)\).
- For each row \((d)\) in \(D\), determine minimum distance \((dp1)\):

\[
(1) \quad dp1 = \min_{i=1,...,m} d_{i,m}
\]

Where:

\(i=1,...,m\) denotes the indexes of the rows of vector \(d\)

- if \(dp1 > cd\), no action required for this class, move to next class.

If not:

- Determine the closest class to \(d\) \((P1)\)
- Determine minimum distance for class \(P1\)
  - If closest class to \(P1\) is not \(d\), eliminate \(P1\)
  - If closest class to \(P1\) is \(d\):
- Determine the second lowest distance in class \(d\) \((c2)\)
- Determine the second lowest distance in class \(P1\) of \(D\) \((dp2)\)
  - If \(c2 > dp2\), eliminate \(P1\)
  - If \(c2 < dp2\), eliminate \(d\)

When applied to centroid data, this method removes close, non-remote taxa. In this particular instance, a thinning algorithm was applied first to the ST centroid data set, then the thinned ST was combined with the ASC and the thinning algorithm run again. This avoids the superfluous calculations and decision making on taxa that are unnecessarily close within ST.

**6.4.5 Determining level of equivalence**

Although SPDs and even taxa have been compared numerically in the past (Arkley, 1976), and the idea of building a taxonomy based on this has been posited before (Webster, 1972), the steps involved
are not well worn. There is no rule of thumb for a suitable taxonomic equivalence. Unpublished research from our department, has been posited that 3% numeric similarity is the taxonomic equivalent of ST Subgroup. This is below the Great Group level in ST. Any measure used should be above this .03*distance units equivalence threshold. In this particular setting, the performance of this number is unknown. The measure of equivalence is the average minimum distance of the Order being appraised/added to. This number should be above the 3% suggested, and considering that the largest minimum distance between taxa is on or around 100 units, the threshold could be calculated as distance between pairs \((pd)\) compared to the maximum distance \((md)\) in a given data set:

\[
\frac{pd}{md} > 0.03
\]

This method will be evaluated further in future studies. As the current size of the soil property universe almost exactly matches a distance of 100 units between the furthest two points, 0.03 is a suitable point of reference. If the soil property universe size is altered by the addition of new data or systems, the figure of 0.03 may need to be supplemented with \(0.03*\text{max distance}/100\).

6.4 Results

6.4.1 Thinning Soil Taxonomy and Australian Soil Classification

The mean minimum distance between taxa in ST is around 15 standard taxonomic distance units (STDU). Considering the maximum distance between the combined taxa of ST and ASC is little over 100 STDU, each STDU can be considered to be at or around 1% of maximum variability. Histosols have large inter taxa distances and exert more influence on the data than other groups. A cut-off distance of five STDU was trialed, roughly equating to 5% variability, which negates the influence of Histosols but is close to the estimated ST family average differentiation. With this cut-off distance in place, the US data excluded eight Great Groups from the combined data set: Argiustolls, Haplustalfs, Hapludults, Haplohumults, Paleustalfs, Paleudults, Palehumults, and Paleustolls. All these are derivatives of Alfisols, Ultisols and Mollisols. The majority are either pale or haplic. This implies that
these qualifiers are taxonomically similar to each other in property space. When performing the same calculation, 31 ASC Great Groups are eliminated. This larger number of exclusions implies that ASC Great Groups are more tightly bound in data space. The majority of eliminated groups are Tenosols, with a smaller number of Rudosols being excluded. A small number of Sodosols are cut, a still smaller number of Vertosols are removed with an equally small number of Hydrosols. Half of the excluded Hydrosols have a “rudic” qualifier. The 5% cut-off excludes more ASC than ST at the Great Group level which, is expected considering it was already established that the ASC Great Group minimum distances were too small. The cut-off figure although arbitrary, seems to be a good estimate of the calculated taxonomic equivalence, placing these two groups at or around 5% variation which is above the presumed species limit of 3%.

6.4.2 Australian Soil Classification to Suborder
As the ASC Suborders are seen to be the equivalent of ST Great Groups by distance calculation, adding thinned ASC Great Groups would possibly be a retrograde step. The ASC Great Group centroids were therefore supplanted with Suborder means, which were in turn calculated from the original ASC Great Groups. This provides 81 rows in line with the number of Great Groups that have sufficient data available. The thinning algorithm demonstrates that at the suborder level, the ASC groups are sufficiently isolated.

6.4.3 Combining the two sets
TRex is a combined database of ST and ASC that has 772 taxa rows, taking into account Orders and Suborders. Culling TRex and removing the ST Order means and ST Suborder means, then supplanting the ASC suborder means as a substitute for Australian Great Groups, then thinning with a limiting criterion of five provides a database of 291 rows (Table 23). Once combined with Suborder means, no additional centroids/taxa are eliminated. This new set of taxa theoretically encapsulates the majority of property ranges extant in ASC and ST at roughly the level of detail expected of ST Great Groups (Figure 48). This lends credence to the idea that ASC suborders are at or around the same level as ST Great Groups. The eight ST Great Groups initially eliminated shows that the levels at
which the two taxonomies merge are not precise, merged soil groupings behaving more like a continuum than a set of fixed categorical tiers. Presumably all taxonomic systems will operate on variable levels of which there will only be an approximate alignment with others.

![ASC SubOrders combined with thinned ST Great Groups](image)

**Figure 48.** Comparison of thinned data from ST and ASC with un-thinned data from both taxonomic systems. This demonstrates the ability of the reduced data set to capture most of the range of soil taxonomic space.
Table 22. List of soil taxa from both ASC and ST which are a minimum of 5 STDU apart. ASC taxa are coded but ST has full names provided. This table is the full list of data that can be used in the creation of a global taxonomy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AUS</td>
<td>AN HY</td>
<td>USA</td>
<td>PALUANGS</td>
<td>USA</td>
<td>HAPLOXERUltS</td>
<td>USA</td>
<td>ENODIAquiLFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>AN HW</td>
<td>USA</td>
<td>PALACQUOdS</td>
<td>USA</td>
<td>HAPLOXERUltS</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA BD</td>
<td>USA</td>
<td>PETROGysiODS</td>
<td>USA</td>
<td>HAPLOXERUltS</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA CQ</td>
<td>USA</td>
<td>PETROCALcODS</td>
<td>USA</td>
<td>HAPLOXERUltS</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA CV</td>
<td>USA</td>
<td>PETROARiODS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA CY</td>
<td>USA</td>
<td>PETRAQUiEpS</td>
<td>USA</td>
<td>HAPLOXERALFS</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA DA</td>
<td>USA</td>
<td>PALEXERUltS</td>
<td>USA</td>
<td>HAPLOTurBELLs</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA EL</td>
<td>USA</td>
<td>PALEXEROLLs</td>
<td>USA</td>
<td>HAPLOTorBELLts</td>
<td>USA</td>
<td>DYSTRIQUEpS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA FB</td>
<td>USA</td>
<td>PALEXERAiFS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA GS</td>
<td>USA</td>
<td>PALEUSTULtS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CA YY</td>
<td>USA</td>
<td>PALEUDDLls</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CH AA</td>
<td>USA</td>
<td>PALEUDALFS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CH AB</td>
<td>USA</td>
<td>PALEARiODS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CH AC</td>
<td>USA</td>
<td>PALEAUDDLls</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CH AO</td>
<td>USA</td>
<td>NATRUStOLLs</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>CH AE</td>
<td>USA</td>
<td>NATRUStALFS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>DE AA</td>
<td>USA</td>
<td>NATRUdULLS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>DURiSTOLLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>DE AB</td>
<td>USA</td>
<td>VE AA</td>
<td>USA</td>
<td>NARiTRULdiOdS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>DE AC</td>
<td>USA</td>
<td>VE AB</td>
<td>USA</td>
<td>NARiTRULdiOdS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>DE AD</td>
<td>USA</td>
<td>VE AO</td>
<td>USA</td>
<td>NARiTRULdiOdS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>DE AE</td>
<td>USA</td>
<td>VE AE</td>
<td>USA</td>
<td>NARiTRULdiOdS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>FE AA</td>
<td>USA</td>
<td>VE AM</td>
<td>USA</td>
<td>NARiTRAGiDS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>FE AB</td>
<td>USA</td>
<td>XERORThENTS</td>
<td>USA</td>
<td>NARiTRAGiDS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>FE AC</td>
<td>USA</td>
<td>XEROPIsSAMMENTS</td>
<td>USA</td>
<td>NARiTRAGiDS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>FE AE</td>
<td>USA</td>
<td>XEROFLUENTs</td>
<td>USA</td>
<td>NARiTRAGiDS</td>
<td>USA</td>
<td>HAPLOXERANOdS</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>HY AB</td>
<td>USA</td>
<td>VITRIKEXANODS</td>
<td>USA</td>
<td>NARiTRULdiOdS</td>
<td>USA</td>
<td>DYSPIUSiERTTs</td>
<td>USA</td>
<td>CRYOFLUENTs</td>
</tr>
<tr>
<td>AUS</td>
<td>HY BT</td>
<td>USA</td>
<td>VITRIKORDRANODS</td>
<td>USA</td>
<td>MELANiURiiODS</td>
<td>USA</td>
<td>GLOSSUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY CS</td>
<td>USA</td>
<td>VITRIKEXANODS</td>
<td>USA</td>
<td>MELANiURiiODS</td>
<td>USA</td>
<td>GLOSSUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY CW</td>
<td>USA</td>
<td>VITRIQUAiNdS</td>
<td>USA</td>
<td>KANHAPiUStULLtS</td>
<td>USA</td>
<td>GLOSSUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY DT</td>
<td>USA</td>
<td>VERMUDOLLs</td>
<td>USA</td>
<td>KANHAPiUStULLtS</td>
<td>USA</td>
<td>FULViUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY ED</td>
<td>USA</td>
<td>VERMIQUAliFS</td>
<td>USA</td>
<td>KANHAPiUStULLtS</td>
<td>USA</td>
<td>FULViUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY EG</td>
<td>USA</td>
<td>USTORiThENTS</td>
<td>USA</td>
<td>KANHAPLOHUMULtS</td>
<td>USA</td>
<td>FRAGLOSSUDALFs</td>
<td>USA</td>
<td>CACOLiXERODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY EW</td>
<td>USA</td>
<td>USTiViRANODS</td>
<td>USA</td>
<td>KANHiUSTULLtS</td>
<td>USA</td>
<td>FRAGiKEXEPtS</td>
<td>USA</td>
<td>CACoGISODS</td>
</tr>
<tr>
<td>AUS</td>
<td>HY YY</td>
<td>USA</td>
<td>USTiPSiSAMMENTS</td>
<td>USA</td>
<td>KANHiUSTULLtS</td>
<td>USA</td>
<td>FRAGiKEXEPtS</td>
<td>USA</td>
<td>CACoGISODS</td>
</tr>
<tr>
<td>AUS</td>
<td>KA AA</td>
<td>USA</td>
<td>USTiFLUENTs</td>
<td>USA</td>
<td>KANHiUSiLTs</td>
<td>USA</td>
<td>FRAGiKEXEPtS</td>
<td>USA</td>
<td>CACoGISODS</td>
</tr>
<tr>
<td>AUS</td>
<td>KA AB</td>
<td>USA</td>
<td>UMBRAQUlALFS</td>
<td>USA</td>
<td>KANHiUSiLTs</td>
<td>USA</td>
<td>FRAGiKEXEPtS</td>
<td>USA</td>
<td>CACoGISODS</td>
</tr>
<tr>
<td>AUS</td>
<td>KA AC</td>
<td>USA</td>
<td>UDOiRiThENTS</td>
<td>USA</td>
<td>KANHUAUfiLs</td>
<td>USA</td>
<td>FRAGiKEXEPtS</td>
<td>USA</td>
<td>CACOQuiLFS</td>
</tr>
<tr>
<td>AUS</td>
<td>KA AD</td>
<td>USA</td>
<td>UDIViRANODS</td>
<td>USA</td>
<td>KANPiPEROx</td>
<td>USA</td>
<td>FRAGBOrTHODs</td>
<td>USA</td>
<td>ARGiKEXEPtS</td>
</tr>
<tr>
<td>AUS</td>
<td>KA AE</td>
<td>USA</td>
<td>UDIpsiSAMMENTS</td>
<td>USA</td>
<td>KANHiUMULtS</td>
<td>USA</td>
<td>FRAGiMAQUiLTs</td>
<td>USA</td>
<td>ARGiKEXEPtS</td>
</tr>
<tr>
<td>AUS</td>
<td>KU AA</td>
<td>USA</td>
<td>UDIiOLiFS</td>
<td>USA</td>
<td>HYRXUAdALFs</td>
<td>USA</td>
<td>FRAGiMAQUiLTs</td>
<td>USA</td>
<td>ARGiKEXEPtS</td>
</tr>
<tr>
<td>AUS</td>
<td>KU AB</td>
<td>USA</td>
<td>UDIiFLUENTs</td>
<td>USA</td>
<td>HYRXUAdALFs</td>
<td>USA</td>
<td>FRAGiMAQUiLTs</td>
<td>USA</td>
<td>ARGiKEXEPtS</td>
</tr>
</tbody>
</table>

179
<table>
<thead>
<tr>
<th>AUS</th>
<th>KU AC</th>
<th>USA</th>
<th>TORRISAMMENTS</th>
<th>USA</th>
<th>HUMISAMMENTS</th>
<th>USA</th>
<th>HUMIDETTS</th>
<th>USA</th>
<th>FLUVIAQUENTS</th>
<th>USA</th>
<th>ARGIGRIVOLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUS</td>
<td>KU AO</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>KU AC</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>OR BW</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>OR CE</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>OR GH</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>PO AL</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
<tr>
<td>AUS</td>
<td>PO AM</td>
<td>USA</td>
<td>TORRISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>HUMISAMMENTS</td>
<td>USA</td>
<td>FLUVIAQUENTS</td>
<td>USA</td>
<td>ARGIGRIVOLS</td>
</tr>
</tbody>
</table>

### 6.5 Discussion

#### 6.5.1 Levels of Australian Soil Classification and Soil Taxonomy

There are a large number of taxa that can be drawn from to create a numerical taxonomic system. However, simply having the data does not mean that it should be used. Using distance calculations to determine the level at which ASC and ST reduces complexity and provides a picture of how one taxonomic system operates with reference to another. In the case of ST and the ASC, there is increasing evidence that the ASC suborder level has the same function as the ST Great Group level. Previously it was discovered that ASC Great Groups coalesce with other ASC Great Groups of the same Order more often than in ST. This greater coherence is probably associated with the inherent simplicity in ASC descriptions. The idea that ASC suborders have the same level of generalization as ST Great Groups is given some credence by the relatively low amount of ST taxa removed at the Great Group level, compared to the larger number of ASC taxa removed at that level- contrasted by no eliminations of ASC suborders.
6.5.2 Soil Taxonomy Orders and sub categories
Of the eight ST Great Groups eliminated, three were haplic. This constitutes a large proportion of the eliminated rows. Users of ST think of Hapl from Greek haplous, simple. They make up the central concept of the Suborder, left because all other Great Groups in the Suborder (those with some morphology or character that is not "simple") have been allocated previously and eliminated from consideration. Haplic qualifiers are often describing taxa that are difficult to define or have a single errant property, rather than a series of properties or descriptors that fulfil a specific purpose (Nachtergaele et al., 2001). It is interesting, therefore, that these qualifiers prove themselves to be almost indistinguishably similar to other taxa in property space. This suggests that the priority of haplic qualifiers should be changed. This is an issue for pedologists well versed in ST. The other exclusions happen at the Order level for ST but the majority of the sub qualifiers are based on climate regime. It is possible that climate regime is not a good differentiating factor in numerical analysis. Accounting for this is difficult as the purpose of climatic descriptors is not necessarily as a property per se but a method of identifying areas with the required rainfall, temperature etc. for agronomic or other human purposes. Removing climate information would reduce the effectiveness of ST.

6.5.3 Australian Soil Classification Orders and sub categories
In ASC, the majority of the taxonomically uniform Great Groups (which were ultimately not included in the final product) are Tenosols and Rudosols. Distinguishing between these two Orders is a challenge mathematically. ASC is an excellent system, but there are those in states dominated by sandy soils that are progressing an argument to have an order that caters more specifically for sand. Perhaps there is an opportunity here to augment or requisition the Tenosol class for this purpose. Adding descriptions that encompass different sandy soils could result in more measurable differentiation. This is advantageous for both ASC and ST. Some Hydrosols were also removed from the Great Group set. Commensurate with close Hydrosols are taxa close to Spodosols (ST Podosols), a taxon that occurs on the coast and also in the deep, arid interior of Australia. This reflects the dual nature of ASC Hydrosols that sometimes occupy a niche that in ST would be taken by Aridisols.
6.5.4 Turning data into a system

This work has been able to highlight areas of improvement for both ST and the ASC. It is obvious that combining and comparing these two systems this way will assist in the creation of a classification based on common principles. The taxa from ST and ASC have been combined and thinned, the techniques used to establish this combination imply there is a relationship between all taxa in this data set but how each ASC Suborder relates to the ST Great Groups in a way that can be used in a taxonomy needs to be established. With the calculation of Euclidean distances on principal components, dendrograms are the most probable solution (Moore et al., 1972). A dendrogram is a visual method of demonstrating similarities between groups and is almost ubiquitous when considering biological taxonomies. Using such a device, a user can identify relationships between taxa within a system they are familiar to and a system that is unknown. Nomenclature can be based on the results of dendrogram analysis, possibly using a sequential system not unlike the IUPAC system found in chemistry (Muller, 1994).

6.6 Conclusions

In this chapter it has been shown that a 'universal' comprehensive system of soil classes can be built up sequentially by taking the centroids for Great Groups from Soil Taxonomy, and applying a thinning algorithm so that no centroid was closer than five STDU from any other. This results in a data set of 210 ST taxa. It was decided that as the taxonomic level of ASC Suborders align mathematically to Great Groups in ST, any combined taxonomy should be based around ST Great Groups and ASC suborders. This idea was strengthened when no ASC Suborders were culled from the initial list prior to augmentation with ST. Once combined, the conjoined data was thinned again, no further eliminations were required indicating a relatively seamless merger. The addition of the 81 ASC Suborders resulted in a data set containing 291 individual taxa from both ST and the ASC. This can be thought of as a cloud of 291 reference points in a 414 dimensional space. The combined set occupies a similar range to the original 8000 strong data set (including individual SPDs), giving a more or less uniform coverage over the entire data space. This coverage can be made more complete with the addition of extra data (e.g. New Zealand Landcare Centroids). Sequential centroids
representing taxa at Great Group level from other significant and national taxonomies (e.g., WRB, Brazil, China etc.) can be thinned, tested and incorporated in the same way. They can supplant the data, provided they are more than five STDU from existing taxa in the system. In this way new kinds of soil profile can be identified, which have always been known to a particular taxonomic system, and their place within taxonomic space can be identified, through a cloud of reference points, that is, the centroids representing the quantitative centres of the SMT (the soil mesotaxa), the operational taxa of the system.
6.7 References


Chapter 7

Creation and Nomenclature for a New Classification System

Summary

Much work has been put into the creation of a simple combined taxonomic database consisting of selected Soil Taxa from ST and the ASC. These chosen taxa have been referred to as Taxonomic allocations. This calculation and allocation of these two soil classifications, can be capitalized on, creating a distance based taxonomic system and nomenclature, which can be more clearly understood by an operator familiar to either ST or ASC. The relationships between soil taxa from two systems can be explored, but finding logical methods of identifying and characterizing common features and differences between the taxa of differing systems is difficult. In this thesis a system of classification and nomenclature based on dendrogram analysis was presented. A Ward dendrogram is calculated from a distance matrix determined from principal components of soil data that is derived from a culled variant of the TReX data set, which consists of taxonomic descriptions of Great Groups for both ST and ASC, splined into 18 standardized increments to 150cm depth, consisting of 23 independent properties. This dendrogram is then cut at various tiers, creating logical end groups, which contain taxa from both ST and ASC. The properties that differentiate groups at each node of the dendrogram are identified. It has emerged that the allocated taxa from, Soil Taxonomy and the Australian Soil Classification system coalesce within these groups according to common properties or pedological processes. The nodes of the dendrogram that created these groups can be linked back to specific properties and they can be used to create a taxonomic key. This property association could form the basis of a property based description scheme and taxonomic key for this combined group of soil taxa. Gypsic soils within extreme conditions tend to occur in the most remote taxonomic group, and some gypsic taxa have been relegated to an extragrade cluster. Other features of Soil Taxonomic allocations that can be defined in order of their influence in a dendrogramatic key (bearing in mind that the same factor can be used several times going down the nodes of a dendrogram) are bulk density, acid
saturation, soil texture and colour. These factors are enough, in this data set to distinguish taxa, for example Oxisols from Mollisols, without the need for other identifying tests or properties creating a quick and relatively accurate method of allocating an unknown SPD. Using a full set of 23 properties and 18 depth increments, a precise place among the soil groups can be calculated if necessary. The groups these soils fall into are given a designation and description of their own, forming a perfectly reasonable taxonomic key and system. The result is a new, multidisciplinary taxonomic system that is simple, transferrable and familiar to its constituent taxonomies. It is detailed enough for complex analysis, or alternately, if a few properties are required, i.e. texture and CEC requirements for a specific crop, this information is also easily and simply obtained.

7.1 Introduction and previous work
The classification of different soil types over the world is a difficult exercise as taxonomies and requirements are not the same the world over, leading to siloed systems. Over the last few years, work has begun in earnest to solve this dilemma, attempting to create a truly global classification scheme. Such a scheme, if created, should be simple, repeatable, transferrable and based on concepts or comparisons already understood by the larger scientific community. This research capitalizes on the existence of a huge USDA database which is classified using the US based Soil Taxonomy (ST), combined with TERN which has taxa allocated using the Australian classification system (ASC). These data sets have been standardized to 23 properties at 18 consistent depth increments. Each data set contains the mean values for a given taxonomic allocation e.g. Order, suborder or Great Group. These mean values are referred to as Standardized Taxonomic Allocations (STA). Data are linked with their respective taxonomic systems and they have been combined in such a way that it is possible to retrieve the mean values for a defined set of properties for each STA in each system to the Great Group level. These basic descriptions can be called soil taxonomic allocations (STAs). With the use of spline functions, depths can be harmonized and direct taxa to taxa comparisons made. These comparisons have created a point of reference between two very different taxonomies. This point of reference is the basis for numeric comparisons. Numeric techniques such as distance matrix calculations combined with convex hull overlap area calculations have already been used to highlight
similarities and differences between ASC and ST. This taxonomy to taxonomy comparison has reaped rewards in discovering details such as the taxonomic level each system tends to operate on (Chapter 6), also highlighting areas where each respective taxonomy have gaps. Overlaps of convex hulls have again been used to glean details on how both taxonomies function, their strengths and their potential weaknesses. It is possible to extend the gains so far, using numeric analysis to find if these two taxonomies are linked by any simple sets of properties or principles that can then be applied to a combined taxonomic system or soil profile description (SPD). This would assist in the creation of a taxonomy, which is global, repeatable and simple, yet does not conflict with any current methodologies.

7.1.1 Creation and use of a combined database
A combined database (TRex) was created by adding STAs from ST and ASC together. Upon creation it was obvious that TRex is large, cumbersome and owing to the fact that both taxonomic systems occasionally describe STAs that address similar problems. TRex contains STAs that are so mathematically close as to be considered near duplicates. The data was therefore optimized using a thinning algorithm in which taxa from both systems were selected and arranged in such a way that duplication within and between systems was kept to a minimum and the majority of the gaps in both taxonomies were addressed. The methods used are simple and repeatable, so with this new set of reference taxa that spans much of the soil data space, it is now possible to identify via principal component plots where there are large aggregates of STAs and also where there are gaps in soil taxonomic space. Taxa from other taxonomic systems can be recruited to further augment this database if necessary. This augmented set of reference taxa will be scrutinized in detail for the purposes of the creation of a world taxonomy, one free from the biases of any localized issues. The supplementation of data from other taxonomic systems has been demonstrated to be possible from this exercise, but is beyond the scope of this study.
7.1.2 Taxonomic keys and computer aided classification

In creating a new taxonomy, a method for understanding it must also evolve. For the majority of taxonomic systems, there is a taxonomic key, a set of basic binomial questions that progressively eliminate or elucidate particular reference taxa. These are simple but highly effective methods of determining the most appropriate STA to a great amount of detail with very few questions. Computer power can assist this process. An example of the power of using binomial questions digitally in a decision tree to achieve an accurate specific answer can be seen online with the web akinator (http://en.akinator.com), a computer program designed to identify celebrities within a small number of yes/no questions. While the purpose of this program is trivial, the power of its capabilities, achieving an accurate answer from such simple principles is obvious. It can be seen from other examples such as machine learning and cluster analysis (Anderberg, 2014) that there are computer based methods that can take much of the hard work out of identifying and describing taxa. Calculation of Euclidean distances on principal components is a powerful concept when considering the comparison of taxa for the purposes of creating a world taxonomic classification system.

7.1.3 Distance comparison

It is possible, in the field of soil analysis and taxonomy to compare Euclidean distances between the principal components of a soil profile description and the STAs described in a combined database. While this is simple compared with many statistical analyses, it provides a user with a point of reference between a given profile and the combined experience of (in this case) the wisdom that resulted in two independent taxonomic systems, without said user being well versed in any individual system. The Euclidean component approach is useful for general descriptions and systems, but it needs to be noted that the idea of straight Euclidean distances implies a broader, fuzzier methodology. Distances are calculated on all properties, rather than two or three critical cut-offs to distinguish for example, an Order, and these distances are continuous rather than discrete. A hypothetical taxonomy therefore requires a pH to be over a given threshold, sand fraction to be below another threshold and so on. This is not so when using distances. STAs in the database are derived from the mean values of several SPDs of a given taxonomy which have been calculated at regular intervals to a specific depth.
This means that there is the possibility of the closest unknown SPD to a given STA having a close distance to another STA that has characteristics which would typically exclude it from that group if it were formally described. This is not necessarily a bad thing, knowing, for example an ASC Dermosol is very close to an ST Vertosol would be useful. Dermosols are defined by absence of a clear or abrupt B horizon, which could imply that it has a high clay percentage for the entire depth of the profile, and the difference between the Dermosol and the Vertosol would be in whether there was a sufficient quantity of smectite for shrink-swell behaviour. The ASC currently has bleached-vertic and vertic Great Group designations which are applicable to Calcarosols, Chromosols, Dermosols, Hydrosols, Kurosols and Sodosols, but these definitions can be subjective. If a Vertic Dermosol for example, was more easily identified or it were known how close to a Vertosol the taxa was, that would be useful information for agriculture as a large amount of Australian productive crops are Vertosol derived. Distance calculations can determine similarity, and compare this with its similarity to other groups as a further aid to decision making, giving a set of comparable soil types and a number demonstrating how alike they are.

7.1.4 Dendrograms and cladistics

Decision making for complex assemblages of organisms and minerals, as can be found in soil communities can be done using cladistics or dendrograms (Miltenyi et al., 2015), a manner that other taxonomies such as biology tend to use (Figure 49). Groups that have a common set of properties or principals can be identified, and the properties that associate them can be made into a diagnostic property cluster for that group. These are different from traditional taxonomic keys as the properties are centered around a central value, and it is possible to have one errant property within the cluster, so long as it does not stray too far from the group mean. With this property cluster, an unknown can be quickly allocated on the strength of three or four attributes before more expensive or time consuming tests are carried out. New individuals or groups can be inserted to the same dendrogram structure via distance calculations. The dendrogram can be recalculated if necessary, but this may result in reorganization of the entire structure and subsequently, the taxonomy. When there are taxa that are remote, the use of Euclidean distances will provide a description of the taxa it is closest to, but if the
distance is enormous, it could be a tenuous match, such as Aridisols being closest to Histosols. These would be taxa that are mathematically closest, but in reality not at all alike. The only consolation being that the taxa is not like anything else either. In data space, the mathematical solutions e.g. fuzzy k means with extragrades or akromeson, can resolve outlying individuals or clusters, but an individual SPD would be difficult to allocate in this manner. The most optimal solution for this problem would be the addition of taxa from other systems, which occupy voids within taxonomic property space. This was the purpose of the combination of ASC and ST, and is the rational for later work, hopefully combining taxa from other systems.

Figure 49. The biological tree of life. A commonly used taxonomic structure for the description of biological organisms. (Letunic and Bork, 2006)
7.1.4.1 Studying relationships between tiers and individual units of a taxonomy

More advantageous than simple distance calculations are relationships between groups based on distance. The use of individual and group distances has been known for some time in a variety of scientific endeavors, a useful version being the Ward minimum variance method (Murtagh et al., 2014) for dendrograms. This is a technique that calculates the variances between individual distances and groups them together, then calculates the variances of these groups, grouping the groups and iterating until no differentiation exists, most often represented graphically with a tree like structure. The structure is such that the relationships of individuals within a group or taxonomy are easily identified, and can be made to any level of detail. These mathematical structures are useful because although the calculations are performed using variances of grouped properties, common characteristics in and between groups emerge that can be identified, providing a quick, mathematically robust method of determining common themes within a data set. The individual taxa of ST and ASC can be grouped according to common principles which do not necessarily pertain to the taxonomy of origin. This has therefore, the potential to create a relatively bias free tree in which groups can be compared on sets of common characteristics.

7.1.4.2 Studying relationships between taxonomic systems

Applying these ideas to an unfamiliar taxonomic system, a rough dialogue between the unknowns in the system and what is known in the combined database helps to build a picture of both the soils that the system describes and the processes by which the taxonomy evolved. An unknown STA or SPD can be compared with individuals or the group as a whole and comparisons to any level of detail made. There are a few extra steps required when a group is isolated from the dataset, determining first the properties, then possibly the pedological processes that lead to the differentiation. This opens the way for more detailed investigation, which can then ascertain the operational tier from this information. This converts a relatively subjective allocation into an objective taxonomic class. An appropriate guide to the allocation of levels on future additions are the tiers that are already present in ST and the ASC. These are somewhat arbitrary, but this can be changed at any time, so if an objective
method of determining the correct tier arises it can easily be adopted. Estimating the exact number of
tiers is therefore less important than ensuring that the relationship between all taxa is correct.

### 7.1.4.3 Understanding the tier structure of taxonomic systems
The ASC and ST both commonly operate on tiers, which are designated Order, Suborder and Great
Group. While the descriptions and priorities behind these common designations sometimes differ
greatly, and there are more tiers available for both taxonomic systems, these tiers have been shown to
be useful for describing the capability and usefulness of taxa to a reasonable amount of detail. Ward
dendrograms are more complex because they are calculated on variance. This means that each branch
will split into branches that can be ordered by their level of variance within certain tolerances. In the
case of soil taxonomic groups and Wards calculations, cutting directly to level will produce huge
variation and therefore greater differentiation between groups containing Podosols and Oxisols, but
where massively varied taxa are concerned, such as Histosols, Mollisols and even to a lesser extent,
Vertosols, the groups will be larger and more homogeneous. Splitting a taxonomic diagram into three
levels is too simple. The best approach therefore, is to either split according to level, group sizes or
homogeneity. In the case of a complex data structure such as a combination of ASC and ST, semi-
 supervising this process is wise. It should be possible to create between eight to ten groups out of this-
roughly equivalent to Order in ST or ASC. These should have common properties and may even be
able to be linked to common taxonomic principles. If so, naming according to these principles should
be considered. Such understanding greatly assists when addressing the threats to the underpinning
principal behind water and food security, soil security (McBratney et al., 2014) as the likelihood of
achieving common ground in analysis and policy making is increased and is not lost in confusing
taxonomic language.

### 7.1.5 Taxonomic designations
With the concept of tiers and detail established, the more understandable elements of taxonomy need
to be addressed: designations. The system itself is developing to be a robust soil comparison method,
but ease of use requires familiar elements. Naming the tiers and individuals allows pedologists to give
each unit and or group a metaphysical space in which the soil exists. In biology (Simpson, 1945), Latin names are given to each taxonomic level, and the convention conforms to defined rules which are revised from time to time to ensure the taxonomy is consistent with current ideas. IUPAC (Muller, 1994) uses a more systematic idea of taxonomy, with a defined methodology for naming a particular hydrocarbon by identifying the parent chain of a given molecule (usually by length), naming the main functional group, identifying side chains and so on. The convention has scope for tricky issues such as double bonds and arene rings, all which combine in a methodological system by which a single word precisely defines a molecule, and this word can be used anywhere in the world. ST in many ways reflects this idea, having a system in place, which defines diagnostic horizons and characteristics (a.k.a. differentiae) that include definitive properties and form from systematic processes in similar settings. The descriptive key required is enormous, having many tests, comparisons and descriptions, which cover a range of attributes. This is unsurprising as soil as a system is a more convoluted concept than the order in which atoms are arranged. The difficulty in realizing an IUPAC style system for soil analysis is demonstrated by the complexity that Soil Taxonomy exhibits. In attempting to cover all the taxonomic and pedological options, Soil Taxonomy emerges as a book, which is over 860 pages in length. While ST is one of the best descriptive schemes available, understanding it can be daunting. Therefore, this research results in the challenge of creating a naming convention that is simple, yet covers the myriad of differing biological combinations along with the temperature and time scales that soil descriptions tend to bring with them (McBratney et al., 2003). Identifying key groups in a dendrogram has already been discussed as possible. Separation by common property combinations and allocation of a systematic name based on the alphabet or some other such logical sequence should be appropriate.

7.1.5.1 Naming convention
IUPAC is a good method for describing molecules but it is not appropriate in a soil classification scheme as soil is more complex. ST names are difficult to disaggregate into individual levels, as there is a human element attached. The names themselves, while systematic in nature are not systematic enough for the tiers of ST to be easily extracted. Some descriptors have the spelling changed so the
word is more easily pronounced. While this is good for humans, it is more difficult to do automatically. The easiest example of this is the Order “Oxisols”. This order ends in “OX”, a 2 letter designation e.g. “Hapludox”. Most databases which deal with ST Great Groups typically record the taxa in the plural form which does not change for a taxa ending in “OX”. Every other Order in ST when recorded this way has two unique letters which end with an “s” e.g. “Haplosaprists”. Identifying this sequence of letters automatically requires either a list of recognizable sequences, or a script which is capable of recognizing the “OX” as a unique case, which is not impossible but adds to the complexity of some already complex algorithms. There are many more convoluted examples of names that are easy for humans but hard for computers. This system is therefore a non-systematic, systematic system. Sorting by these designations require extra scripting effort should it be done by computer. The ASC names, as collected from TERN, by comparison are a series of three two letter descriptors e.g. “CA BD DY” (describing in this case a calcic pedcalcalcosol). These are easy to extract with computers but are hard to understand from a human perspective. If analysis and understanding is to be maximized, any convention created should cater for computer recognition and human understanding, if necessary having a name for humans and a designation for computers.

7.2 Aims
The creation of a simple common taxonomy should be possible with the following goals:

- Create a dendrogram from a combined culled database that sequentially groups STAs logically with common, easily recognized characteristics.
- Identify suitable nodes on this dendrogram that can be used to create unique groups for a taxonomic system, identify the common properties or sequence of properties associated with these nodes.
- Create a logical nomenclature based on the position of nodes in the calculated dendrogram and the properties therein.
7.3 Methods

7.3.1 Ward dendrogram of PCA
The reference group of STAs can be compared by distances to any group or SPD. This comparison can be exploited to create groups with common properties. The first step in a Ward dendrogram is the creation of a distance matrix. This can be achieved in the basic packages for R. Wards dendrogram (Sneath and Sokal, 1973; Anderberg, 2014; Murtagh and Legendre, 2014) measures the variance of individuals and then groups tiers by differences in this variance. Individuals are linked to groups, groups are linked to supergroups and a tree diagram can be produced from this. All groups can be easily referenced to the main diagram creating a relatively easy to follow plot of taxonomic relationships. Groups can be determined to any level of variation or complexity.

7.3.2 Finding groups and determining property clusters
In the calculation of dendrograms, the height of the groups is also calculated. Heights in Ward dendrograms are an indication of the similarity between groups. The minimum variance method tends to split groups by variation, so on a given level there will be a group with large variation relative to the other groups, and when groups split off, the proceeding groups have a high variation group and so on. This can be seen by the height of the split in relation to the next. It can be seen from this that using height alone will not capture groups adequately. In creating groups from a Ward structure, there needs to be a level of supervision.

7.3.3 Determining the property differences between groups
The differentiated categories can be defined by differences in properties. As the dendrogram is calculated from a distance matrix produced in PCA space, the original properties need to be grouped according to this dendrogram and individual properties compared between groups. The procedure for this is as follows:

1. Identify groups within the dendrogram according to tiers, identify the STAs within each group.
2. Organize original taxonomic designations and property data according to dendrogram groups.

3. Standardize all property variables within the groups.

4. Calculate standardized property means, determine the difference between these standardized means.

5. Sort from most different to least different property means.

6. The properties with the largest standard differences between means are the properties that are the cause of differentiation between groups.

As with most cluster analysis methods, no individual property is the sole cause of differentiation, it is typically the interactions between groups of properties which are responsible for the various groupings. How large these groups of properties are, and how much they vary is difficult to determine. Currently the first few properties are selected as a guide, in the knowledge that a more formal process will be developed as part of a future study.

7.3.4 Systematic name

The dendrogram is split by twos, this means the simplest method of identifying groups is to use a 1 or a 2 to determine the position on the branches of the tree (Figure 50). The primary groups and therefore the source of the most variance contains 2 groups designated 1 and 2. The next level has four groups which are 11, 12, 21 and 22. The next tier down has eight groups with three letter designations and so on. This can be repeated until the groups are differentiated to the individual STA.
Dendrogram designations

Figure 50. Demonstration of nomenclature of groups prior to group names. This shows how a dendrogram can have groups catalogued in a logical fashion.

This method is easiest for computer aided analysis as it is logical, but its designation is difficult for a human to remember. The best practice in this case would be to split groups by properties, then determine a name which explains these properties in a broad sense. It would be desirable to have each name in each tier the same number of letters, as this aids in the automatic recognition of elements within each group, thus creating a system that is equally recognizable to humans and computers. Although the automatic designations are adequate, future studies will aim to develop better nomenclature.

7.3.5 Link back to original taxonomy

Any nomenclature system that emerges needs to be simple, repeatable and compatible with other systems. Currently the data available pertains to ST and ASC. A convergent/emergent system needs to be as familiar as possible to its component taxonomies. If distance in PCA space is used, the resulting taxonomy is determined by the taxa that have been transcribed within. Any analysis should easily be linked back to the original STA that was used to create the classification. This way adding other taxonomies increases understanding, by demonstrating in-group taxa that each individual within the new taxonomy is closest to and therefore has the most properties in common. The simplest method is to retain the original taxonomic name of the STAs at the end point of the classification. This means
that when allocating a new unknown, there will be multiple examples next to it that can be seen as analogous, or alternately a distance matrix can be used to identify the mathematically closest STA. One of the side benefits to this is the combination of the speed of the ASC and the functionality of ST. An unknown SPD could, theoretically, be identified using the ASC and the low data approach this entails, then the closest ST designation identified. This description could be later amended with the inclusion of lab data. A key can be constructed from the properties that differentiate. Using these properties in the dendrogram key, a user can understand the broader concepts that link an individual STA in the group to the others. This way there is an understanding of the unknown by virtue of a close familiar STA, and there is also a property connection via to the property key. Also, if there is a group of Orders that sit within a calculated group then what is known about these specific Orders can guide the identification of the properties within the group.

7.3.6 Depth

Depth is a problem in that it is an extra dimension to add to all the properties, making simple, easy to understand associations difficult. Depth needs to be taken into consideration when assessing differing taxa. Typically, rotational values of PCAs are summed for all depths to make data easier to understand but this falls over if the difference between a profile is a specific property at a specific depth. When measuring maximum movement of a property to determine which property differs most in the groups, the depth of the property is associated. This means that it is possible to say that the property with maximum difference between group x and group y is e.g. calcium at 100cm. Although this is possible, typically there are 10 or more properties with similar influence, and there can be several similar depths to a given property. Determining when one group of properties should be taken as a group or a series of individuals is a topic that requires further investigation. For the time being, a minimalistic approach is taken to depth, concentrating more on the individual properties. This gives a more coarse determination between groups but that determination is still obvious. Additionally, horizon depth and horizon thickness are similar concepts, and previous research (chapter 3) has demonstrated that horizon thickness is important.
7.4 Results

7.4.1 Dendrogram
The TRex data was converted into a Ward dendrogram (Figure 51). The relationships between each SDA from each taxonomy of origin are similar to that demonstrated in a two dimensional PCA biplot, but represented in tree and branch format. The nodes can rotate, meaning that STAs that are in differing groups appear next to each other, but the association may not be strong. The only associations that can be demonstrated are within groups. Each group is identified with a “G” followed by a binary number indicating where it appears on the Ward dendrogram (Figure 52).
Figure 51. Dendrogram created from combining ST and ASC. This is a diagram showing all the relationships between included STAs from ASC and included STAs from ST.
Figure 52. Dendrogram groupings. This figure shows groups produced from dendrogram and a simplified demonstration of differentiating properties. The two taxa at the top are the most diverse, defined by Gypsum. BD is bulk density, Asat is acid saturation or acidity, Red green and blue are colours, sand% is the sand fraction, Gypsum is gypsum fraction, Silt% is silt fraction, ESP is exchangeable sodium percentage and Clay% is the clay fraction.
7.4.2 Properties that separate groups
The dendrogram can be split many ways. In this instance the dendrogram has been split it into 10 main groups. The rationale behind each group is a mix of property definition and groupings of similar STAs. These groups seem to define what is known about pedological processes in the United States and Australia.

7.4.2.1 G1 and G2
G1 and G2 are the highest level of the taxonomic tree. This group set is easiest to differentiate, having large variation with gypsum at all levels. G1 (Petrogypsids to Haplogypsids) have high levels of gypsum at all depths, G2 (Cryosaprists to Aquiturbs) have low gypsum in all layers. This grouping possesses the most variation. G1 exists in a taxonomic space far away from the rest of the soil groups, meaning a decision has to be made on how to properly classify these gypsum rich taxa. For the time being they are a useful sink for taxonomic variation, even if the group itself is never formally recognized.

7.4.2.2 G21 and G22
G21 and G22 are the next highest level of the taxonomic tree. This group set is split by extreme levels of organic matter which typically separates most of the acidic Histosols and Gelisols from the rest of the taxa. From a pure mathematical perspective, there is clear separation from G21 and the other Histosols, but a close examination of the dendrogram places the remaining Histosols, and possibly three Australian Organosols in a remote space between G21 and the rest of G22. Also the group G22 is defined by large differences in variation down the tiers, as are the histosols and organosols in G21. This could be seen as a strong argument to group them all together. This hybrid group will have common properties, but the cutoffs would be at a much finer scale than the rest of the soil groups. This is one of the inherent difficulties associated with organic, variable soils. Assuming the groups need to be mathematically distinct, the differences between G21 and G22 are bulk density at 0-40cm and colour at depth. At 35cm to depth, one group would be redder than the other. G21 (Cryosaprists to
Haprendolls) have high relative bulk density to 40cm and are more red past 40cm, G22 (Sulfisaprists to Aquiturbels) have low relative bulk density to 40cm and are less red past 40cm depth.

7.4.2.3 G211 and G212
The next level of the tree is the difference between most Andisols (with some Histosols and Organozols) and the rest of the soil classes. Group G211 which contains mostly Andisols is defined by high acid saturation at most depths. From its position in the dendrogram it can be seen that a relatively minor difference in properties, or possibly calculation method would have seen this group join up with G22, the difference is minor and variation at this level is still large. G211 (Cryosaprists to Endoaquands) splits off from G212 (Grey Kandosols to Haprendolls) almost entirely by high base saturation at all depths, with high organic carbon for the first 15cm as well. G212 has average carbon and acid saturation for these depths.

7.4.2.4 G2121 and G2122
The level of G212 can be further differentiated, but G211 is highly variable, and further differentiation is not reasonable. This puts G2121 and G2122 into the next grouping. G2121 contains the majority of Australian taxa with the exception of Organozols and Vertisols. There seems to be a level of pedological development displayed in these groups, the Australian group having either extremely young or extremely old taxa varieties, while G2122 may exist in an ideal zone of pedological development, having soils which in the main, demonstrate properties that would be considered favourable for agriculture. Chromosols, for example, are in G2121 and are often cultivated. G2121 (Grey Kandosols to Kandihumults) can be differentiated by above average sand at all depths while G2122 (Sulfaquents to Haprendolls) can be determined by below average sand at all depths.

7.4.2.5 G21211 and G21212
G21211 (Grey Kandisols to shelly Rudosols) are differentiated by light colour for the majority of depths. This group contains mostly well-developed soils. The exception is a Suborder belonging to
Tenosols, an Australian Order not necessarily associated with profile development. Upon investigation, the Sub Order has an organic, very shallow horizon bounded by a hard pan. This unique taxon would probably fit more with well-developed soils than under developed soils, and should provide a place within an emerging taxonomic system for soils that have somehow become under developed and well developed at the same time. Of the ST groups that are in this zone, there are a small number of Alfisols, but there are larger numbers of Ultisols, and this fits the general pattern. G21212 (Psammaquents to Kandihumults) can be determined by dark colour at the same depths. This group can be further split into 212121 and 212122.

7.4.2.6 G212121 and G212122
The basic defining feature of these two groups are pedological development. Among the group with rudimentary soils, G212121 (Psammaquents to red Kandosols) are a small selection of Australian and US Podosols, presumably Podosols that have been developed to the point of almost pure sand, grouped with Rudosols and Tenosols. The Australian Podosols are in the rudimentary group. This is almost certainly because of overdevelopment. A feature of Australian Podosols is the presence of a large A2 horizon i.e. a horizon that has probably had everything except sand washed out of it. In these circumstances finding Podosols with anything but large sand fractions and low fertility is difficult, making differentiation from Rudosols nuanced. It is possible that finding badly developed Podosols may cause some consternation, as these are typically described as soils with a Spodic horizon, which is usually a sign of advanced pedological processes. Also included in most definitions however, is the caveat that this Spodic horizon can be easily perturbed. Collecting data or samples on heavily cultivated Spodosols may well be an exercise in sand data mining. If data is collected after disturbance, there would not be much separating them from Rudosols or Entisols. The main differentiating properties in G212121 are sand fraction and darker colour than G212122 (Cryaquands to Kandihumults) which also has a high sand fraction. This group contains typical podsols, cast with Ultisols/Oxisols, a few Andisols and the Ferrosol/Oxisol complex. This group further subdivides into Podosols/Entisols/Andisols and Oxisols/Ferrosols/Ultisols.
7.4.2.7 GX and GY
In the middle of this dendrogram, just below G2122, a highly diverse group has split off and is designated GX because of the large variability within. It contains Andisols, Aridisols, Entisols and Inceptisols. This group, having massive diversity and low representation exists in a sparse section of the PCA environment. This is one of the areas that would be assisted by further inclusions from other countries and data sets. This group splits from the main body of soils by a moderate concentration of Gypsum. From the first split in the dendrogram, it is known that where there is gypsum, there is variability. Another group, dubbed GY cleaves from this section. GY is divided into GY1 and GY2.

7.4.2.8 GY1 and GY2
GY1 (Vitraquands to Fragiaquults) differentiates from GY2 (Anhyorthels to Leptic Tenosols) by high silt, and contains mostly Mollisols, a large number of Alfisols, Ultisols and a small number of Entisols and Inceptisols. The incorporated Inceptisols have a reasonable carbon percentage, thus their inclusion in these moderately to well-developed soil groups.

7.4.2.9 GY2X
GY2 has another sub group which has massive variation. It contains Gelisols and Andisols as well as a mix of other Orders. This eclectic mix is why it is dubbed GY2X. It is dominated by a large mean ESP throughout all depths. The average ESP has been generously set at 10 but many of these groups have consistent ESPs of 15 and over. This branching from a variety of other Orders indicates that sodium may be a higher priority property than is currently recognized in ST. It should be noted that even with the high sodium content of these soils, these Natric STAs are still distinct from the Australian Sodosols.

7.4.2.10 GY21 and GY22
GY2 can be split into GY21 and GY22. GY21 contains exclusively Vertisols and Vertosols. GY22 contains primarily Mollisols and Aridisols. The defining criteria is a large clay fraction at all depths (for GY21) as opposed to an average clay fraction at all depths (for GY22). The nonsequitous marriage between the Aridisols and the Mollisols can be understood by the fact that both these groups
can have high base saturation. “A” horizon thickness which is present but not easily expressed in this
data set can create a firm differentiating factor here, but in the final few branches, the Mollisols are
separated from the Aridisols. Whether this is separate enough is a matter of opinion. Should it be
proved that these soils need to be differentiated further up the tree, they can be. Horizon thickness can
be weighted, and this factor has been identified and used successfully in previous chapters.

7.4.3 Creating a classification
There are 10 main groups created with this dendrogram method. Gypsum isolated two taxa and they,
for the time being are undefined. The current groups therefore have dendrogram designations: G211,
G21211, G212121, G212122, GX, GY1, GY2X, GY21, GY22 and G22. As a demonstration, the
means of all the data can be extracted (Table 1 and 2), these means compared with the properties that
differ the most between groups and a taxonomic key can be created.
Table 23. First table of property means. This can be used as a reference when considering differences in properties between groups.

<table>
<thead>
<tr>
<th>Depth</th>
<th>Gravel</th>
<th>Clay</th>
<th>Silt</th>
<th>Sand</th>
<th>R</th>
<th>G</th>
<th>B</th>
<th>H₂O</th>
<th>Ice</th>
<th>BD</th>
<th>OC</th>
<th>Carb</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5cm</td>
<td>5.20</td>
<td>20.23</td>
<td>32.08</td>
<td>42.93</td>
<td>0.37</td>
<td>0.30</td>
<td>0.25</td>
<td>0.19</td>
<td>0.00</td>
<td>1.20</td>
<td>4.90</td>
<td>3.83</td>
</tr>
<tr>
<td>5-10cm</td>
<td>5.43</td>
<td>20.93</td>
<td>32.17</td>
<td>42.76</td>
<td>0.39</td>
<td>0.31</td>
<td>0.27</td>
<td>0.19</td>
<td>0.00</td>
<td>1.24</td>
<td>4.29</td>
<td>3.95</td>
</tr>
<tr>
<td>10-15cm</td>
<td>5.91</td>
<td>21.81</td>
<td>31.87</td>
<td>42.49</td>
<td>0.42</td>
<td>0.33</td>
<td>0.28</td>
<td>0.19</td>
<td>0.00</td>
<td>1.29</td>
<td>3.64</td>
<td>4.17</td>
</tr>
<tr>
<td>15-20cm</td>
<td>6.23</td>
<td>22.72</td>
<td>31.55</td>
<td>42.21</td>
<td>0.44</td>
<td>0.34</td>
<td>0.29</td>
<td>0.20</td>
<td>0.00</td>
<td>1.33</td>
<td>3.17</td>
<td>4.44</td>
</tr>
<tr>
<td>20-25cm</td>
<td>6.55</td>
<td>23.56</td>
<td>31.24</td>
<td>41.73</td>
<td>0.46</td>
<td>0.36</td>
<td>0.30</td>
<td>0.21</td>
<td>0.00</td>
<td>1.36</td>
<td>2.85</td>
<td>4.73</td>
</tr>
<tr>
<td>25-30cm</td>
<td>6.87</td>
<td>24.21</td>
<td>30.92</td>
<td>41.51</td>
<td>0.48</td>
<td>0.38</td>
<td>0.31</td>
<td>0.22</td>
<td>0.00</td>
<td>1.38</td>
<td>2.58</td>
<td>5.07</td>
</tr>
<tr>
<td>30-35cm</td>
<td>7.19</td>
<td>24.83</td>
<td>30.65</td>
<td>41.31</td>
<td>0.49</td>
<td>0.39</td>
<td>0.32</td>
<td>0.23</td>
<td>0.01</td>
<td>1.39</td>
<td>2.28</td>
<td>5.35</td>
</tr>
<tr>
<td>35-40cm</td>
<td>7.46</td>
<td>25.25</td>
<td>30.37</td>
<td>41.27</td>
<td>0.50</td>
<td>0.39</td>
<td>0.32</td>
<td>0.23</td>
<td>0.01</td>
<td>1.41</td>
<td>2.13</td>
<td>5.52</td>
</tr>
<tr>
<td>40-45cm</td>
<td>7.65</td>
<td>25.58</td>
<td>30.09</td>
<td>41.23</td>
<td>0.51</td>
<td>0.40</td>
<td>0.33</td>
<td>0.24</td>
<td>0.01</td>
<td>1.41</td>
<td>1.97</td>
<td>5.70</td>
</tr>
<tr>
<td>45-50cm</td>
<td>7.78</td>
<td>25.80</td>
<td>29.90</td>
<td>41.21</td>
<td>0.52</td>
<td>0.41</td>
<td>0.33</td>
<td>0.24</td>
<td>0.01</td>
<td>1.42</td>
<td>1.85</td>
<td>5.81</td>
</tr>
<tr>
<td>50-60cm</td>
<td>7.91</td>
<td>25.83</td>
<td>29.49</td>
<td>41.37</td>
<td>0.52</td>
<td>0.41</td>
<td>0.34</td>
<td>0.24</td>
<td>0.01</td>
<td>1.43</td>
<td>1.66</td>
<td>5.98</td>
</tr>
<tr>
<td>60-70cm</td>
<td>7.94</td>
<td>25.83</td>
<td>29.20</td>
<td>41.63</td>
<td>0.53</td>
<td>0.42</td>
<td>0.34</td>
<td>0.25</td>
<td>0.01</td>
<td>1.45</td>
<td>1.50</td>
<td>6.08</td>
</tr>
<tr>
<td>70-80cm</td>
<td>7.89</td>
<td>25.68</td>
<td>28.94</td>
<td>42.08</td>
<td>0.53</td>
<td>0.42</td>
<td>0.34</td>
<td>0.24</td>
<td>0.01</td>
<td>1.46</td>
<td>1.37</td>
<td>6.16</td>
</tr>
<tr>
<td>80-90cm</td>
<td>7.66</td>
<td>25.31</td>
<td>28.63</td>
<td>42.92</td>
<td>0.54</td>
<td>0.43</td>
<td>0.35</td>
<td>0.25</td>
<td>0.01</td>
<td>1.46</td>
<td>1.18</td>
<td>6.30</td>
</tr>
<tr>
<td>90-100cm</td>
<td>7.36</td>
<td>24.97</td>
<td>28.31</td>
<td>43.61</td>
<td>0.54</td>
<td>0.43</td>
<td>0.35</td>
<td>0.24</td>
<td>0.02</td>
<td>1.46</td>
<td>1.04</td>
<td>6.36</td>
</tr>
<tr>
<td>100-110cm</td>
<td>7.28</td>
<td>24.48</td>
<td>27.96</td>
<td>44.26</td>
<td>0.54</td>
<td>0.43</td>
<td>0.36</td>
<td>0.24</td>
<td>0.01</td>
<td>1.47</td>
<td>0.93</td>
<td>6.31</td>
</tr>
<tr>
<td>110-130cm</td>
<td>7.26</td>
<td>23.89</td>
<td>27.69</td>
<td>45.20</td>
<td>0.54</td>
<td>0.44</td>
<td>0.36</td>
<td>0.23</td>
<td>0.01</td>
<td>1.45</td>
<td>0.82</td>
<td>6.26</td>
</tr>
<tr>
<td>130-150cm</td>
<td>6.79</td>
<td>23.31</td>
<td>27.29</td>
<td>45.98</td>
<td>0.54</td>
<td>0.44</td>
<td>0.28</td>
<td>0.23</td>
<td>0.02</td>
<td>1.45</td>
<td>0.71</td>
<td>5.54</td>
</tr>
</tbody>
</table>
Table 24. Second table of property means. This can be used as a reference when considering differences in properties between groups.

<table>
<thead>
<tr>
<th>Depth</th>
<th>pH</th>
<th>CEC</th>
<th>Ca</th>
<th>Mg</th>
<th>Na</th>
<th>K</th>
<th>Asat</th>
<th>Bsat</th>
<th>ESP</th>
<th>EC</th>
<th>Gyp</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5cm</td>
<td>6.09</td>
<td>14.23</td>
<td>20.29</td>
<td>4.03</td>
<td>2.94</td>
<td>0.82</td>
<td>14.97</td>
<td>68.31</td>
<td>3.49</td>
<td>3.57</td>
<td>0.40</td>
</tr>
<tr>
<td>5-10cm</td>
<td>6.12</td>
<td>13.87</td>
<td>19.48</td>
<td>3.98</td>
<td>2.79</td>
<td>0.76</td>
<td>14.29</td>
<td>67.62</td>
<td>3.88</td>
<td>3.63</td>
<td>0.58</td>
</tr>
<tr>
<td>10-15cm</td>
<td>6.16</td>
<td>13.51</td>
<td>18.63</td>
<td>3.95</td>
<td>2.60</td>
<td>0.70</td>
<td>13.53</td>
<td>66.79</td>
<td>4.35</td>
<td>3.68</td>
<td>0.62</td>
</tr>
<tr>
<td>15-20cm</td>
<td>6.21</td>
<td>13.21</td>
<td>17.97</td>
<td>3.96</td>
<td>2.71</td>
<td>0.64</td>
<td>13.01</td>
<td>66.28</td>
<td>4.67</td>
<td>3.80</td>
<td>0.72</td>
</tr>
<tr>
<td>20-25cm</td>
<td>6.25</td>
<td>13.16</td>
<td>17.64</td>
<td>4.09</td>
<td>2.79</td>
<td>0.59</td>
<td>12.48</td>
<td>66.01</td>
<td>4.94</td>
<td>3.78</td>
<td>0.76</td>
</tr>
<tr>
<td>25-30cm</td>
<td>6.29</td>
<td>13.19</td>
<td>17.22</td>
<td>4.26</td>
<td>2.90</td>
<td>0.56</td>
<td>12.09</td>
<td>66.07</td>
<td>5.20</td>
<td>3.83</td>
<td>0.79</td>
</tr>
<tr>
<td>30-35cm</td>
<td>6.32</td>
<td>13.27</td>
<td>17.02</td>
<td>4.37</td>
<td>2.85</td>
<td>0.54</td>
<td>11.71</td>
<td>66.14</td>
<td>5.62</td>
<td>3.76</td>
<td>0.86</td>
</tr>
<tr>
<td>35-40cm</td>
<td>6.34</td>
<td>13.46</td>
<td>16.84</td>
<td>4.48</td>
<td>2.95</td>
<td>0.52</td>
<td>11.24</td>
<td>66.30</td>
<td>5.87</td>
<td>3.84</td>
<td>0.95</td>
</tr>
<tr>
<td>40-45cm</td>
<td>6.36</td>
<td>13.44</td>
<td>16.73</td>
<td>4.58</td>
<td>3.00</td>
<td>0.51</td>
<td>10.70</td>
<td>66.43</td>
<td>6.06</td>
<td>3.89</td>
<td>1.00</td>
</tr>
<tr>
<td>45-50cm</td>
<td>6.38</td>
<td>13.44</td>
<td>16.37</td>
<td>4.65</td>
<td>3.04</td>
<td>0.49</td>
<td>10.23</td>
<td>66.75</td>
<td>6.28</td>
<td>3.97</td>
<td>0.99</td>
</tr>
<tr>
<td>50-60cm</td>
<td>6.41</td>
<td>13.46</td>
<td>15.80</td>
<td>4.67</td>
<td>3.02</td>
<td>0.47</td>
<td>9.74</td>
<td>67.05</td>
<td>6.44</td>
<td>4.03</td>
<td>0.97</td>
</tr>
<tr>
<td>60-70cm</td>
<td>6.45</td>
<td>13.61</td>
<td>15.24</td>
<td>4.70</td>
<td>3.07</td>
<td>0.46</td>
<td>9.11</td>
<td>68.14</td>
<td>7.04</td>
<td>3.97</td>
<td>0.98</td>
</tr>
<tr>
<td>70-80cm</td>
<td>6.47</td>
<td>13.68</td>
<td>14.67</td>
<td>4.72</td>
<td>3.10</td>
<td>0.45</td>
<td>8.54</td>
<td>69.00</td>
<td>7.39</td>
<td>3.96</td>
<td>0.96</td>
</tr>
<tr>
<td>80-90cm</td>
<td>6.49</td>
<td>13.89</td>
<td>14.11</td>
<td>4.62</td>
<td>3.03</td>
<td>0.43</td>
<td>8.03</td>
<td>69.76</td>
<td>7.72</td>
<td>3.84</td>
<td>1.05</td>
</tr>
<tr>
<td>90-100cm</td>
<td>6.49</td>
<td>13.91</td>
<td>13.59</td>
<td>4.47</td>
<td>2.93</td>
<td>0.42</td>
<td>7.70</td>
<td>70.09</td>
<td>7.85</td>
<td>3.82</td>
<td>1.00</td>
</tr>
<tr>
<td>100-110cm</td>
<td>6.49</td>
<td>13.77</td>
<td>13.20</td>
<td>4.39</td>
<td>2.94</td>
<td>0.41</td>
<td>7.50</td>
<td>70.08</td>
<td>8.05</td>
<td>3.71</td>
<td>0.96</td>
</tr>
<tr>
<td>110-130cm</td>
<td>6.47</td>
<td>13.71</td>
<td>12.73</td>
<td>4.22</td>
<td>2.83</td>
<td>0.40</td>
<td>7.28</td>
<td>69.02</td>
<td>8.17</td>
<td>3.32</td>
<td>0.96</td>
</tr>
<tr>
<td>130-150cm</td>
<td>6.46</td>
<td>13.67</td>
<td>12.29</td>
<td>4.11</td>
<td>2.80</td>
<td>0.40</td>
<td>7.27</td>
<td>68.71</td>
<td>8.39</td>
<td>3.28</td>
<td>0.94</td>
</tr>
</tbody>
</table>

7.4.4 Organizing the data

Each node in the dendrogram creates two groups. These groups can have their individual properties standardized and compared, and the most divergent properties identified. Once these properties are identified, they can be directly compared with the mean properties table and an identifying property mean estimated. Several estimations can be presented together to give a centroid value for each taxonomic group.

Groups of taxa combined with their defining properties operate in a similar fashion to a taxonomic key. Although they operate as a taxonomic key, they also represent a multidimensional space with fuzzy associations, thus there are no true cutoff values. The property cluster that has properties that
matches most with a given SPD is the probable group to which the SPD belongs. So if an SPD has an acidity (acid saturation) of 10% all depths, 55% sand all depths and has a low blue saturation, but its bulk density is a little under 1, there is still a good chance this SPD is within group G212122. It’s more a measure of distances than hard cutoffs.

7.4.4.1 The taxonomic key

_G22_:  
BD of approx. ≤0.5 g/cm to a depth of 40 cm

_G211_:  
BD from 0-40cm of approx. 1.2 to 1.4 g/cm  
Asat of approx. 50% all depths

_G21211_:  
BD from 0-40cm of approx. 1.2 to 1.4 g/cm  
Asat of approx. 10% all depths  
Approx. 50-60% sand all depths  
Blue colour saturation of approx. 50% all depths

_G212121_:  
BD from 0-40cm of approx. 1.2 to 1.4 g/cm  
Asat of approx. 10% all depths  
Approx. 50-60% sand all depths  
Blue colour saturation of approx. 25% all depths

_G212122_:
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10% all depths
Approx. 50-60% sand 0 - 80cm
Approx. 45% sand 80 - 100cm
Blue colour saturation of approx. 25%

GX:
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10% all depths
Approx. 30-40% sand all depths
Gypsum of approx. 2% all depths

GY1:
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10% all depths
Approx. 30-40% sand all depths
Gypsum of approx. 0.5% all depths
Approx. 50% silt all depths

GY2X:
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10% all depths
Approx. 30-40% sand all depths
Gypsum of approx. 0.5% all depths
Approx. 30% silt all depths
ESP of approx. 10% all depths

GY2I (Vertisols/Vertosols):
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10% all depths
Approx. 30-40% sand all depths
Gypsum of approx. 0.5% all depths
Approx. 30% silt all depths
ESP of approx. 2% all depths
Approx. 30% silt all depths
Approx. 40% clay all depths

GY22:
BD from 0-40cm of approx. 1.2 to 1.4 g/cm
Asat of approx. 10%
Approx. 30-40% sand all depths
Gypsum of approx. 0.5%
Approx. 30% silt all depths
ESP of approx. 2%
Approx. 30% silt all depths
Approx. 20% clay all depths

7.5 Discussion

7.5.1 Grouping
The dendrogram structure identifies groups of properties that explain an underlying feature. This feature could be as simple as sand fraction, but is more likely to be something far more complex. The properties that differ the most between dendrogram groups have been identified. There can be one or two emergent properties that are different to mean changes, which are in fact highlighting a different aspect or aspects of the feature the dendrogram is describing. The differences between properties therefore can only be used as a guide, indicating a larger scale of interaction between groups of
properties or possibly a pedological concept. The generalized property differences are useful however, for the purposes of allocation in the field. There have been 10 basic groups isolated and the underpinning pedological processes can be estimated. More groups can be identified and extracted if necessary up to and including individual STAs.

7.5.2 Groups and pedological/property processes

- G211 contains primarily Andisols, so volcanic and organic activities seem to be the prime movers.
- G21211 contains Ultisols and Australian soils barring types such as Rudosols and Tenosols. This could imply the underlying process is pedological development- these being old, ultra developed soils.
- G212121 contains mostly Entisols and Rudosols- all soils that demonstrate under development. The inclusion of Podosols and some Spodosols makes this group an area for further discussion.
- G212122 contains the remaining Spodosols/Podosols and has Ultisols, Oxisols and Ferrosols. This implies another pedologically well developed soil type, leaning towards processes that bring out oxides.
- GX contains Entisols and Aridisols. This group also has many of the more soluble components such as carbonates, gypsum and sodium. These soils are most likely arid soils.
- GY1 contains Mollisols and Ultisols plus a large proportion of Alfisols. This is probably a cooler climate well developed soil series. This would be one of the groups that seem to occupy an ideal pedological zone for agriculture.
- GY2X are soils that have Natric or near-Natric horizons. They are separate from Australian Natric soils (Sodosols) by larger proportions of clay and base saturation. As such they are not as prone to dispersion.
- GY21 Soils contain almost all the Vertisols/Vertosols. There are also a curious mix of Hydrosols and Aquic soils in this group. A tantalizing theory on this is that these Aquic soils
have good shrink swell capacity, but as they are waterlogged, they don’t develop the same pedological features and retain a massive appearance in their horizon(s).

- GY22 contains mostly Aridisols and Mollisols. The other soils that appear here have a xeric or water limiting climate regime. The differentiation between the Mollisols and Aridisols would most likely be assisted if horizon thickness was added in a way that gave it greater influence.

- G22 contains Histosols and Gelisols. These are most likely cool climate organic rich soils.

7.5.3 Allocation issues
Although the data collation, mean calculation, and distances on PCA method has produced surprisingly robust results, there are areas in which improvements could be made. For example some groups have dispersed over the classification series, and there are measures that could be used to either push them into an independent group or align them more logically to an existing group. The primary groups for which this occurs are Andisols and Podosols/Spodosols.

7.5.4 Andisols
The Andisol class has proved troublesome. The mineral allophane is one of the key identifying factors in this soil class, however, it can be confused with kaolinite when tested using XRD (Parfitt, 1990), which is the reason why the oxalate test is prescribed when allophane is suspected. There is a specific set of diagnostic criteria for Andisols which are rarely used with Australian soils, for good reason. Andisols are not probable in most parts of the continent. This means, however, that an Andisol can be missed. If it were missed, it would possibly appear with Ferrosols or Organosols or possibly another Order. Misclassified Andisols or soils with Andic properties would therefore spread their influence over a variety of inappropriate classifications, making differentiation by dendrogram difficult. This inconsistency can be overcome. The TERN dataset has many other chemical and physical attributes that can be extracted. Properties that can be linked to Andisols can be used to double check suspect taxa. Any taxa that have Andic properties can then be re-designated and the dendrogram run again. A procedure such as this should be able to better define these end points. Having said this, the Andisols
group together relatively consistently. The class is spread over four of ten groups. The majority of these fall into an Andisol specific group, another group is typified by organic content and pH, while the other two are groups which have swept up some of the more variable soils. Considering the potential for mis-classification, the robustness of this outcome is surprising.

7.5.5 Spodosols
The Spodosol/Podosol group retains coherence barring one example for four orders of differentiation. At this point Spodosols split between group G212121 and G212122. Both of these groups contain roughly equal proportions of Spodosols. This differentiation is acceptable, but also understandable as the most likely cause is degradation of the Spodic horizon at the time of sampling. This means that there may not be any differentiating features in what should be a well-developed soil type. The one example that falls outside the main group is Cryods. Frozen conditions and the excess carbon this creates, gives it some variability and affinity with Histosols. There are Andisols in this group as well, and 212122 contains a roughly equal measure of Spodosols and Andisols, suggesting that there is some kind of taxonomic commonality here. Most likely, the measures suggested to better define Andisols will have a flow on effect that sharpens the resolution of Spodosols/Podosols as well. As far as determining the group, identifying Spodosols can be done by checking for low gypsum, average acid saturation, above average sand and dark colour. Group 212121 should be less accommodating to Spodosols and the high difference between Spodosols and the rest of the taxa in the group reflect this, with massive differences in bulk density detected. Spodosols sit much more easily with the other, well developed soils such as Ferrosols and Ultisols, demonstrating milder differences in sand content.

7.5.6 Mapping other taxonomies
Using the ideas of ST and ASC a comprehensive yet simple classification emerges. If the underlying principles hold true, then a soil can be placed into one of these categories relatively quickly using a minimum of information. Using more information such as a detailed SPD, the description can be linked back to the closest STA in the system. The linked STA has a description based in either ST or ASC which if necessary gives a wealth of detail. This has the potential to reduce much of the hard
work in understanding SPDs. Any taxonomy can be placed by distance within these groups. It is therefore conceivable that the broad simple principles can add understanding to obscure taxonomies.

7.5.7 Reduction in conflict
The TRex classification system allows for nations/groups to keep their classification scheme, and use it the same way they always have. Placing a simple qualifier such as the international group at the end of their STA, it is instantly provides a small list of reference taxa to which their STA is closely associated. More importantly, if more detail is present and more explanation required, the distance to the nearest reference taxa can be retrieved. This presents us with a new paradox, where a new taxonomy emerges, that is not designed to replace any individual taxonomy, yet has the potential to be helpful to all, is based on simple principles, yet can describe SPDs to any level of complexity required, can accommodate simple data, or can facilitate complex analysis.

7.6 Conclusion
Two disparate soil classifications were successfully merged by means of PCA, mathematical thinning and addition. A distance matrix of this data was calculated and a Ward dendrogram produced. This dendrogram can be considered an objective tool to determine the relationships in and between various taxa of two classification schemes. It is a simple matter to add other schema to this dendrogram system, as each system is added, more detail on the world’s soils and their relationships between each other and each taxonomy becomes clearer.

7.6.1 Future work

7.6.1.1 Addition of other taxonomies
Significant progress has been achieved combining ST and ASC. Each step has been programmed and therefore can be repeated. This reduces the workload should any other data become available. There are similar data sets in existence for New Zealand, China, Brazil, India, Tibet, Canada and a host of other regions. The addition of this data will assist in the creation of a broader ranging classification
scheme. Such a scheme should be far simpler to compile than the original comparison between ST and ASC. However, each new system, may have differences, be it procedural, unit, data, experimentation method, particle size or other which will need to be accounted for.

7.6.1.2 Fixing Andisols
In this study Andisols were found to be scattered in PCA data space. This is probably because of the lack of a clear Andisol classification in the ASC. There is much data in the TERN database, it therefore could be interrogated to determine if there are ST Andisol classifications associated with any described ASC taxa, or alternately any physical or chemical data that could indicate Andic properties in ASC STAs. If such associations are found, the Andisol like taxa could be subset and the analysis run again. This course of action may help resolve Andisols in general.

7.6.1.3 Fixing Spodosols/Podosols
Podosols are technically a well-developed taxonomic Order. It is incongruous that some Spodosols and Podosols are relegated to groups in which there are Rudosols or Entisols; taxa that are on the opposite end of the soil development spectrum. This could be associated with Andisol dispersion in the soil groups, but also could be because some classifications classify heavily degraded Podosols as such and these may be close taxonomically to sands. Another example is a coastal plain soil with a very thick E horizon. It classifies as an Entisol if the argillic starts below 200cm, but an Ultisol if the argillic horizon starts at 200cm. There are some who would like to see more sand classifications in the ASC, this reallocation of Podosols that are primarily sand to another category may allow for better groupings.
7.7 References


Chapter 8

8.1 Conclusions

From chapter two it can be seen that with new mathematical methods and sufficient data, it is possible to conceptualise a given landscape into categories. These categories reflect more than the sum of data acquired, as evidenced by the expression of geological features that resulted from the akromeson output. This initial success was the rationale for bringing these techniques into the realm of the US based Soil Taxonomy.

In chapter three, the akromeson algorithm was used to partition surface horizon classes. Silty A horizons were delineated in the areas of the United States such as Nebraska and Iowa. To the south, end point horizon classes were discovered. The general theme of the end point horizons delineated by akromeson was a typical horizon with a single property or attribute that was extreme. The reasons for this extreme shift in a single property are unknown, but it could be anthropological. This means akromeson could be used to track anthropogenic land degradation. This success with the analysis of USDA data led to more ambitious objectives.

Chapter four asked the question of whether two extremely different soil taxonomies could be harmonized. Utilizing a Soil Taxonomy data base that had been carefully created by a research team in Hungary, Australian data was standardized and compared for the first time. This methodology is important, and may become the basis for harmonization of soil data the world over. The resulting comparisons required a large number of new mathematical methods. A set of taxonomic distances between individual soil taxonomic allocations was created as well as taxonomic distances between the Orders of differing taxonomies. As a result of the need to analyse these distance matrices, the algorithm for taxonomic accuracy was born. Surprisingly, it was determined that Australian taxa tends to coalesce more strongly together than its US counterpart. It is theorized that as ST is a more general classification, the individual soil taxonomic allocations are further apart in taxonomic space. Also as the ASC uses simple criteria, this makes easier correlations. The climate zones on offer in Soil
Taxonomy proved to hamper its performance in taxonomic space, which may be another reason for more compact groupings within the ASC. The data harmonization however, was a big leap forward, enabling more targeted analysis and comparisons between Soil Taxonomy and the ASC. This can also be applied to other soil systems creating a methodology of harmonization and analysis.

Chapter five took data from both taxonomies and used the distances between Orders as a guide to the similarity between Orders of opposing taxonomies. The individual Soil taxonomic allocations could be used to form convex hulls, delineating the areas in which each taxonomic Order tends to reside in soil taxonomic space. This could be used as a tool for determining the accuracy of the harmonization process. The convex hulls of Vertisols and Vertosols, being nearly identical taxa, were overlaid. The areas they occupied were nearly identical. The same applied to Spodosols and Podosols. These too were taxa that had nearly identical descriptions. Interestingly, there was overlap between ASC Hydrosols, an Australian taxon that is defined by prolonged periods of saturation and ST Aridisols, a US classification defined by prolonged periods of no saturation. This suggests problems both with definitions of Hydrosols and with climate criteria used to define soil taxa in ST. Similarly there was overlap between ST Andisols and ASC Ferrosols (an Australian taxa that shares common features with the ST Oxisol). This overlap could be attributed to the fact that the ASC has no capacity to properly classify ST Andisols, scattering Australian soils with Andic properties all over the data space. This is an important issue. As well as comparing convex hulls, next nearest neighbour distances were compared. Using these distances it was determined that ASC Suborders have the same level of taxonomic generalization as ST Great Groups. This work was crucial to later taxonomic algorithms.

Chapter six was devoted to the standardization and merging of STAs from ST and the ASC. A thinning algorithm was applied to ST centroid data, removing several STAs. The ASC data was then up-merged to the Suborder level. Mean values of all taxa belonging to the same Suborder taken, creating approximations for each Suborder. These Suborder means were then added to the Soil Taxonomy data and the thinning algorithm applied. There were no further taxonomic eliminations.
from this indicating an acceptable level of spacing. Distances were then recalculated in preparation for the creation of a world taxonomy.

The research from chapters two to six coalesce in chapter seven. A distance matrix was created from the recombined, averaged and thinned database. This was then analysed. From the data a Ward dendrogram was produced. From this dendrogram a series of 10 soil groups were identified. These groups were then isolated, their properties standardized, the means taken and subtracted from each other group to determine which properties were responsible for their divergence. From this a summary of divergent properties for each group was created. As it is based on cluster analysis, this property divergence is far more flexible than in most taxonomies, making a group or soil individual “most like” rather than a set of hard cut-offs.

It was demonstrated that data from disparate taxonomies can be standardized and combined. From this a schematic was developed to show how every STA has been placed within this framework and its interaction with every other STA or SPD demonstrated. This new taxonomic scheme is calculated on properties, but the individual STAs are determined by the principles from their original taxonomy. This means that although this system is novel, it is also based on concepts that are validated in the field. The resulting product is a new innovation, yet takes into account the previously developed taxonomies. Once properties are collected and harmonized, the place of any new STD or STA in the soil universe can be made known, its properties are familiar to its originating taxonomy, and its associations with similar taxa provide understanding for others. What is more, the closest STA from another taxonomy can be determined; someone unfamiliar with the closest STA can find the next closest, and if this is unfamiliar, interrogate the data until a suitable point of reference can be found.

### 8.2 Future directions

The work that has been performed to standardize two completely different taxonomies is a unique process. There is a huge amount of data that has been converted from one format/depth increment/
unit to another. The work of checking data, comparing standards and ensuring accurate translation should be an ongoing process.

8.2.1 Andisols
Andisols are rare in Australia and as such there is not much scope for them in the ASC. Just because they are rare does not mean they should not be taken into consideration. The ASC would possibly classify most Andisols as Ferrosols (an Order similar to the ST Oxisols), but this is by no means certain. Undiscovered Australian Andisols, or even soils with Andic properties have the potential to scatter this soil group. The TERN database contains many more properties than the 23 chosen for this study. The next step in creating a world classification is to ensure Andisols are grouped with Andisols which requires the mining of key soil attributes associated with Andic properties, once this is done, the Andic soils of Australia may have to be grouped before the ASC designations are used. This should assist in the creation of an exclusive Andisol group, and may have the benefit of removing Andic properties from where they don’t belong, thus tightening the associations of all the soil Orders in the dataset.

8.2.2 Spodosols/Podosols
There are those in Australia that would like to see more sandy soils recognized in the ASC. This is a controversial topic. I agree that if the definition of a soil requires some kind of evidence of pedological development, then a pile of sand can never truly be called soil, but it is possible that some Podosols in the database are probably not the well weathered, highly developed soils most pedologists may visualize. Agricultural practices may have altered some soils to the point where they are indistinguishable from sand. If these are currently in the database, designated as Podosols, then numerically it would be very difficult to find differences between Podosols and unconsolidated materials. Adding sand classes to the ASC may have the effect of rescuing the taxonomic definition of soil rather than polluting it. As Podosols/Spodosols tend to be highly developed/weathered soils, and can have affinity with ASC Ferrosols, the improvements to Andisols may well assist in the better delineation of Podosols as well in a Ward dendrogram.
8.3 Closing Remarks

This body of research has created a system that, although needing further refinement could be used as a world soil taxonomic scheme. The data as it stands, using all of the methods described, can create logical categories that fit in well with pedological or property descriptions. In the future a superior nomenclature could be developed, possibly encompassing a designation that is systematic and character driven, or using a name that reflects the general characteristics of the taxa within each group.

The procedures highlighted in this body of work, present a novel approach to the classification of the world’s soils. There are a number of different regional classifications schemes in existence, and more importantly, as they have similar data sets as the USDA soil characterization database and TERN they could be easily added to the TReX database. The groundwork has been laid for conversion of these databases into regional soil taxonomic allocations. By understanding the operational level of each classification system and determining which tier of the classification should be utilized, we can harmonize the data with TRex. Soil taxonomic allocations can be calculated at the correct operational level, culled and added to the database. Although this procedure was initially time consuming, the majority of the programming has now been developed and unification of different systems is not only possible, but can be achieved in a short amount of time. This means that it would be simple to add soil taxonomic allocations from other taxonomies, provided there are no regional differences that require extra conversions. Each group that is added to this system makes a more robust, global taxonomic system that should retain its simplicity. End point STA’s, grouped with similar taxa, with a set of properties that define the group from which it came should be easy for most pedologists to understand. The list of properties associated should create meaning for those whose primary concern are the specific property features of a soil, such as the requirements for a specific crop. In conclusion, this is a scheme that will allow for universal communication between soil scientists, whilst preserving regional schemes.