"Mathematics and Archaeology" rediscovered

Michael Greenacre

Universitat Pompeu Fabra, Barcelona, and Barcelona Graduate School of Economics

The book "Mathematics and Archaeology", consisting of 25 chapters by a range of international scholars in archaeology, will be published by Chapman & Hall in 2014. The present document, written as an invited Epilogue to the book, recounts the rediscovery of the book 275 years later by an archaeolinguist. The remnants of the book have been found in the Universitat Pompeu Fabra's dilapidated library, which fell into disuse after books were abandoned in favour of electronic publishing. The archaeolinguist explains how statistical methods found in old texts on quantitative archaeology helped to piece together the basic content of this book, using the words in the chapters as artefacts.

Key words: archaeology, clustering, correspondence analysis, linguistics, multivariate analysis, textual data mining.

"Mathematics and Archaeology" rediscovered

Michael Greenacre

Universitat Pompeu Fabra and Barcelona Graduate School of Economics, Barcelona

It was July in the year 2289 and in Gallia, the northern part of what used to be France a century earlier, they were celebrating the 500th anniversary of the French Revolution. It was also almost a hundred years since World War 4, the so-called "great technological war", which had decimated the world's scientific, technical and engineering community as well as destroying all the digital information stored in the 55 huge data centres scattered across the earth. These data centres, mostly underground, housed the "cloud", in which mankind had entrusted all its data and creative work, and were conveniently concentrated for the ensuing attacks. In a short space of time the totality of electronic records, including all books written after 2100, when paper printing was abolished, were obliterated and lost forever. All that was left of the written word were the millions of books stored in dusty warehouses, called "libraries", as well as some ancient storage media that had long fallen out of use – these were mostly disk drives with very limited storage, less than a petabyte, which had passed from one generation to another as heirlooms, usually quite deteriorated and no longer readable. Many philosophers had forecasted this disaster after the technological singularity was reached in the mid 22nd century, but nobody paid much attention to these "prophets of doom", as they called them.

I worked as an archaeolinguist at the prestigious Mediterranean Centre of Advanced Research in Barcelona. Our job was to try to understand the contents of the remnants of books discovered in various libraries across the southern Mediterranean region, from Catalonia in the west through Occitany and as far east as Lombardy and Venicia. Most books had been poorly preserved after cloud storage became universal, so we were faced with documents that were often in a highly damaged state. Nevertheless, we had excellent facilities, including robotic equipment with the new *Ocula* vision system that had been programmed to page through the documents more carefully than any human hand and to automatically scan and interpret every letter, diagram and table that was still readable. We only had to intervene when the system stopped with a warning that some pages were stuck together, in which case we decided whether we should intervene or omit reading the hidden content, because detaching the pages might destroy them completely.

After several years working on old statistical texts, I was now working on books about archaeology itself, which then had a twofold interest for me as an archaeological linguist. I had previously reconstructed the contents of a book called "Correspondence Analysis and West Mexico Archaeology", published in 2013 by the University of New Mexico[1], and which had been found in the Pompeu Fabra University's subterranean library in Barcelona, having been donated to the library by a statistician who worked there at that time. This book explained how artefacts could be coded and how the data on several artefacts could be compared through a visualization method called correspondence analysis. There were several chapters by another statistician named de Leeuw who explained the methodology, and having decyphered several statistical texts I could more or less follow his explanation and the mathematical calculations involved. So in my spare time I had programmed the method using the **`VerbalR`** language[2] and tested it out. It really was quite impressive how correspondence analysis showed the main differences between the objects as well as the features that distinguished them from one another. So I thought for my next book I would try it out, using the words that I found in the book as archaeological indicators – after all, I was in a sense excavating through these old printed artefacts ravaged by time, and the words were their features.

The next book was called "Mathematics and Archaeology", a book edited by two researchers in Barcelona at about the same time as the one I mentioned previously, i.e. early 21st century. This book was in quite bad condition, with many pages so deteriorated that they sometimes disintegrated with handling. Even the whole of chapter 1 was missing, probably torn out secretly by some reader prior to the year 2100. Nevertheless, the contents pages were intact, and I could gather from them that the book was quite varied in its treatment of the subject, and thus a good example for my empirical exercise. The robotic scanning took a full morning, as I had to intervene several times to cope with many torn and stuck pages, but by midday I had a file of most of the text of the book stored on the atomic RAMchip that I carried permanently

---

[1] *Correspondence Analysis and West Mexico Archaeology: Ceramics from the Long-Glassow Collection*, edited by C. Roger Nance, Jan de Leeuw, Phil C. Weigand, Kathleen Prado and David S. Verity, University of New Mexico Press, 2013.

[2] **`VerbalR`** is a verbal programming language whereby the user describes the algorithm verbally to the computer, and the computer then creates the code while automatically verifying its syntax.

on my wrist[3]. I had scanned different chapters, written by different authors, separately, and soon, with a simple command to **`VerbalR`**, I had computed word counts of all the words in each chapter – this was similar to what I had read in the archaeology book about how quantitative archaeologists counted how many different types of artefacts were discovered at each site. The program also incorporated a lemmatization algorithm, grouping different forms of the same word together into a single unit.

The first thing I did was to list the most frequent words, shown in Fig. 27.1, which was a simple statement to **`VerbalR`**: "plot a horizontal barchart of the 40 most frequent words, in descending order of frequency".

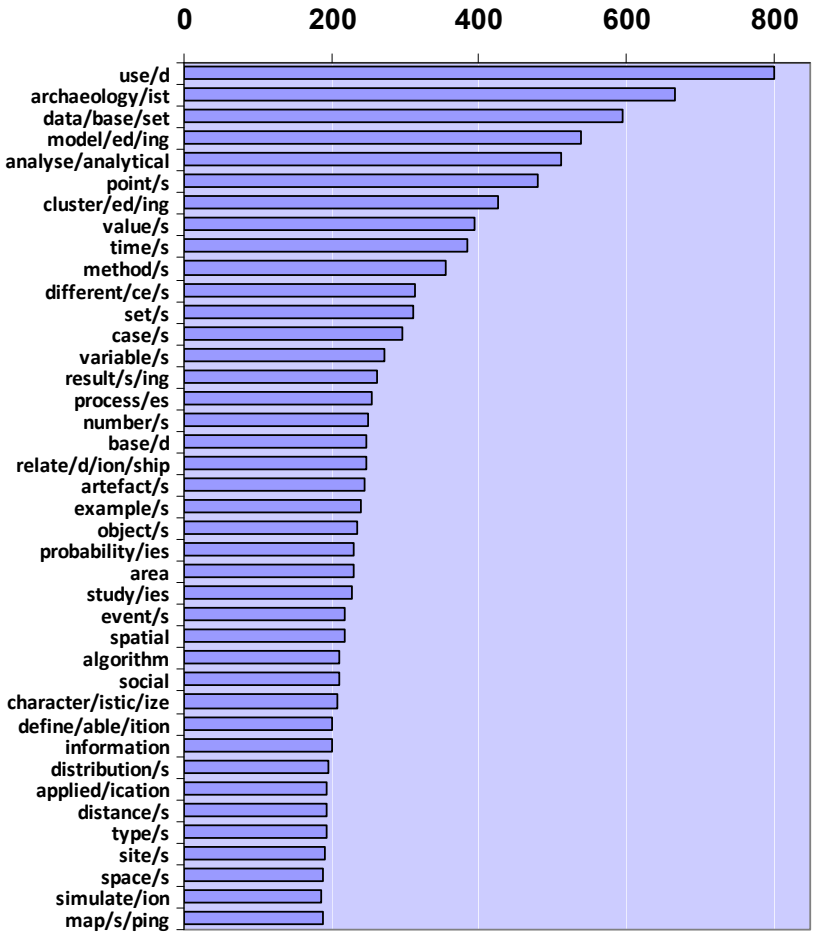

Figure 27.1 Most frequent words (top 40 out of the 751 used at least twice) in the "excavation" of the text of "Mathematics and Archaeology"

[3] Atomic RAMchips are now carried by every human as well as some pet animals. They carry the holder's complete personal information, including personal genome, educational and medical history, as well as a complete record of every transaction conducted by the holder (physical money was not used since the late 21st century). They were the first examples of so-called indestructible digital storage, invented as a consequence of the Great Technological War. Originally, they contained one zettabyte of memory, but nowadays you can find them with more than 100 Zb – most of this memory can be used by the holder for storing data of his or her choice.

The three most frequent terms were *use/d*, *archaeology/ist* and *data/base/set*. This gave me the idea that the book was practically oriented around different aspects of archaeology and that the use of data by archaeologists was a central concept. This was confirmed by other frequent words in this "top 40" list, words such as *analyse/analytical*, *value/s*, *method/s*, *case/s*, *example/s*, etc. The word *statistical* was the 41st, just after *map/s/ping*, with 185 mentions, while the word *mathematical* was mentioned only half as much, 95 times. This led me to believe that the book was more about statistics in archaeology than mathematics. In fact, mathematics as "practised" in earlier centuries had now become a subject more of philosophical abstraction and esoteric debate – statisticians and data engineers had long since gleaned all the useful mathematical results they needed for their myriad of practical problems.

Correspondence analysis would show me how the distributions of words differed across the chapters, so I issued the instruction to **`VerbalR`** to "count the words in each chapter and perform a correspondence analysis on the table of chapters by words". In an instant the computer responded with a verbal summary and description of the results – here I transcribe the essential part:

> ***Correspondence analysis of chapters by words***
> *Total inertia = 2.668*
> *Inertia dimension 1 = 0.226 (8.5%) P<0.001*
> *Inertia dimension 2 = 0.174 (6.5%) P<0.001*
> *15.0% inertia on first two dimensions, but dimensions 3, 4, 5 and 6 also non-random.*
> *38.0% of the inertia on dimensions 1 to 6.*
> *62.0% of the inertia can be considered random variation.*

The way I understood these results is that 38% of the information in the word counts across chapters reflected real differences between the chapters, but I would need to see the results in six dimensions to fully appreciate these differences. Nevertheless, I was interested to see the best two-dimensional view, and told the computer using **`VerbalR`**: "plot the best two-dimensional view of the chapters and connect them in their numerical order; then plot the best two-dimensional view of the most interesting words that distinguish the chapters" – these commands gave the results in Figs 27.2 and 27.3.

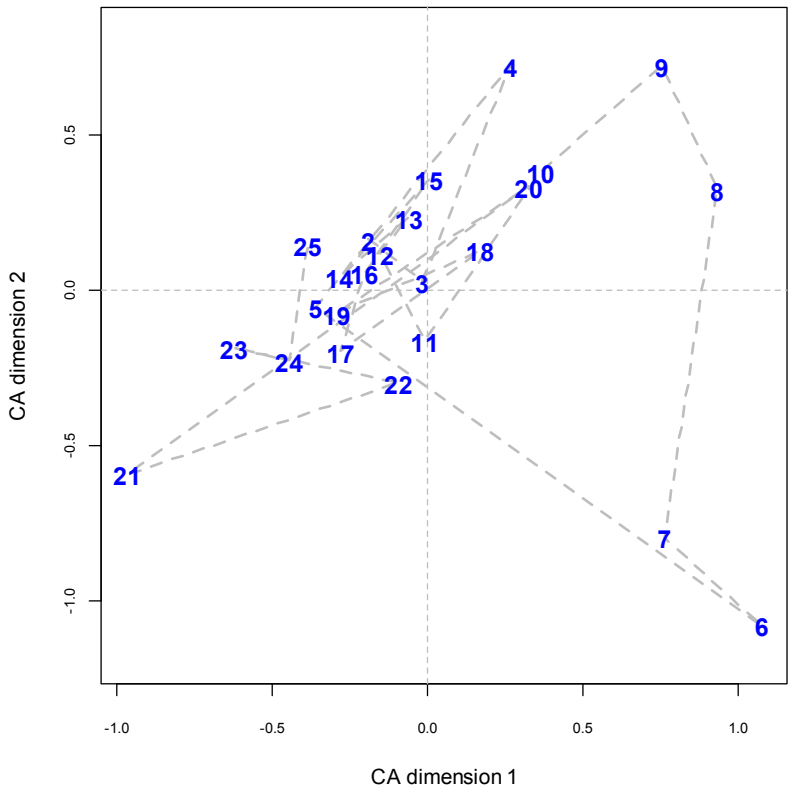

Figure 27.2 Positions of chapters 2 to 25 of the book "Mathematics and Archaeology" according to first two dimensions of the correspondence analysis of their word counts.

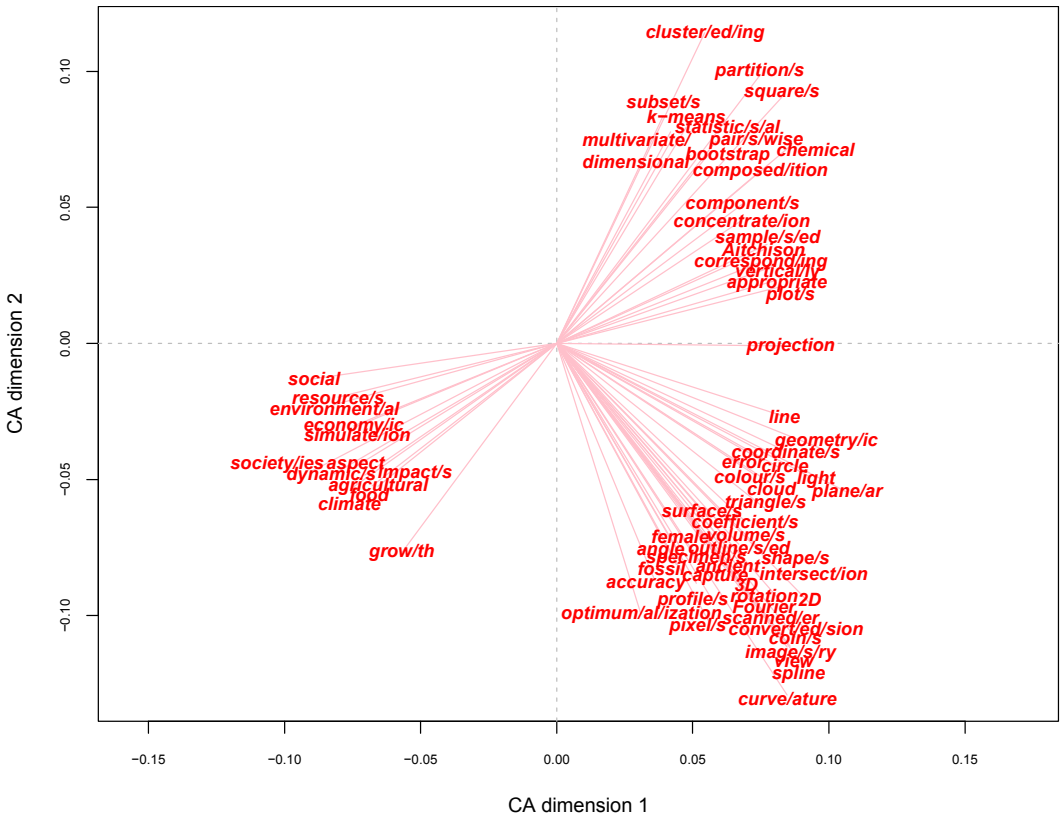

Figure 27.3 Words that make the major contributions to the distinction between chapters observed in the two-dimensional solution of Fig. 27.2.

Straight away I noticed in Fig. 27.2 that chapters 6 and 7, and to a certain extent 8 and 9 as well, were radically different from the remainder of the book. In Fig. 27.3 I further noticed that there were three important bunches of words, those pointing to bottom right corresponding to the separation of chapters 6 and 7, and then two bunches at bottom left and top right coinciding with the diagonally oriented dispersion of the chapters in Fig. 27.2 from chapter 21 (bottom left) to chapters 4 and 9 (top right). Important words associated with chapters 6 and 7 were, for example, *curvature*, *spline*, *images/ry*, *coins*, *convert/ed/sion*, *scanned/er*, *pixel/s*, *profiles*, *3D* and *2D*. In fact, just from the titles of these chapters I could confirm that they were exclusively dealing with the shape analysis of archaeological artefacts, and the coding of two-dimensional images and three-dimensional objects. Moving upwards on the right hand side of Fig. 27.3 came many terms associated with geometry such as *triangle/s*, *plane/ar*, *cloud*, *circle*, *coordinates*, *geometry/ic* and *line*. The more these words moved up on the right the more they were shared with chapters 9 and 8, in fact the word *projection* appeared as a word common to chapters 6 to 9. The only proper nouns that appeared in Fig. 27.3 were the names *Fourier* and *Aitchison*. Fourier at bottom right was clearly involved in the chapters on shape analysis, while it turned out that Aitchison had been mentioned 19 times in Chapter 8 and in no other chapter. This name referred to a statistician who was the founder of a school of compositional data analysis that was much appreciated in Girona, north of Barcelona, where the authors of this chapter worked. At top right I could recognize many statistical terms such as *cluster/ed/ing*, *partition/s*, *square/s*, *subset/s*, *k-means*, *statistics/al*, *multivariate/dimensional* and *bootstrap*, which implied that chapters 8 and 9, as well as the chapters at top right (e.g., chapters 4, 10 and 20), were more statistically technical, especially in the area of multivariate analysis. In contrast, the chapters towards bottom left, notably chapter 21 but also chapters 22, 23 and 24, were characterised by non-statistical words such as *growth*, *climate*, *agricultural*, *society*, *economy/ic*, *environment/al*, *resources* and *social*. It was no coincidence then that these chapters, along with the last chapter 25 were classified together as a separate section "Beyond Mathematics: Modeling Social Action in the Past". The position of the last chapter 25 seemed to tend back to the top right towards the statistical terminology – looking at this chapter in more detail I confirmed that it was indeed using more technical language than the others in this section.

The correspondence analysis gave me a good overview of the book's contents and main themes, but I was rather concerned about seeing only 15% of the inertia-type measure of

variance in the word counts, whereas the results pointed out that as much as 38% was non-random, and thus obviously worth exploring. The most important statistical term in the book appeared to be "cluster/ed/ing" (432 mentions – see Fig. 27.1) and I knew that clustering methods do analyse data in higher dimensions. Moreover, the term "k-means" was also important, prominent in chapters 20, 9 and 4, and appeared to be a type of clustering adapted to large sets of objects. So, in order take all the significant dimensions into account, I asked **`VisualR`** to "perform a k-means clustering of the six-dimensional solution of the words in the previous correspondence analysis" and, not knowing exactly how the results might be reported, I followed this with a vague "plot some standard results". The graphical results are shown in Fig. 27.4, and the verbal results were simply as follows:

k-means clustering of 751 words according to six-dimensional CA coordinates
Total variance = 1.104
Number of clusters explored from 2 to 30.
Explained variance between clusters is 14.2% (3 clusters) to 82.1% (30 clusters)
Estimated number of clusters can be 8 (58.5%) or 12 (68.5%).

The method had automatically formed different numbers of clusters and proposed some possible solutions. I decided to take the larger number of clusters recommended by the program, namely 12 clusters, since it was at that point that the increments in the variance explained started to tail off, as shown in Fig. 27.4.

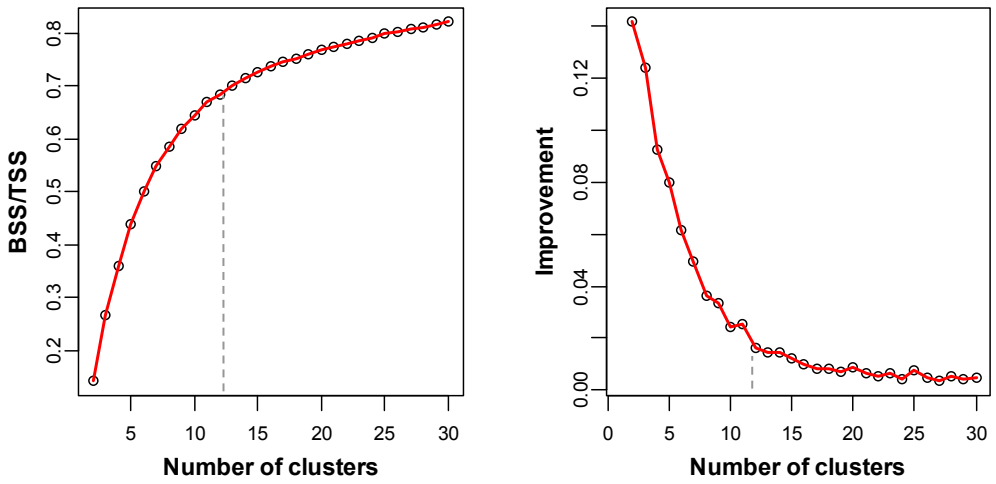

Figure 27.4 (Left plot) Increasing proportion of variance explained between word clusters as number of clusters varies from 2 to 30. (Right plot) Increments in the variance explained. 12 possible clusters of words are suggested.

Now I was faced with the interpretation of these word clusters, but first I wanted to see their relationship to the chapters, using the idea of a "heat map" that was commonly used in the *DataVision* utility on my atomic RAMchip to show me graphically the history of my expenses on different items. I tested **`VerbalR`**'s intelligence with this instruction: "compute percentages of words in each chapter for the previous 12-cluster solution; re-order the chapters and the clusters to be to be as similar as possible to their neighbours; plot a heat map of the table of standardized percentages". There was hardly a hesitation from the program to produce the plot in Fig. 27.5.

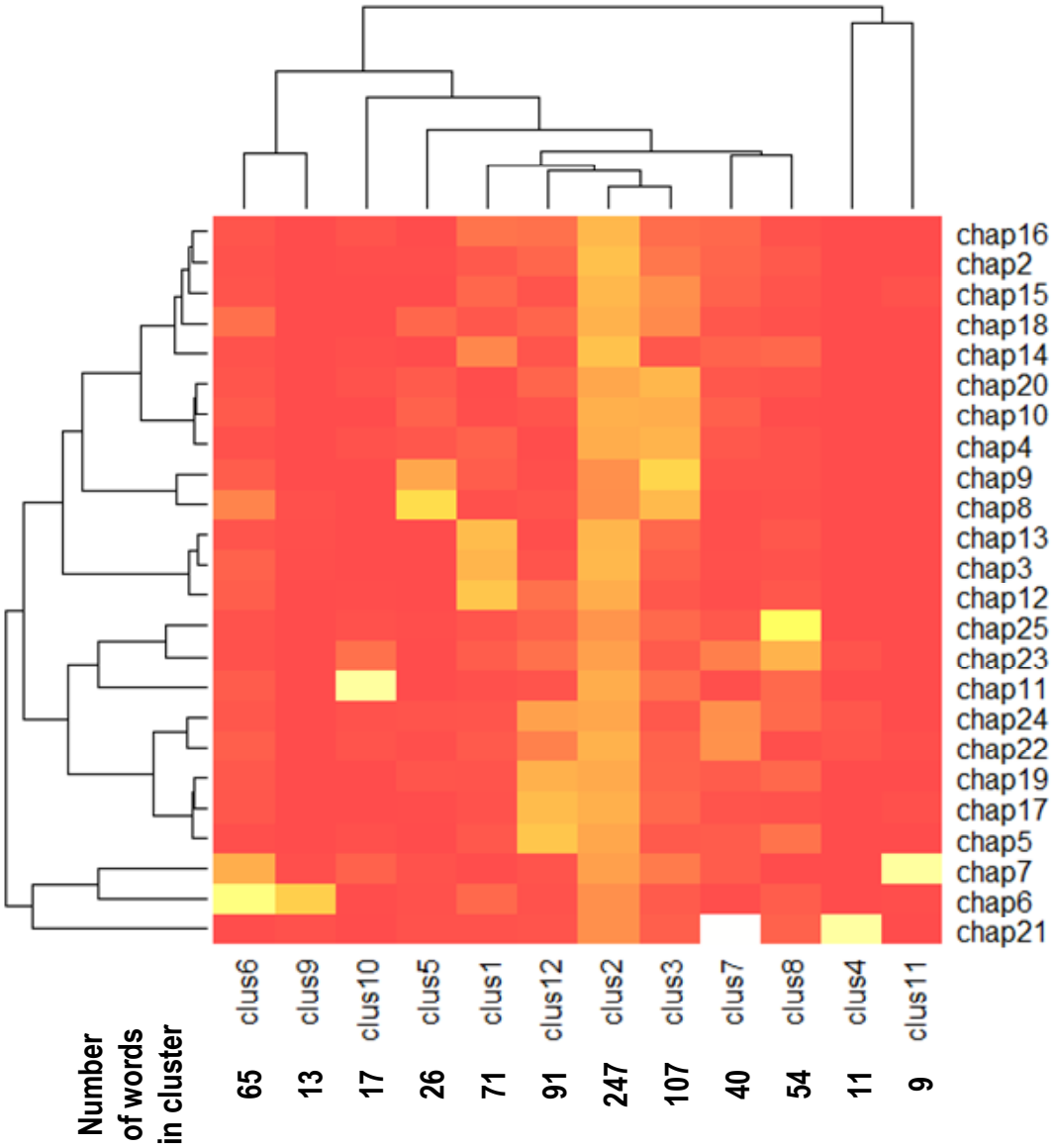

Figure 27.5 Heat map of the chapters and word clusters, re-ordered according to a hierarchical clustering of each set. Colours go from red (low) to yellow (medium) to white (high) in terms of standardized percentages of words.

I got much more than I expected, which testified to the advanced way **`VerbalR`** could respond to my requests. I could see more detail now why certain chapters were different from others and what characterized their vocabularies. In addition, the program had automatically chosen to re-order the chapters and word groups according to a hierarchical clustering of each, shown neatly in the graphical result of Figure 27.5. Several associations observed in the two-dimensional correspondence analysis were similarly apparent here, for example the exceptional nature of chapters 6 and 7. In addition, chapter 7 was unique in having words of cluster 11, a small cluster of the nine words with *outline/s/ed* and *fossil* being the most frequent, 60 and 30 occurrences respectively, words that were hardly mentioned in the other chapters. Chapter 21 had a unique vocabulary concentrated into clusters 4 and 7: cluster 4 had only 7 words, most importantly *farmers/ing*, *agriculture/al* and *climate*, whereas cluster 7 had 40 words, most importantly *population*, *simulate/ion*, *density/ies* and *transition*. The heat map showed clearly that the words in cluster 4 were exclusive to chapter 21 but the cluster 7 words were shared with chapters 22, 23 and 24, all in the same section of the book mentioned above. Chapters 3, 12 and 13, each from a different section of the book, were tightly clustered due to the words in cluster 1, such as *object/s*, *event/s*, *date/s/ed/ing*, *place/s*, *domain/s* and *chronology/ies/ical*, suggesting that these chapters were dealing with space-time aspects of archaeology. Cluster 2, the biggest one with 247 words, was clearly consisting of a vocabulary common to the whole book: the most frequent words in this cluster were those dominating the word list in Fig. 27.1, *use/d*, *archaeology/ist*, *model/ed/ing*, *data/base/set*, *analyse/analytical*, etc. Although I could comment on many other features of this heat map, I finally mention one that was impossible to see in the correspondence analysis results. In Fig. 27.2, chapter 11 was near the centre and seemed to have no special association with the words. In the heat map, however, that took into account more dimensions and formed several word clusters, chapter 11 appeared to be exclusively associated with word cluster 10, another small cluster consisting of 17 words, the most important in terms of frequency being *tree*, *lineage/s*, *ancestor*, *hypothesis/es/tical*, *evolve/evolution/ary* and *clade/istics*. The title of the chapter "Phylogenetic Systematics" confirmed its particular vocabulary and thus content.

Time precluded me from carrying on with my investigation of the text of "Mathematics and Archaeology" – my next job was to scan a related but older book "Mathematics in

the Archaeological and Historical Sciences", which had also been found in a highly deteriorated state in the same library. My reconstruction of the text, where it had been possible, of "Mathematics and Archaeology" would now go to another section of our research group. There specialists in archaeology and mathematical modelling would try to fill in the gaps and establish as complete a text as possible, which would then be added to the archives of a subject that would otherwise have been well and truly buried in the past. Nevertheless, I felt satisfied that I could apply many of the statistical techniques that the archaeological specialists themselves had applied to their artefacts of interest, be they coins, ceramics or fossils, to the words I found in the text of this old book. In the same way I had been able to discover patterns in the vocabulary in the various chapters in order to determine the book's structure, and to re-establish content just like the archaeologists of the past. The reconstructed text will now be stored in the new indestructible atomic databases in our research centre, and a copy of all my findings will remain in the atomic RAMchip on my wrist, be buried with me one day and never be lost to future generations.

My experience of rediscovering the statistical and archaeological texts gave me much food for thought. It seemed that although archaeology dealt with physical fragments of the past, which could be inspected, described and discussed amongst experts, and conclusions drawn, it lent more credence to their research to actually code the information in the objects in some type of quantitative or qualitative format and apply mathematically-inspired statistical tools to the resultant data. Sometimes the patterns in a corpus of objects emerged naturally from the data, using methods that "handled" the data minimally, allowing for objective inference, whereas in other cases researchers had a distinct hypothesis in mind, which could then be confronted with the numerical data.

This reminded me of my holidays in Samos Island, Greece, where I had stayed several times in Pythagorion, the home village of Pythagoras. Once I met a retired mathematical philosopher there who told me that the dictum of the Pythagoreans was "*All is number*" – all things in the universe had numerical attributes that uniquely described them. And another surprising thing was that this dictum had been passed along generations from ancient times to the 23rd century! Word-of-mouth, story-telling, the process of teacher-student communication, all these ways of transferring knowledge had become absolutely crucial after the technological war, when it was realized that

knowledge was “perdurant” and not “endurant”. The solid objects studied by archaeologists – ceramics, coins and fossils, amongst others – were enduring reminders of past societies. The written word itself had proved to be as ephemeral as the nebulous cloud that used to house it, and even the Rosetta stone and its various copies had outlived most of the writings of the succeeding millennia.

Apart from the verbal transfer of knowledge, the remnants of books remained as artefacts of mankind’s scientific progress. Understanding this progress required a particular technology for the “data” collection and – crucially, when the books had extensively deteriorated, as in the case of “Mathematics and Archaeology” – statistical tools to help reconstruct content and form. Thanks to the work of archaeolinguistic groups like ours and the application of these statistical methodologies, we were painstakingly piecing together the lost knowledge of the past.