Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited

Adalbert F. X. Wilhelm[1], Edward J. Wegman[2], and Jürgen Symanzik[2]

[1] Institut für Mathematik, Universität Augsburg, D–86135 Augsburg, Germany
[2] Center for Computational Statistics, George Mason University, Fairfax, VA 22030, USA

Summary

Interactive statistical graphics can be effectivly used to find natural groupings in observations. In this paper we want to demonstrate how clustering and classification can be done with three approaches based on highly interactive graphical environments: high–dimensional scatterplots as available in XGobi, parallel coordinate plots as available in ExplorN, and linked low–dimensional views as available in Manet. We will point out the strengths and the weaknesses of these techniques by comparing their behavior when applied to the Oronsay particle size data set.

Keywords: High Interaction Graphics, Grand Tour, Parallel Coordinate Plots, Linked Views, XGobi, ExplorN, Manet.

1 Motivation

Clustering and classification have always been highly related to the geometric structure of a data set. Visualization of that structure has been a main goal. Graphical devices have been widely used to support this visualization, but typically the use of graphics is restricted to the presentation of results and not to the exploration of clusters. In the last three decades many display types have been proposed for analyzing multivariate data and almost all creators of a new plot type claim that their method is powerful in clustering and classification. So, why do these graphical techniques play a minor role in the classification community?

A main reason might be that classification using these plots is hard to achieve in a static environment. The rise of high interaction graphics provides the user with the flexibility and power needed to detect and assign clusters.

The outline of the paper is as follows. In Section 2 we describe the basic notions of the high interactive techniques that we use and relate them to a dynamic graphics program: scatterplot rotation with projection pursuit and grand tour (XGobi), parallel coordinate plots combined with a d–dimensional grand tour (ExplorN), and linked low–dimensional views with user specified subsetting (Manet). The Oronsay particle size data set, used throughout this paper, will be introduced in Section 3. In Sections 4, 5, and 6, we describe our analyses in XGobi, ExplorN, and Manet, respectively. We conclude with a summary in Section 7. Analyses by high–interaction color graphics do not translate well into static grayscale pictures in a paper–based publication. For this reason, all of the images referred to in the following sections plus additional material to clarify intermediate steps are available in full color on our webpage

`http://www.galaxy.gmu.edu/papers/oronsay.html`

2 High Interaction Graphics

High interaction graphics have been developed within the last 15 years to enhance the graphic facilities developed for exploratory data analysis since the early 1960's. Eick & Wills (1995) define "an *Interactive Graphic View* as a pictorial representation of some form of data or information which the analyst can manipulate in real time". Main features of dynamic statistical graphics packages have been described in Cleveland & McGill (1988). Most of today's interactive statistical software packages contain some of these features, but unfortunately, they are often inconsistently and ineffectively implemented (Wilhelm, Unwin & Theus 1996).

2.1 High Dimensional Rotating Scatterplots

XGobi (Swayne, Cook & Buja 1998) is a dynamic statistical graphics program for high–dimensional data visualization, implemented in the X Window System™. Some of XGobi's main features (see Buja, Cook & Swayne (1996) for more details) that are particularly useful for classification and clustering, are the scatterplot tools, the grand tour, the projection–pursuit–guided grand tour, and linked brushing in multiple windows.

Scatterplots, and scatterplot matrices, may be the most obvious method to visually detect clusters or other structures. In case of a d–dimensional data set, we create $\begin{pmatrix} d \\ 2 \end{pmatrix}$ plots of variable i versus variable j, $i \neq j$, either one plot after another or all plots simultaneously arranged in a scatterplot matrix, and one can hopefully detect clusters in some of these plots. Examples of scatterplots (and linked brushing in scatterplots) can be found in any textbook on statistical graphics (e. g., Cleveland (1985) and Wegman & Carr (1993)).

The grand tour, introduced in Asimov (1985) and Buja & Asimov (1986), has been described as follows: "*The grand tour is a method for viewing multivariate statistical data via orthogonal projections onto a sequence of two–dimensional subspaces. The sequence of subspaces is chosen so that it is dense in the set of all two–dimensional subspaces.*" (Asimov 1985)

In addition to the grand tour, XGobi supports the projection–pursuit–guided tour (Cook, Buja, Cabrera & Hurley 1995), a combination of two complimentary methods into an interactive and dynamic framework. Projection pursuit (Kruskal 1969, Friedman & Tukey 1974, Huber 1985) results in a series of static plots of projections that are classified as "interesting" with respect to a particular projection pursuit index. The combination of grand tour and projection pursuit helps directing the grand tour towards "interesting" projections. This combination does not only show the "interesting" projections but it maintains the motion so the user has a feeling how successive "interesting" projections have been obtained. More details on projection pursuit indices available in XGobi can be found in Cook, Buja & Cabrera (1993) and Cook et al. (1995). Additional index functions that result in speed improvements of the calculations have been presented in Klinke & Cook (1997).

Finally, linked brushing in multiple XGobi windows allows to select a subset of points, i. e., a visual cluster, in one window, and to mark it with a different symbol and color. Plots in all associated XGobi windows will be updated automatically, using the same symbol and color for this marked set of points.

In addition to examples presented in the previously mentioned references on XGobi, it has been successfully used to display and cluster shopping–

X Window System is a trademark of MIT.

frequency data (Koschat & Swayne 1996). In combination with the Geographic Information System ArcView^{TM}, XGobi also has been used to detect structure and abnormalities in geographically referenced data sets such as satellite imagery, forest health monitoring, and precipitation data (Cook, Majure, Symanzik & Cressie 1996, Cook, Symanzik, Majure & Cressie 1997, Symanzik, Majure & Cook 1996).

XGobi is freely available from the following Web site:

`http://www.research.att.com/~andreas/xgobi/`

2.2 Parallel Coordinate Plots

The parallel coordinate display (Inselberg 1985, Wegman 1990) is a geometric device for displaying points in high–dimensional spaces, in particular, for dimensions greater than three. The idea is to sacrifice orthogonal axes by drawing the axes parallel to each other resulting in a planar diagram where each d–dimensional point $(x_1, \ldots, x_d)$ is uniquely represented by a broken line. The parallel coordinate representation enjoys some elegant duality properties with the usual Cartesian coordinates and allows interpretations of statistical data in a manner quite analogous to two–dimensional Cartesian scatterplots. This duality of lines in Cartesian plots and points in parallel coordinates extends to conic sections. This means that an ellipse in Cartesian coordinates maps into a hyperbola in parallel coordinates. Similar, rotations in Cartesian coordinates become translations in parallel coordinates.

The individual parallel coordinate axes represent one–dimensional projections of the data. We can isolate clusters by looking for separation between data points on any axis or between any pair of axes. Because of the connectedness of the multidimensional parallel coordinate diagram, it is usually easy to see whether or not this clustering propagates through other dimensions.

Since it is easy to see pairwise relationships for adjacent variables, but less easy for nonadjacent variables, a complete parallel coordinate investigation would require to run through all possible permutations. Instead of this, we recommend using a d–dimensional grand tour as introduced in Wegman (1991) and implemented in ExplorN (Carr, Wegman & Luo 1997). An important interactive procedure for finding clusters using parallel coordinate diagrams is brushing a subset of the data points with color. In addition, the level of saturation corresponds with the degree of overplotting creating a kind of parallel coordinate density plot (Wegman & Luo 1997*a*).

ExplorN runs on SGI^{TM} workstations and can be downloaded for free from the following ftp site:

`ftp://www.galaxy.gmu.edu/pub/software/`

ArcView is a trademark of Environmental Systems Research Institute, Inc.
SGI is a trademark of Silicon Graphics, Inc.

2.3 Linked Low–dimensional Views

To make the discovered structure as widely understandable as possible, it is desirable to use simple and easily interpretable views of the data. Extracting an easily understandable statement from some high–dimensional data projected on two–space is typically hard to achieve. It is therefore useful to give a description in terms of the original variables. Linking univariate views of individual quantities enhances this interpretation step. Our analysis by linked low-dimensional views will be done with the MANET (Unwin, Hawkins, Hofmann & Siegl 1996) software. This software provides a range of graphical tools specially designed for studying multivariate features with low-dimensional views. MANET grew out of a project to keep track of missing values in statistical graphics. It now provides many new interactive features including special graphs like spine plots and mosaic plots for categorical data. In MANET all displays are fully linked and instantaneously updated.

The standard use of linked views is to highlight clusters that are apparent in one dimension and to see these one–dimensional clusters in the light of other variables. By systematically subsetting the sample points, we can also detect two– and higher–dimensional clusters. Once we have detected a cluster a classification rule can be set up by taking the boundary values of the cluster. In MANET those values can easily be obtained by interrogating the plot symbols.

One–dimensional views show the one–dimensional clusters directly. Two–dimensional clusters become visible by highlighting a subset in one variable and conditioning another plot on this subset. For three– and higher–dimensional clusters we have to combine various subsets in different plots to one conditioning set and then to look at the remaining plots to check for clusters. The generation of such combined selections is not only possible in MANET but it is also very efficiently implemented as selection sequences (Theus, Hofmann & Wilhelm 1998). In a selection sequence we can easily jump from one branch of the hierarchic selection tree to another by just changing the relevant part in the sequence.

Linking bar charts and histograms is an effective way to analyze data sets with both discrete and continuous variables. It is also much easier to keep track of missing values in low–dimensional views than it is in high–dimensional graphics which typically restrict plotting to complete cases only.

MANET runs on Macintosh® computers and is freely available from the following Web site:

`http://www1.math.uni-augsburg.de/Manet/`

Macintosh is a registered trademark of Apple Computer, Inc.

3 The Oronsay Particle Size Data Set

The Oronsay particle size data, taken from Timmins (1981) and first analyzed in Olbricht (1982), are interesting multidimensional data for purposes of comparing three high interaction graphics tools. However, these data tell an interesting story that provides a useful background for our analysis. The story begins on the Oronsay island in the Inner Hebrides, North and West of the Scottish mainland. Here several important archeological sites are found, including two labeled *Caisteal nan Gillean I* and *Cnoc Coig* in the map displayed as Figure 1. This map is taken from Fieller, Gilbertson & Timmins (1987). These sites date from the mesolithic period or middle stone age, 10,000 to 8,000 BC, and contain middens, which are in effect stone–age land fills. Because these middens are near sites where prehistoric man lived, it is of interest to know how these locations were situated with respect to the mesolithic beaches and sand dunes. The prevailing scientific opinion is that there has been a seaward shift of the beach–dune interface based in part on the analysis of the Oronsay particle size data. This is indicated in Figure 1 by a dashed line that is intended to represent the mesolithic coastline vice the solid line that is intended to represent the modern coastline.

If the sand below the midden were beach sand and the sand from the upper layers were dune sand, this would indicate a seaward shift of the beach–dune interface. This would imply that the small mesolithic communities lived on the beaches. If on the contrary, the sand below and in the midden more strongly resembled dune sand, then the conclusions about the seaward shift of the beach–dune interface would be unwarranted and both the views of geologists about the seaward migration of the interface and of anthropologists about the living habits of mesolithic communities would have to be altered. In order to answer such questions samples of both modern dune and beach as well as samples of sands from the two midden sites were taken. The intent of these samples is to discover distinguishing characteristics between dune sands and beach sands and to classify midden sands into one of these two clusters.

There are a total of 226 sand samples. "Modern" samples have been taken at the two locations *Cnoc Coig* (119 samples) and *Caisteal nan Gillean* (30 samples). 77 archaeological samples of unknown (but to be determined) type from the midden locations are included as well. The 149 samples of known type can be further classified as beach at *Cnoc Coig* (90 samples), beach at *Caisteal nan Gillean* (20 samples), dune at *Cnoc Coig* (29 samples), and dune at *Caisteal nan Gillean* (10 samples). The full archaeological background is given in Fieller, Gilbertson & Olbricht (1984) and Fieller et al. (1987). Table 1 is an adaptation of Table 6 in Fieller, Gilbertson, Olbricht & Timmins (1983). It shows the group number, the number of observations per group, an environmental classification for the modern samples and the context clas-

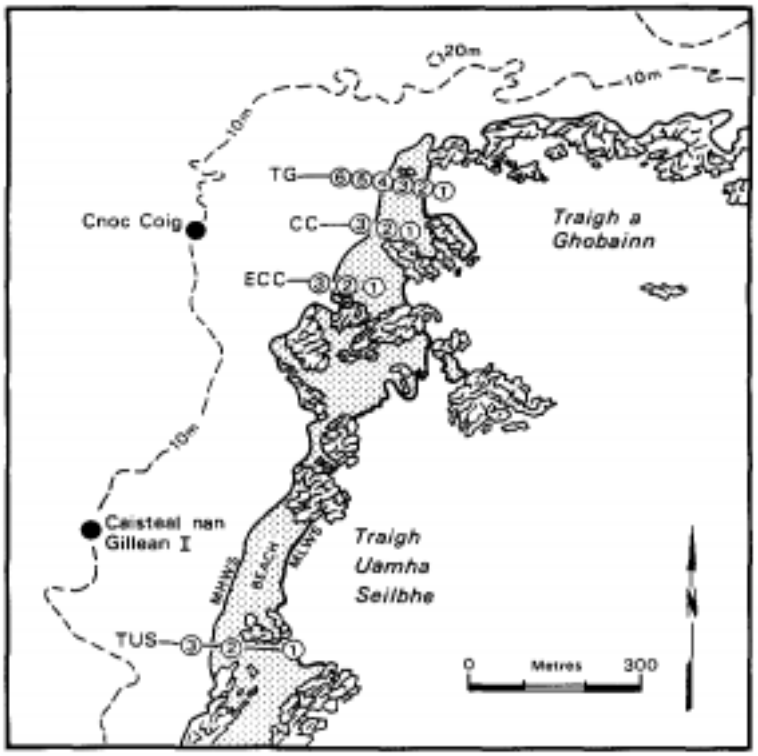


Figure 1: An excerpt from a map of the Oronsay island, Inner Hebrides. It shows the locations of the two archeological sites labeled *Caisteal nan Gillean I* and *Cnoc Coig* and the four transects TG, CC, ECC, and TUS where "modern" sand samples have been collected. Reprinted with permission from Fieller et al. (1987).

sification for the archaeological samples, and the abbreviation for the site code. Even though site EC–C3 (group 17) appears to be a dune location in Figure 1, it is pointed out in Fieller et al. (1987) that these samples come from open sand sheets and not dunes.

The data results from sieving 60g or 70g of sand through a stack of 11 sieves of sizes 0.063, 0.09, 0.125, 0.18, 0.25, 0.355, 0.5, 0.71, 1.0, 1.4, and 2.0 millimeters and weighing the sand left on each of the sieves and the sand that went through the smallest sieve. This results in weight measurements for 12 classes of particle sizes — the ten inner classes and the extreme classes [0, 0.063) mm and [2.0, ∞) mm. Selecting sieve sizes that are in approximate geometric progression with ratio $\sqrt{2}$ is considered standard in applications that are based on particle size data. Additional chemical and shape data are available for some of the samples (Fieller et al. 1983).

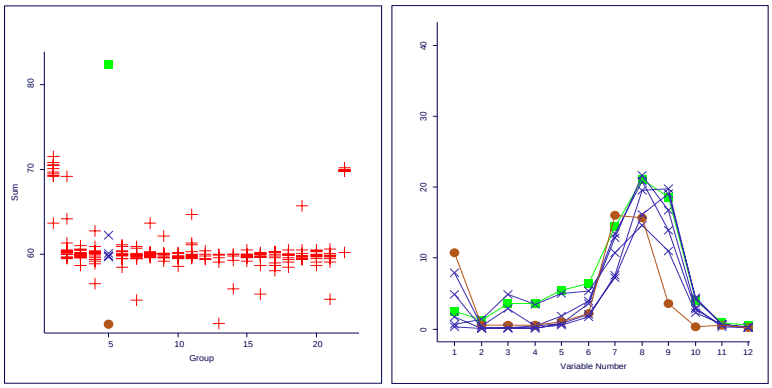
Group	# Obs	Classification	Site Code
1	10	Sands above CC Midden	CCJ6N (layers 1–10)
22	7	CC Shell Midden	CCJ6N (layers 11–17)
2	14	Sands below CC Midden	CCH17
3	18	Sands below CC Midden	CCJ6
4	13	CC Soil Pit	CCSP1
5	7	CNG Shell Midden	CNGIE (layers 1–5)
21	8	Sands below CNG Midden	CNGIE (layers 6–15)
6	15	CC Mid Beach	CC1
7	15	CC Upper Beach	CC2
8	14	CC Upper Dune	CC3
9	10	CC Lower Beach	TG1
10	10	CC Mid Beach	TG2
11	10	CC Upper Beach	TG3
12	5	CC Base of Dune	TG4A
13	5	CC Face of Dune	TG4B
14	5	CC Top of Dune	TG4C
15	10	CC Mid Beach	EC–C1
16	10	CC Upper Beach	EC–C2
17	10	CC Upper Beach	EC–C3
18	10	CNG Lower Beach	TUS1
19	10	CNG Upper Beach	TUS2
20	10	CNG Dune	TUS3

Table 1: Group number, number of observations per group, classification, and site code. CC stands for *Cnoc Coig* and CNG stands for *Caisteal nan Gillean*. Site codes for the "modern" sand samples directly relate to locations shown in Figure 1.

The classical statistical analysis of particle size data such as the Oronsay data tries to identify particle size distributions. While originally these distributions have been typified by their sample moments, more recent studies tried to fit parameter estimates. In the latter case, a classification of the (unknown) samples will be based on the estimated parameters for these distributions. In Fieller et al. (1984), log–hyperbolic and log–skew Laplace models have been discussed. See Fieller, Flenley & Olbricht (1992) for a summary on statistical methods for particle size data and a comparison of several parametric models that have been fit to the Oronsay data.

While Fieller et al. (1992) distinguish only between beach and dune sand for

their classification, Flenley & Olbricht (1993) consider a four–group classification that distinguishes among beach at *Cnoc Coig*, beach at *Caisteal nan Gillean*, dune at *Cnoc Coig*, and dune at *Caisteal nan Gillean*. In addition, multivariate classification techniques such as principal component analysis and projection pursuit methods have been applied to the Oronsay data in Flenley & Olbricht (1993).



(a) (b)

Figure 2: (a) Scatterplot of summed up weights (vertical) versus group (horizontal). While most weights scatter around 60g, values around 70g have been observed for most samples from groups 1 and 22. Group 5 has the smallest sum (51.8g) marked with a "•" but also the largest sum (82.4g) marked with a "■". (b) A parallel coordinate plot of group 5 reveals that the "•" sample has an unusual small value for variable 9, i. e., particle size [0.125, 0.18) mm, but it also has higher values for variables 1 and 7. The "■" sample has a shape that matches the shape of the remaining samples of group 5.

For our graphical analysis of the Oronsay data within this paper, we used the original weight measurements for the 12 classes but did not consider any of the chemical and shape data. However, we also did not exclude the extreme classes a priori, as suggested in some of the previous analyses of this data. One inconsistency within the data should be noted. While in Timmins (1981) and other publications sample weights of 60g or 70g have been reported, the summation over the 12 classes results in approximately 60g or 70g for most of the samples, but extreme values of 51.8g and 82.4g have been found as well (Figure 2(a)). We contacted the authors of one of the previous papers. They were aware of this discrepancy but could not provide an explanation. Within this paper, we did not do any adjustment, standardization, or transformation but kept the original data to check for the robustness of our graphical methods. However, a graphical representation might provide some explanation for some of the unusual observations. In Figures 2(a) and 2(b) we highlighted the observation with the smallest sum (51.8g) with a "•" and the observation with the highest sum (82.4g) with a "■", both from group 5 (CNG Shell Midden). It becomes immediately visible

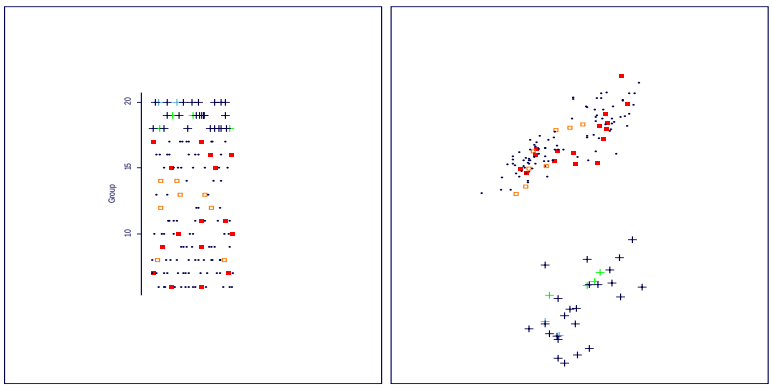
in Figure 2(b) that the smallest observation has a considerably smaller value for variable 9, i. e., particle sizes of [0.125, 0.18) mm, when compared with the remaining 6 samples from group 5. Actually, this sample has only 3.6g for variable 9 while all other samples from group 5 have weights in the range from 11.0g to 19.8g for variable 9. So, we conjecture that this is just a recording error and the correct measurement should be 13.6g which would raise the sum for this sample to 61.8g, a plausible value. However, this cannot be answered definitively since this sample shows quite a different pattern than the other samples from this group. This sample has higher values than the other 6 samples from group 5 for variables 1 and 7, i. e., for particle sizes >2.0mm and [0.25, 0.355) mm. On the other hand, the pattern associated with the highest observed sum looks very similar to the remaining samples from group 5. We conjecture that this is just a human error and in fact 82.4g of sands have been sieved instead of 60g.

4 Analysis in XGobi

Our graphical analysis in XGobi can be split into two parts: (i) a visual clustering/classification entirely based on the grand tour, and (ii) a reevaluation of some of the results of Fieller et al. (1984) based on the projection–pursuit–guided grand tour.

For part (i), we only used the 149 samples from the known sampling locations, i. e., groups 6 to 20 (see Table 1). Within each of the 15 groups we randomly selected 2 samples and brushed them according to their classification as beach at *Cnoc Coig*, beach at *Caisteal nan Gillean*, dune at *Cnoc Coig*, and dune at *Caisteal nan Gillean*. This sampling/brushing was done during the preparation of the XGobi data files by assigning different color information to the first two samples of each group. We make use of these training samples to classify visible clusters in the different XGobi projections. Given that clusters can be distinguished and that only one type of training samples appears in a particular cluster, we classify the entire cluster with respect to this type. Therefore, in the ideal case it does not matter which samples have been initially selected. We now started the grand tour and were looking for visible clusters.

Almost immediately, we found a projection that clearly separated two clusters (Figure 3(b)). We brushed one of these clusters with a "+" symbol. All 6 marked training samples from *Caisteal nan Gillean* fall into this cluster while all training samples from *Cnoc Coig* fall into the other cluster. We use this projection to distinguish between the two sites. Actually, we notice in the dotplot of "Group" (Figure 3(a)) that the brushed points fall all into groups 18 to 20 — the *Caisteal nan Gillean* locations. The clear separation between the two sites described in Flenley & Olbricht (1993) has been detected. From



(a) (b)

Figure 3: (a) A dotplot of group for the 149 samples from known sampling locations. Within each of the groups 6 to 20 two training samples have been selected. The "+" symbols for all observations in groups 18 to 20 have been obtained through linked brushing in (b) a projection of the grand tour. Here, two clusters are clearly distinguishable that separate *Caisteal nan Gillean* samples (the"+" samples) from *Cnoc Coig* samples.

here on, we did a hierarchical analysis, first considering the *Cnoc Coig* samples (Figures 4(a) to 4(d)) and thereafter the *Caisteal nan Gillean* samples (Figures 5(a) to 5(e)). Within this hierarchical analysis, we erased the other subset (but could easily restore the points later on) such that the grand tour is only based on the visible points. This is different from masking points. There, points that are temporarily invisible in a projection still contribute to the internal calculations of this projection. Hence, in Figures 4(a) to 4(d) only the 119 *Cnoc Coig* samples are shown while in Figures 5(a) to 5(e) the 30 *Caisteal nan Gillean* samples are shown.

Not every visible cluster in one projection necessarily results in a correct classification or subsetting of the samples. Figure 4(b) shows a very narrow cluster within the *Cnoc Coig* samples (groups 6 to 17) that has been brushed with a "+" symbol. Figure 4(a) shows that the brushed points relate to 9 (out of 10) mid beach (group 15) and 4 (out of 10) upper beach (group 17) samples from location EC–C. However, as Figure 4(d) indicates, the "+" cluster is not really homogeneous but separates into two subclusters in another projection. Probably, from this projection it should be easy to indicate which of the 13 points are mid beach and which are upper beach samples. It should be possible as well to guess which other points are most likely to fall into these sand classes. However, this static projection does not provide a unique answer to this since the two "+" subclusters do not clearly separate from the main cluster of points. Only the continuation of the grand tour (or the knowledge of its previous stages) might provide the correct answer.

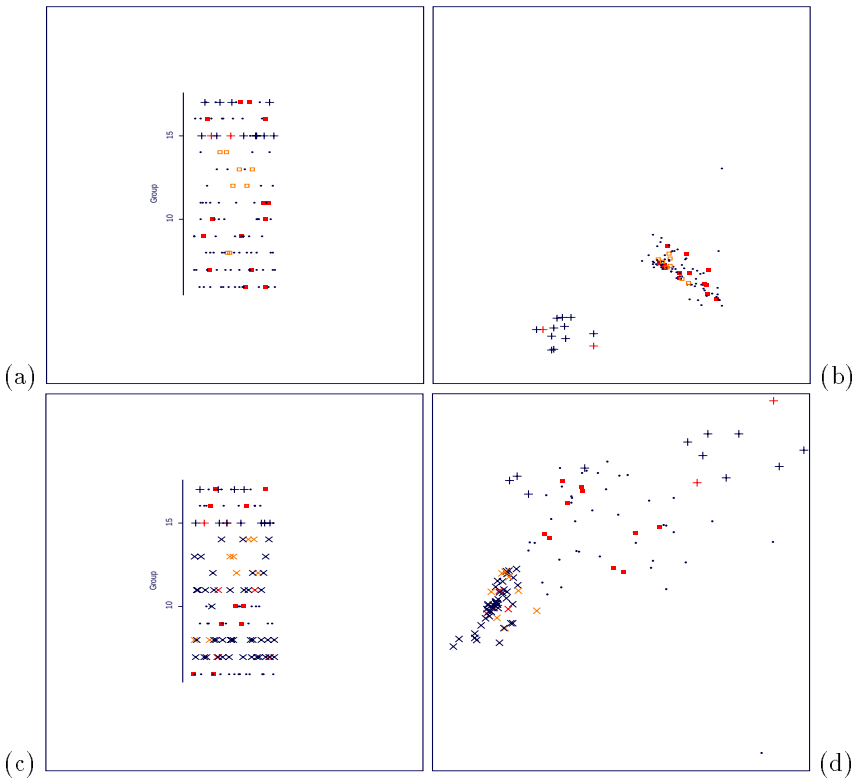


Figure 4: (a) When considering only the *Cnoc Coig* samples (groups 6 to 17), a dotplot reveals that the cluster brushed with a "+" symbol in (b) a projection of the grand tour contains mid beach (group 15) and upper beach (group 17) samples. (c) A dotplot of the final clustering of the *Cnoc Coig* samples shows the correct classification of dune samples (groups 8, 12 to 14) and the misclassification of two entire groups of upper beach samples (groups 7 and 11) and 1 mid beach sample (from group 10) based on (d) a projection of the grand tour where a homogeneous group of points has been marked with an "×" symbol, assuming these are all dune samples. This projection also shows how the "+" cluster brushed in (b) separates into two subclusters.

In addition, Figure 4(d) shows a cluster (brushed with an "×" symbol) that behaved homogeneously over a longer time period and in this projection moved into a different direction than the nearby (but otherwise marked) points. This cluster contains all 8 training samples that initially have been marked as dune at *Cnoc Coig* but it contains also 4 training samples that initially have been marked as beach at *Cnoc Coig* (out of our set of 24 training samples for the *Cnoc Coig* location). None of the training samples marked as dune at *Cnoc Coig* is located anywhere else. Therefore, we used a majority vote and decided that all the points in this cluster have to be considered as

possible samples from dune at *Cnoc Coig*. The result of this classification can be seen in Figure 4(c). All dune samples (groups 8, 12 to 14) have been correctly classified as dune. However, all upper beach samples from 2 locations (groups 7 and 11) and 1 mid beach sample (group 10) have also been classified as dune, resulting in a total of 26 misclassifications. We further tried to find any structure or subcluster in the group of samples that we just classified as dune but did not succeed. This matches the results of Flenley & Olbricht (1993), page 484, who state that "*amongst the samples that were wrongly classified as dune, we find in both cases entire "subclusters" of "upper beach" samples*". However, as Figures 4(b) and 4(d) suggest, it is even possible to find further subclusters for the beach sand. This is demonstrated through the use of parallel coordinates in ExplorN in the next section.

In a second step, we looked at the *Caisteal nan Gillean* samples (groups 18 to 20). Figure 5(b) shows a cluster that has been brushed with a "•" symbol. Figure 5(a) shows that the brushed points relate to lower beach samples (group 18). Figure 5(d) shows another cluster that has been brushed with a "∘" symbol. Figure 5(c) shows that the brushed points relate to 9 (out of 10) dune samples (group 20). The remaining dune sample has been brushed with a "■" symbol and the projection in Figure 5(d) shows that this sample is located close to the points that previously have been classified as lower beach (the "•" symbols).

Figure 5(e) finally shows a projection where all three clusters are separated and the sample brushed with the "■" symbol falls among the other dune samples. However, would we consider this projection as a good one if we know about the other projections we obtained during our interactive work with XGobi? — The answer is no. Trying to do visual clustering within XGobi is an alternating sequence of brushing, looking at additional projections from the grand tour, brushing, and so on. We can only be sure that a cluster visible in one projection really is a cluster if its points remain close to each other in a series of projections and these points move similarly when the grand tour is activated.

Unfortunately, Figure 5(e) is only a snapshot. The "+" and "∘" symbols moved in the same direction and the "■" symbol moved outside the area of "∘" symbols almost immediately. According to the majority of other projections we have seen, we would have misclassified this sample as lower beach. Thus, we misclassified a total of 27 (out of 149) samples — exactly the same number of samples that have been misclassified in the best case reported in Flenley & Olbricht (1993) (page 484) and better than the classifications based on parametric models reported in Fieller et al. (1992) (page 140).

However, the real strength of the visual clustering in XGobi is the fact that we are able to see if samples are mixtures of two sand types even though we assume that they originate from one sampling location only. In the case of the

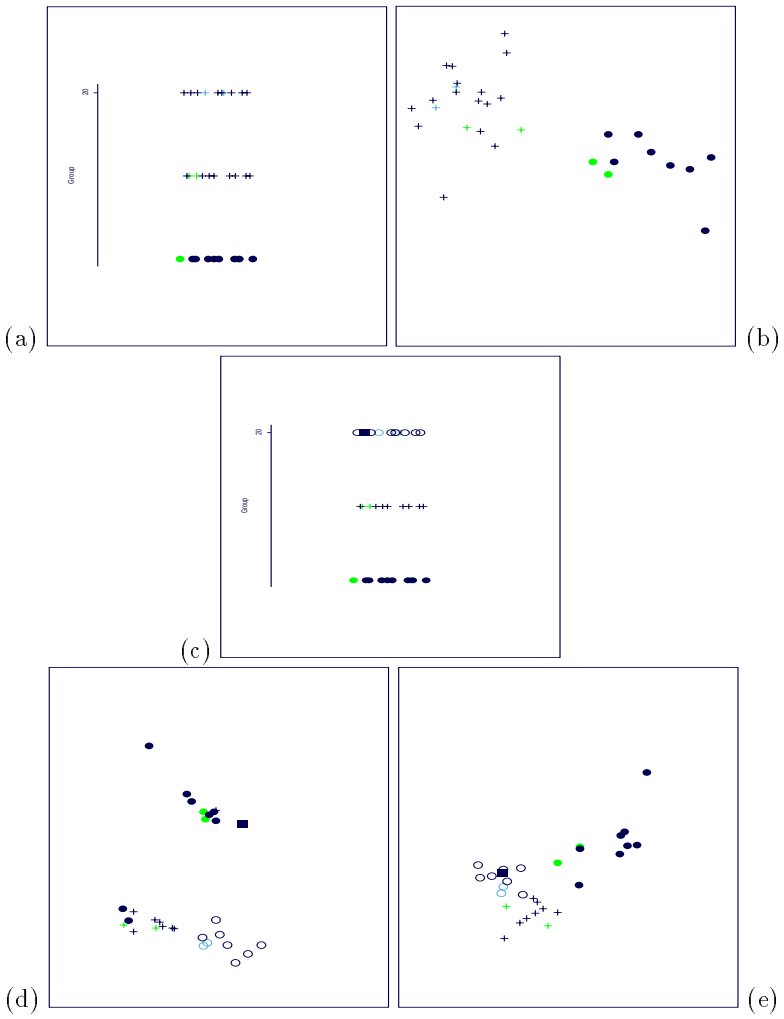
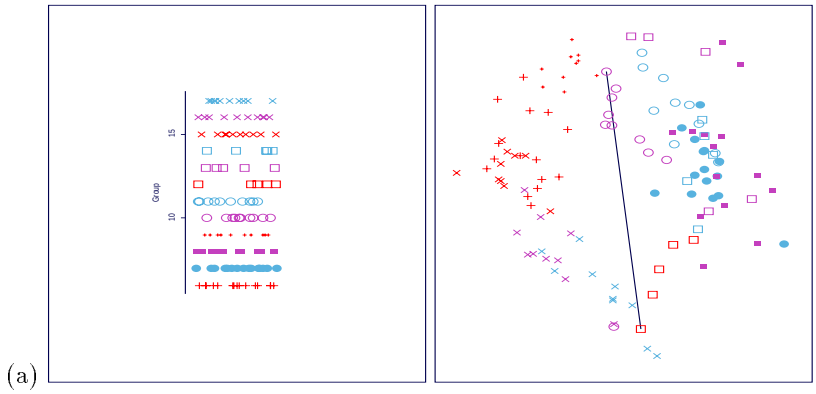


Figure 5: (a) When considering only the *Caisteal nan Gillean* samples (groups 18 to 20), a dotplot reveals that the cluster brushed with a "•" symbol in (b) a projection of the grand tour contains all lower beach samples (group 18). (c) A dotplot showing how the dune samples (group 20) brushed with a "o" symbol can be separated from upper beach samples (group 19) based on (d) another projection of the grand tour. The "■" sample is not identified as a dune sample even though it originates from group 20. (e) Even though this projection of the grand tour shows an instance of a separation among the three groups, it cannot be considered as useful for clustering in an interactive environment since the "+" and "o" symbols moved in the same direction and the "■" moved outside the area of "o" symbols almost immediately.

30 *Caisteal nan Gillean* samples it is reasonable to assume that the two "•" samples from Figure 5(d) that are located on top of the "+" cluster contain a mixture of lower and upper beach sand. It would have been interesting to look at a map and see how close these two samples from the lower beach are located to the upper beach sampling area but unfortunately this spatial information is no longer available. Similarly, the "■" sample seems to be a mixture of lower beach and dune sand. There are many possible explanations what has happened with these samples. Sand may have been blown away by wind, washed off by water, or carried to another location by animals or humans, thus resulting in the observed mixture of sand types.



(a) (b)

Figure 6: (a) When considering only groups 6 to 17 for the reevaluation of Figure 4c in Fieller et al. (1984) a dotplot shows the symbols used to mark the 12 known groups at the *Cnoc Coig* site. (b) A projection showing a local optimum based on the projection–pursuit–guided grand tour and a manually drawn dividing line shows a separation between beach samples (left of the line) and dune samples (right of the line). However, upper beach locations from the CC transect (group 7) and from the TG transect (group 11) also fall right of the line.

In part (ii) of our analysis with XGobi we wanted to graphically reevaluate some of the results of Fieller et al. (1984). First we looked at the three transects CC (groups 6 to 8), TG (groups 9 to 14), and EC–C (groups 15 to 17) from Figure 4c in Fieller et al. (1984). We tried to find a distinction between modern beach and dune sands. In addition to the classification into the three transects used in Fieller et al. (1984), we marked each of the 12 known groups with a different symbol (Figure 6(a)). Then we made use of XGobi's projection–pursuit–guided grand tour using the "Holes" projection pursuit index as optimization criteria. This option is responsive to projections containing very few data points in the center — exactly what we are looking for when trying to locate clusters. One should note that we could not use all of the 12 classes of the sieve sizes — the projection pursuit algorithm detected singularities within the data. We randomly deleted one or multiple of these variables from our data matrix (just by clicking on the corresponding

variable circle in XGobi) but in general obtained similar results no matter what choices we made. Figure 6(b) shows one of many very similar projections where the projection–pursuit–guided grand tour stopped with a local optimum. We manually added a dividing line. All dune sampling locations (the "■" symbol for group 8, and the three different "□" symbols for groups 12 to 14) are located right of this line. In addition, all upper beach locations on the CC transect (the "•" symbol for group 7) and most of the upper beach locations for the TG transect (the light "∘" symbols for group 11) fall right of the dividing line. Moreover, the mid beach locations for the TG transect (the dark "∘" symbols for group 10) fall onto this line.

Recall that exactly the observations from groups 7, 10, and 11 were the samples that have been misclassified in part (i) of our XGobi analysis. We assume that these are also the samples that have been misclassified in all other analyses throughout the literature. Since we could not find any further difference among samples visually classified as dune sand, one should consider to relabel these samples as dune sand even though they originate from beach locations before trying to classify the historic samples. Figure 2 in Fieller et al. (1984) which shows the spatial closeness of the sampling locations and Figure 7.2 and Table 7.1 in Andrews, Gilbertson & Kent (1987) which provide wind exposure rates at different locations might be a reasonable explanation that the sand found in the upper beach locations of transects CC and TG in reality is dune sand blown away by wind. It appears that these two transects fall into a region (nearby transect 22 in Andrews et al. (1987)) where the wind exposure has been rated as "Exposed", a fact also noted in Fieller et al. (1984). The EC–C transect almost coincides with transect 20 in Andrews et al. (1987) which has neither been rated as "Exposed" nor "Sheltered". Thus, there is no strong evidence to believe that dune sand has been blown to the beach. Actually, it is not even clear where the dune is located for the EC–C transect since the last samples on this transect originate from open sand sheets and not from dunes. The rating in Andrews et al. (1987) also provides an additional explanation why there was no dune sand found at the upper beach at the *Caisteal nan Gillean* site. Not surprisingly, this segment (near transect 16 in Andrews et al. (1987)) of the beach has a rating of "Sheltered", implying that only little or no sand has been blown from the dune to the beach.

Finally, we looked at the samples from *Caisteal nan Gillean* presented in Figure 4d in Fieller et al. (1984). This subset includes modern samples from lower beach (the small "+" symbol for group 18), upper beach (the large "+" symbol for group 19), dune (the "×" symbol for group 20), unknown sands from the shell midden (the "•" symbol for group 5), and unknown sands from below midden (the "□" symbol for group 21) as displayed in Figure 7(a). We encountered the same problems with singularities when running the projection–pursuit–guided grand tour in XGobi. Again, we ran-

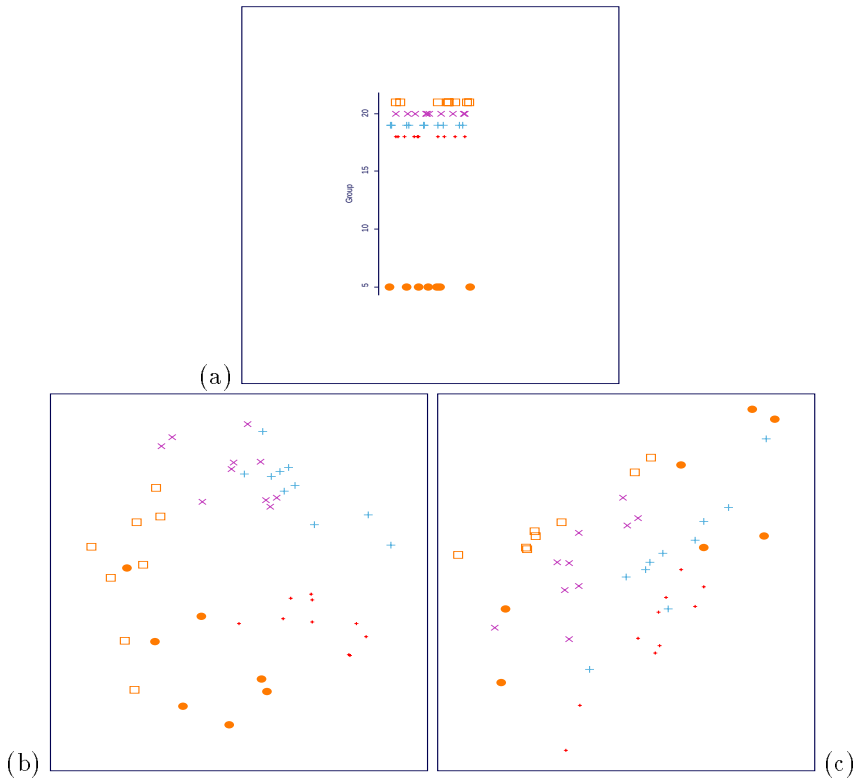


Figure 7: (a) A dotplot shows the symbols used to mark the 5 groups from the *Caisteal nan Gillean* site used for the reevaluation of Figure 4d in Fieller et al. (1984). Projection (b) shows a circular arrangement and projection (c) shows a linear arrangement, each obtained as a local optimum based on the projection–pursuit–guided grand tour. These and many similar projections show more differences than similarities between archaeological samples (groups 5 and 21) and modern samples (groups 18 to 20).

domly eliminated some of the variables but got similar results in each of the cases. Figure 7(b) shows a circular arrangement of the groups lower beach, upper beach, dune, sands from below midden, and shell midden. Figure 7(c) shows a similar linear arrangement starting with the lower beach samples in the lower right corner and ending with the sands from below midden in the upper left corner and the shell midden samples spread among several groups.

In addition, Figure 7(b) allows to draw a separating line between modern and historic samples. Based on our analysis in XGobi, we cannot agree with the conclusion from Fieller et al. (1984), page 650, that "the sands within the CNG I shell midden resemble more closely deposits found on modern mid–beach, rather than upper beach". Interpreting our circular and linear

arrangements and the (clear) separability in Figures 7(b) and 7(c), we might conclude that the historic samples have little in common with modern dune or beach sand.

Otherwise, interpreting these arrangements as a spatial arrangement (assuming that closeness in the projection resembles geographic closeness, thus, assuming that lower beach and upper beach samples are closer to each other than lower beach and dune samples, for example), one might conclude that the sands from below midden samples (group 21) are located closest to the modern dune — perhaps on the other side of the dune, i. e., further land inwards? The shell midden samples (group 5) might be located on the upper beach or dune according to Figure 7(c) but Figure 7(b) does not support such a classification. It might be reasonable to consider the chemical information that is available for these modern and historical samples before a final conclusion is drawn.

We did not reevaluate Figure 4b from Fieller et al. (1984) using XGobi. This is done in the next section using ExplorN. However, we did a readjusted reevaluation of Figure 4a from Fieller et al. (1984) considering all 22 groups, i. e., archaeological and modern samples from *Cnoc Coig* and *Caisteal nan Gillean*, at the same time. According to our previous findings, we reclassified samples from groups 7 and 11 as "dune–like" samples. Figure 8(a) shows a local optimum once again obtained through the projection–pursuit–guided grand tour in XGobi. This projection separates between sites (the big "+" and "×" symbols are *Cnoc Coig* samples, the small "+" and "×" symbols are *Caisteal nan Gillean* samples) and sands within sites (the "+" symbols are beach samples and the "×" symbols are "dune–like" samples). The small "." symbols represent the archaeological samples. Separation lines have been added manually.

A dotplot (Figure 8(b)) shows the symbols assigned to the 149 modern and 77 archaeological samples. In Figure 8(c), we see that the archaeological *Caisteal nan Gillean* samples (the "■" and the "□" symbols for groups 5 and 21, respectively) fall relatively close to the modern *Caisteal nan Gillean* samples. Historical samples appear on the modern beach and dune side. No consistent pattern can be detected in this projection.

The archaeological *Cnoc Coig* samples are located on the side of the modern *Cnoc Coig* dune samples. Sands below CC Midden (the small "•" symbol for groups 2 and 3) have most in common with modern *Cnoc Coig* dunes, Sands above CC Midden (the large "•" symbol for group 1) have still something in common with modern *Cnoc Coig* dunes. The CC Shell Midden (the large "∘" symbol for group 22) and the CC Soil Pit (the small "∘" symbol for group 4) samples have some overlap with the other archaeological *Cnoc Coig* samples. They have very little in common with modern *Cnoc Coig* dunes and they are clearly distinguishable from modern *Cnoc Coig* beach. Archaeological *Cnoc*

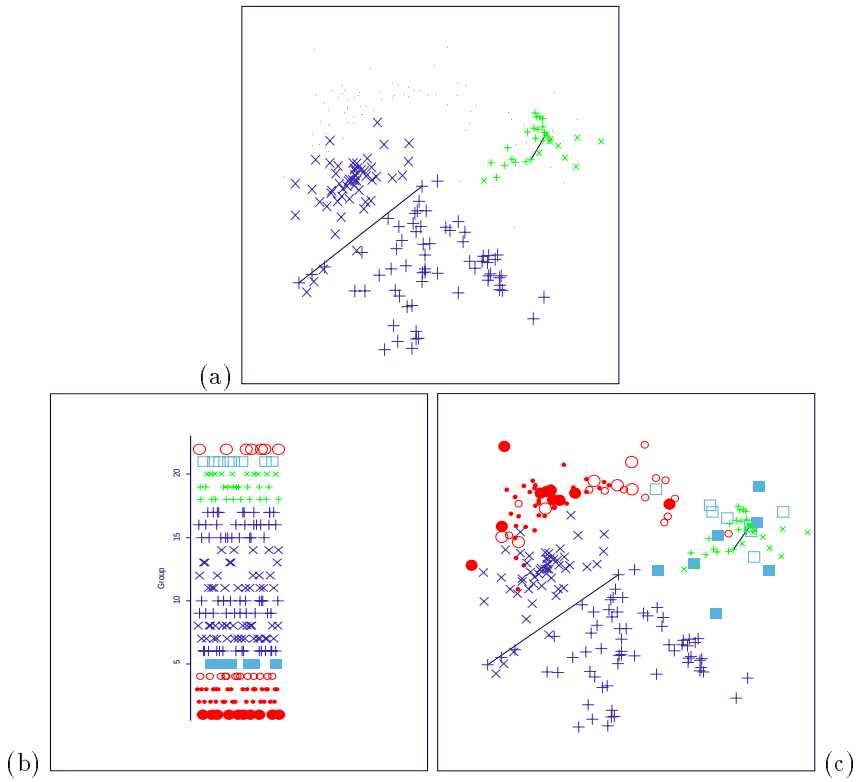


Figure 8: (a) A projection that separates between sites (the big "+" and "×" symbols are *Cnoc Coig* samples, the small "+" and "×" symbols are *Caisteal nan Gillean* samples) and sands within sites (the "+" symbols are beach samples and the "×" symbols are "dune–like" samples). The small "." symbols represent the archaeological samples. Separation lines have been added manually. (b) A dotplot shows the symbols used for all 226 samples. (c) When adding the symbols used in (b) to the projection in (a), we see that archaeological *Caisteal nan Gillean* samples fall close to modern *Caisteal nan Gillean* samples (beach and dune). Archaeological *Cnoc Coig* samples are clearly distinguishable from modern *Cnoc Coig* beach. Sands above CC Midden (group 1) and Sands below CC Midden (groups 2 and 3) are close to modern *Cnoc Coig* dunes. CC Shell Midden (group 22) and CC Soil Pit (group 4) have some overlap with the other archaeological *Cnoc Coig* samples but they have very little in common with modern *Cnoc Coig* dunes.

Coig and archaeological *Caisteal nan Gillean* samples also separate well.

5 Analysis in ExplorN

ExplorN is a multidimensional graphical analysis tool written for Silicon Graphics platforms using either the GL or the OpenGL tools. The code combines scatterplot matrices, parallel coordinate displays, icon–enhanced three–dimensional stereoscopic plots, d–dimensional grand tours and partial grand tours (i. e., tours based on a subset of the variables with the remaining variables being held fixed), and saturation brushing all in a high interaction graphics package. Parallel coordinate methods as an exploratory analysis tool are described in Wegman (1990), d–dimensional grand tours are described in Wegman (1991) and also in Wegman & Carr (1993), while saturation brushing is described in Wegman & Luo (1997*a*, 1997*b*). The ExplorN software is intended to demonstrate principles as opposed to be an operational tool so that some refinements normally found in operational software are not there. These include history tracking, easy point identification, identification of mixture weights in the grand tour, relabeling of axes during and after a grand tour as well as simultaneous multiple window views. On the other hand, the software is capable of handling much larger data sets than commonly used with other packages such as S–Plus (Becker, Chambers & Wilks 1988). We have used the software with more than 300,000 observations in 12–dimensional space. The saturation brushing techniques are most applicable to medium to huge data sets (as defined in Wegman 1995). The Oronsay particle size data is on the order of 200 observations and qualifies as a tiny data set. Although ExplorN is capable of developing an analysis based on conventional scatterplots, for purposes of this paper, we have relied entirely on parallel coordinate displays and partial grand tours except for an occasional presentation in scatterplot matrix mode for readers less familiar with the interpretation of parallel coordinate displays.

ExplorN has been developed on high resolution graphics–oriented workstations. In contrast with general purpose personal computers and workstations, the use of color graphic tools allows for a very subtle high resolution analysis. By using a black background and subtle colorings, we have been able to code data with single pixel and single–pixel–width lines. While this allows a rather delicate high–interaction analysis on the screen, it does not translate well into static pictures in a paper–based publication. For this reason, the reader should visit our webpage on which all of the images referred to in the following analysis are available in full color 1280 x 1024 screen shots.

In ExplorN, the parallel axes are drawn as horizontal lines in contrast with XGobi where they are drawn as vertical lines. While it may be argued that this difference is immaterial from a mathematical point of view, the wider aspect ratio in the horizontal mode coupled with a more usual sense of plotting data along an abscissa rather than along the ordinate tends to allow for an easier human interpretation. In this version of a parallel coordinate

display, a multidimensional point is plotted by plotting its components on the appropriate axis and joining the components by a broken line segments. Detailed interpretations are given in Wegman (1990) and the reader is referred to that work. It is sufficient to note here that any gap in the data plot along a horizontal axis corresponds to a separation between clusters. If the separation forms on several of the parallel axes, then a cluster is evident in more than one dimension.

By using a full or a partial grand tour, one can find orientations where one or more clusters are evident. The general strategy for detecting clusters is as follows. Begin with a static plot of the data in parallel coordinates. If there are any gaps along a horizontal axis (which incidentally does not need to coincide with the coordinate axes), then color the individual clusters with distinct colors. Once all clusters are identified in the original coordinate system, initiate a full or partial grand tour until an orientation of the axes is found in which another gap in one of the horizontal axes is found. Again color the individual subclusters with distinct colors. This procedure is repeated until no further subclusters can be identified. We shall refer to this as the brush–tour strategy. Indeed, this is a matter of judgement, since the procedure can be repeated until practically every data point can be individually colored. The crucial issue, which really depends on the dynamic graphic, is to see that clusters identified in this manner track coherently with the grand tour animation. That is, data points of the same color, red, for example, stay together as the grand tour rotation proceeds. If they do not, then there are likely substructures that can be identified through further grand tour exploration.

Slopes of parallel coordinate line segments can also be used to distinguish clusters. That is, if a group of line segments slopes, say, at 45 degrees to the horizontal and another group slopes at, say, at 135 degrees to the horizontal, then even though the lines fully overlap in both adjacent parallel coordinate axes and there is no horizontal gap, these sets of lines represent two distinct clusters of points. Fortunately, when such indication of clustering exists, the grand tour will also find an orientation of axes in which there is a horizontal gap. Thus the general strategy is to alternate color brushing of newly discovered clusters with partial grand tour rotations until no further clusters can be easily identified. This is the process we followed for purposes of the analysis of the Oronsay data.

Figure 9 is the original parallel coordinate plot of the Oronsay particle size data for which the classification is known. There is a clear separation, even in the unmodified data, which corresponds to the data from the *Cnoc Coig* site and that from the *Caisteal nan Gillean* site. The data from the *Cnoc Coig* site is colored black while the data from the *Caisteal nan Gillean* site is gray. Generally the sand at the *Cnoc Coig* site tends to be much finer with heavy weights developed with the fine sieves (.063 to .18 mm) and light weights

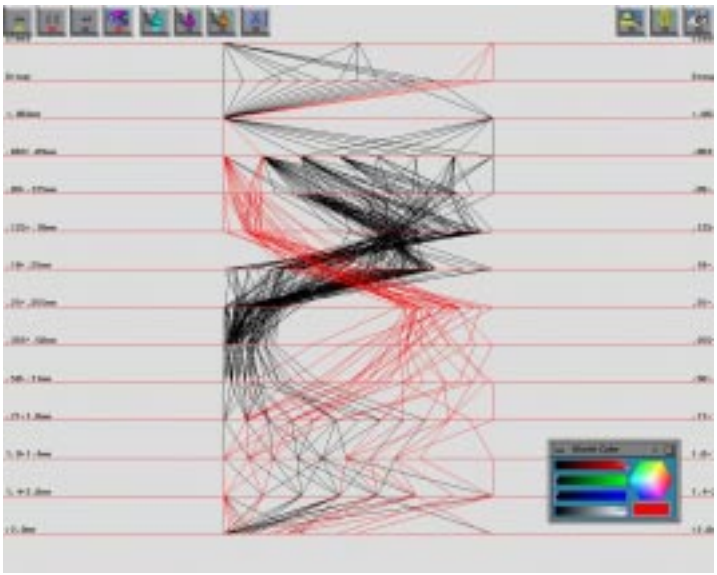


Figure 9: Original parallel coordinate plot of Oronsay *Cnoc Coig* and *Caisteal nan Gillean* data. The *Cnoc Coig* data is in black, the *Caisteal nan Gillean* data in gray (red). Data from the two sources strongly separate with the *Cnoc Coig* sand generally being much finer than the sand from the *Caisteal nan Gillean* site.

from the coarser sieves (.25 to .71 mm). The sand from the *Caisteal nan Gillean* site tends to be much coarser with heavier weights from the coarse sieves (.25 to .71 mm) and lighter weights from the finer sieves (.063 to .18 mm). The sands are so different in character that we decided to analyze the two clusters separately. The gap in the .25–.355mm axis and also in the .355–.50mm axis separates the two clusters perfectly. Also worth noting is that the measurements on the following axes, $<$ .063mm, .063–.09mm, .71–1.0mm, 1.0–1.4mm, 1.4–2.0mm, all are highly quantized as can be seen from the equal spacing of the data along these axes. Since quantization effects can masquerade as clusters during a grand tour, we decided to use only a partial grand tour based on the axes .09–.125mm through .50–.71mm. This leads to a six–dimensional partial grand tour. However, even though the other data was ignored for purposes of classification, the classifications were nearly perfect.

The results of the brush–tour strategy applied to the sand from the *Cnoc Coig* site are illustrated in Figure 10. Although the colors are not apparent in the reproduction in this paper, the illustrations with colors are available on the webpage. We shall, nonetheless, refer to the colors in the following description of our analysis. All clustering was done solely by the brush–tour strategy, ignoring any side information from the other sources. Figure 11

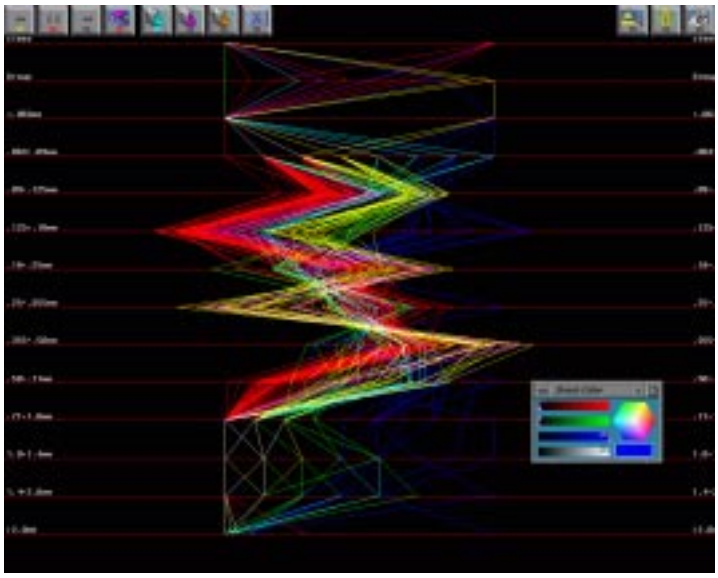


Figure 10: Parallel coordinate plot of *Cnoc Coig* data after completing the BRUSH–TOUR strategy. The data are divided into six clusters with red and magenta being "dune–like" sand. In the present image, all points are rendered in grayscale. The reader is referred to the webpage for full color illustrations.

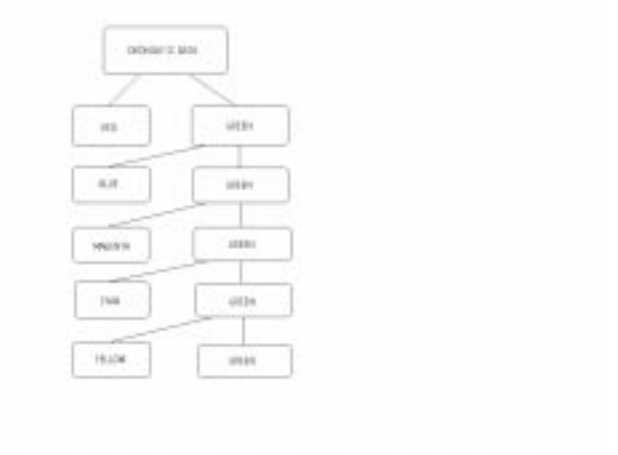


Figure 11: Sequence of decompositions for *Cnoc Coig* sand data. Red and magenta are basically the "dune–like" sands, other colors represent the "beach–like" sands. The strongest splits tend to occur early in the BRUSH–TOUR strategy. Thus, the "dune–beach" split was the most evident.

illustrates the sequence of decompositions of the subclusters. In some sense this tree structure indicates the relative similarity of subclusters. That is the discrepancy between red and green was the most obvious. After the red data had been removed from consideration, there was a subcluster, colored blue, of the green that was apparently different from the remaining green. After the blue data had been removed from consideration, there was a new subcluster apparent, now colored magenta, that was different from the remaining green. This process was repeated until the yellow subcluster was obtained. After these six subclusters were obtained of the Oronsay *Cnoc Coig* data, the grand tour was allowed to run so that any remaining subclusters could be identified. It was decided not to pursue any further splits.

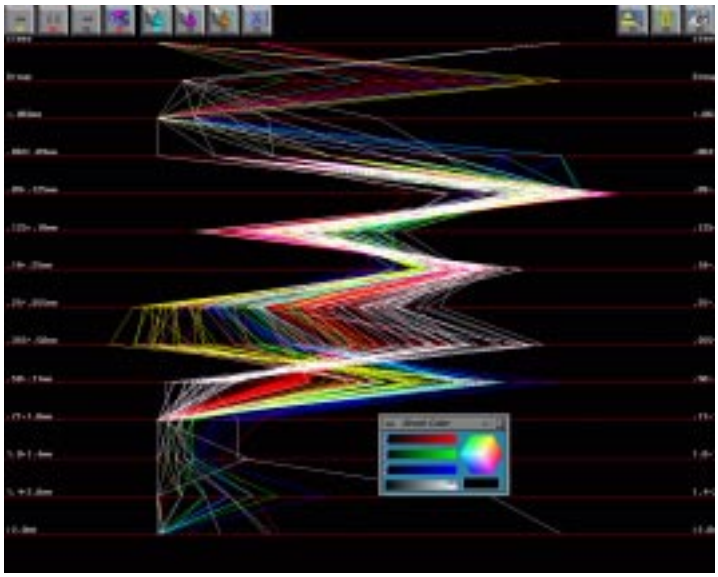


Figure 12: Parallel coordinate display of *Cnoc Coig* known (training) data and *Cnoc Coig* groups 2 and 3 after partial grand tour. The group 2 and 3 data are given in white. The group 2 and 3 data generally follow the red–magenta "dune–like" sand data. However the group 2 and 3 data clearly depart significantly in certain dimensions, notably along the .50–.71 mm axis in this illustration.

The clusters were then examined to see if they characterized any of the known subclasses. The characterizations were as follows:

RED	Groups 7, 8, 10, 11, 13, 14	Exception: 1 blue in 10
BLUE	Group 15	Exception: None
MAGENTA	Group 12	Exception: None
CYAN	Group 9	Exception: 2 reds in 9
YELLOW	Group 16, 17	Exception: 1 magenta in 16, 2 magenta in 17
GREEN	Group 6	Exception: None

Thus, there was an almost perfect classification with only six misclassifications in 149 observations. The one blue in group 10 corresponds to classifying a Mid–Beach observation from group 10 as a Mid–Beach observation from group 15. Similarly the three magentas in groups 16 and 17 correspond to classifying Upper–Beach observations as Base–of–Dune observations. The red cluster is most accurately characterized as the "dune–like" cluster. Thus the two red observations in group 9 correspond to finding two "dune–like" observations among the Lower–Beach data. These are the most anomalous misclassifications. Given the wide range of possible transport mechanisms for sand from one site to another which could easily explain the misclassifications, the overall conclusion must be that the brush–tour clustering strategy is capable of making remarkably subtle distinctions. It should be noted that the "dune–like" cluster consists of beach (groups 7, 10, and 11) and dune (groups 8, 13, and 14) samples while the remaining dune samples (group 12) form a separate cluster. In terms of the "classical" classification problem as presented in Flenley & Olbricht (1993) and Fieller et al. (1992), we therefore have reclassified at least three (or even four) entire groups. However, as already pointed out in the previous section on XGobi, it is reasonable to believe that the beach sand for these groups in fact is dune sand. This is supported by our results obtained in ExplorN.

Having thus classified the training data into "dune–like" (red and magenta clusters) and "beach–like" (other clusters), the desire is to classify the *Cnoc Coig* Midden samples as to whether they are "dune–like" or "beach–like" or, perhaps, neither. This was done in two stages. First, each of the individual samples of the archaeological *Cnoc Coig* Midden data (groups 1, 2, 3, 4 and 22) was compared against the already classified modern day *Cnoc Coig* data. Groups 2 and 3 (Sands below CC Midden) were aggregated and compared against the modern day *Cnoc Coig* data in Figure 12. On the .50–.71mm axis, the white cluster (groups 2 and 3) is clearly separated from the red cluster ("dune–like") which in turn is separated from the remaining clusters. The conclusion from this graphic and other graphics not reproduced is that groups 2 and 3 are different from either the "dune–like" sands or the "beach–like" sands. However, if one was forced to categorize the group 2 and 3 sand, it is more like the "dune–like" sand than the "beach–like" sand. This type of conclusion holds individually for group 22 data, for group 1 data, and for group 4 data as well. This result matches our findings reported in the

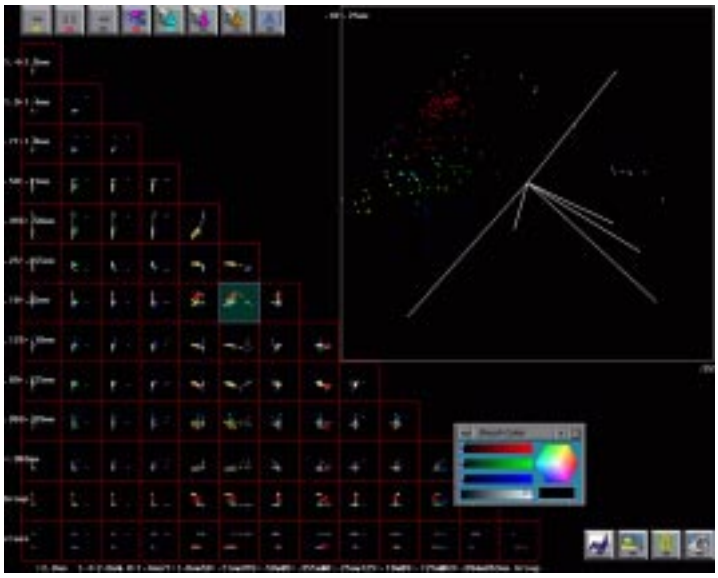


Figure 13: Scatterplot matrix of *Cnoc Coig* known (training) data against *Cnoc Coig* group 4 (unknown) data. The group 4 data are shown in white while the known *Cnoc Coig* data are shown in colors. The scatterplot diagram shown in the upper right–hand side of this illustration shows that the group 4 data are essentially orthogonal to all of the training data. Group 4 is definitely a different cluster, although if forced to characterize group 4 data, they would be closest to the "dune–like" sand data.

previous section on XGobi.

Figure 13 shows the scatterplot matrix (after grand tour rotation) for group 4 data against the training data. In the projection in the upper right corner, the group 4 data forms a linear feature orthogonal to the linear feature formed by the training data.

Finally we refer to Figure 14 and Figure 15. In these images, we have returned to classify the entire *Cnoc Coig* data set. The red cluster is again our "dune–like" cluster, the green cluster is our "beach–like" cluster and the white cluster is the entire unknown *Cnoc Coig* data (groups 1 to 4 and 22). From the white cluster, we have removed a number of outliers. Figure 14 represents the parallel coordinate plot which again clearly separates the unknown *Cnoc Coig* sands from the other two clusters. Again if one were forced to choose, the unknown sands seem to be closest to the "dune–like" sands. Figure 15 is the scatterplot matrix with a density plot for the highlighted scatterplot. The tallest mode in the density corresponds to the tightly clustered "dune–like" red cluster. The two smaller modes on the right–hand side correspond to the local modes in the "beach–like" green cluster. The mode on the left–hand

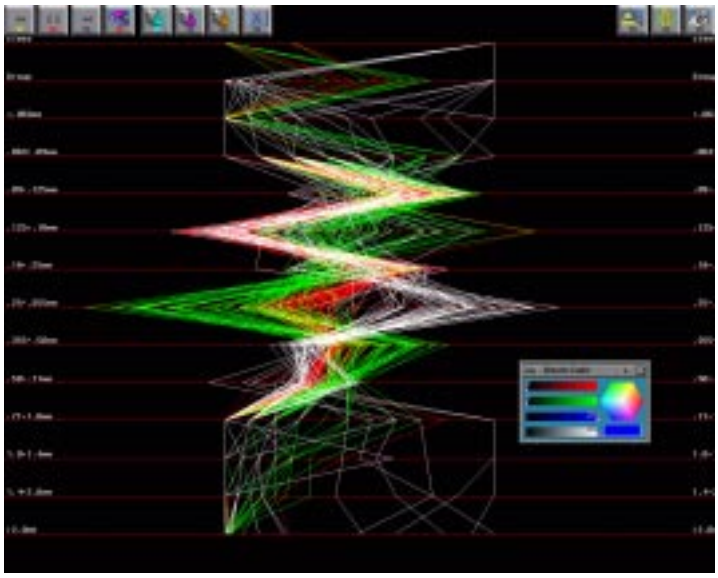


Figure 14: Simplified parallel coordinate display of all *Cnoc Coig* data after partial grand tour. "Dune–like" sands are shown in red, "beach–like" sands are shown in green, and "unknown" sands shown in white. The unknown class is distinct from both "dune–like" sand and "beach–like" sand. This is particularly clear in the .25–.355 axis.

side of the density corresponds to the "unknown" white cluster.

The analysis of the *Caisteal nan Gillean* data follows a similar pattern. The *Caisteal nan Gillean* data is considerably simpler since there are only three classes for the training data. We will not pursue the analysis of this data in the present paper using the ExplorN tool although the full set of graphics for this analysis is available on our webpage. Again there is a similar conclusion. The unknown data forms a cluster which can be separated from both the "dune–like" cluster as well as the "beach–like" cluster. Forced to make a choice, the unknown *Caisteal nan Gillean* data most resembles the "dune–like" *Caisteal nan Gillean* data. Once again, this matches our results obtained through XGobi.

6 Analysis in MANET

Our graphical analysis with Manet aims at establishing classification rules to distinguish between dune and beach sand samples of the Oronsay particle size data. Analogous to the classical discriminant analysis, we first try to find a good discrimination rule. Then we use this rule to classify pre-

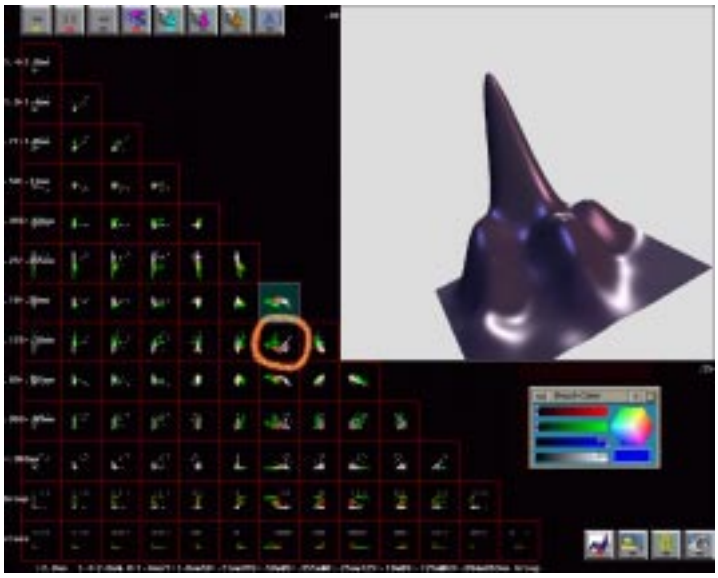


Figure 15: Simplified scatterplot matrix of all *Cnoc Coig* data after partial grand tour. The coloring is as in Figure 14. A density plot of the circled scatterplot is shown in the upper right. In the density plot, the tallest bump/mode corresponds to the red "dune–like" sand, the two smaller bump/modes on the right correspond to the green "beach–like" sand, and the smaller bump/mode on the left corresponds to the white "unknown" sand. The "unknown" sand is most like the "dune–like" sand, but still rather distinct.

viously unknown samples. Various automatic procedures are available for this purpose, which typically yield rules but no interpretation. The classic graphical approach just shows the clusters but does not attempt to indicate a characterization of the clusters.

For a data set with twelve variables only boxplots and dotplots need little enough space to be shown at one time on a regular screen. So we focus on them, even when histograms and scatterplots usually will show the clusters much clearer. We started with the 149 known samples and looked at dotplots for all particle sizes in Figure 16.

Dotplots tell a lot about the internal structure of the data. For the Oronsay particle size data the granularity in many variables is apparent. This is mainly a problem of measurement accuracy (rounding values to just one significant digit). All of these points represent a number of observations and the overplotting of points is visualized in MANET by tonal highlighting, i. e., the brightness of a point shows the frequency of its occurrence. A bright color represents many points while a dark color represents just a few points. Granularity is more present for the extreme particle sizes. Also the skewness

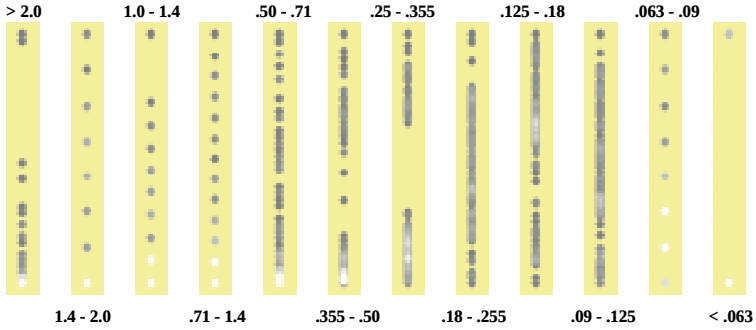


Figure 16: Dotplots of all variables for 149 known samples. Bright colors represent many points and dark colors only a few points. Measurements on the extreme particle sizes > 2.0mm, $1.4 - 2.0$mm, $1.0 - 1.4$mm, $.71 - 1.4$mm, $.063 - .09$mm, and $< .06$mm are highly quantized. Large gaps in the data are apparent for variables $.355 - .50$mm and $.25 - .355$mm.

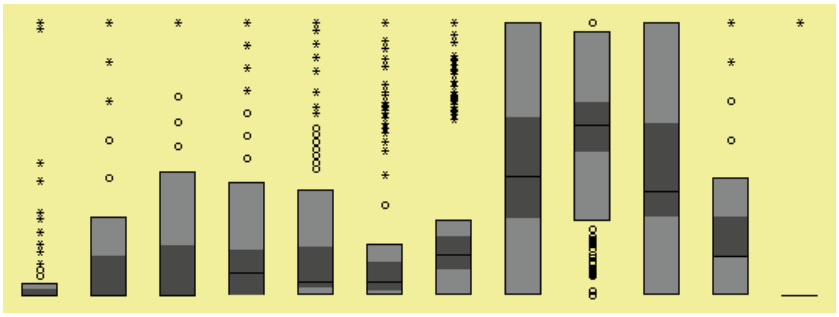


Figure 17: Boxplots for all 149 training samples. Variables > 2.0mm, $1.4 - 2.0$mm, $1.0 - 1.4$mm, ..., $.063 - .09$mm, and $< .06$mm are displayed from left to right. Only two variables, $.18 - .25$mm and $.09 - .125$mm, show no outliers. These two distributions are also only slightly skewed. Data on all the other variables is skewed to the right, but data on particle size $.125 - .18$mm is skewed to the left.

of the distributions can be seen in the dot plots. This becomes even clearer when looking at the boxplots in Figure 17. We also see that the extreme particle sizes produce much more outliers. Due to the discreteness and the outlier frequency we mainly ignore the extreme particle sizes and focus on the "inner" variables: particle sizes between .09g and .50g.

With linked views, three approaches for finding classification structure are viable: first, one can start with one–dimensional clusters and check via linked highlighting how these clusters can be interpreted and whether they also show up in other variables. Second, one can toggle between the known classes and check in the other variables whether one can find significant changes

between two consecutive pictures. And third, one can systematically increase a highlighted subset and check when a significant change shows up in the class variable.

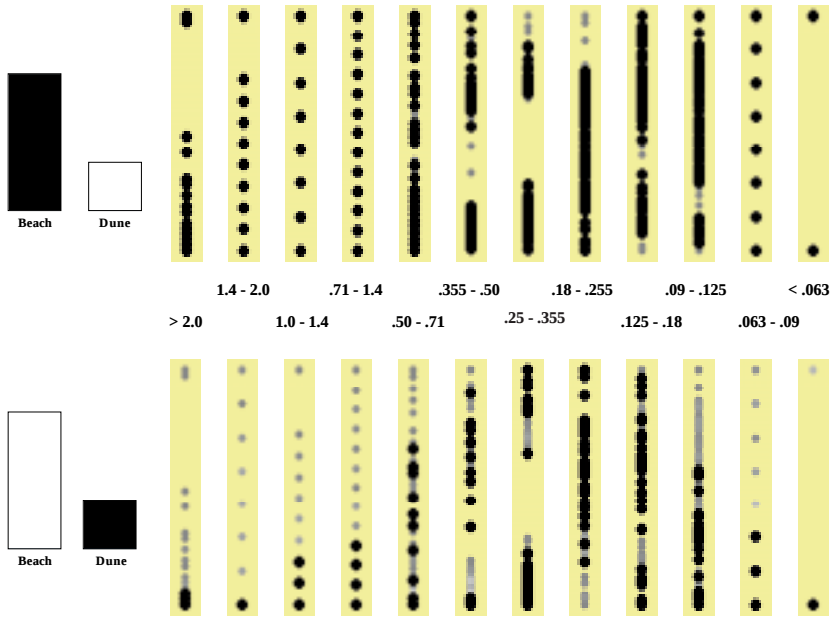


Figure 18: No classification in dotplots between dune and beach for the 149 known samples for both locations when selecting beach (top row) and dune (bottom row). Due to overplotting it appears that the same points have been selected more than once. In reality, a visible point sometimes represents a larger number of samples and it is already highlighted if just one of these samples is selected.

We started with the second approach and looked at univariate dotplots for all particle sizes and alternately selected in a bar chart the two sand types, see Figure 18. In the top row of this figure we selected the beach samples, in the bottom row the dune samples. Unfortunately, we cannot see any clustering structure in the highlighting of some variable in this figure. One reason for this is the huge amount of overplotting, so that slight differences cannot be seen. The other reason is that — as already noted in the literature and the two previous sections on XGobi and ExplorN — the two locations *Cnoc Coig* and *Caisteal nan Gillean* have rather different values.

On the other hand, low–dimensional clusters are especially apparent for the variables '0.355 – 0.5 mm', '0.25 – 0.355 mm' and '0.125 – 0.18 mm'. Pursuing the first visual classification approach we selected the top cluster in variable '0.25 – 0.355 mm' in Figure 19 (bottom row) and saw that these points also constitute clusters for particle sizes '$[0.355, 0.5)$ mm', '$[0.125, 0.18)$ mm', and '$[0.09, 0.125)$ mm'. In the bar chart on the right in Figure 19 (bottom row)

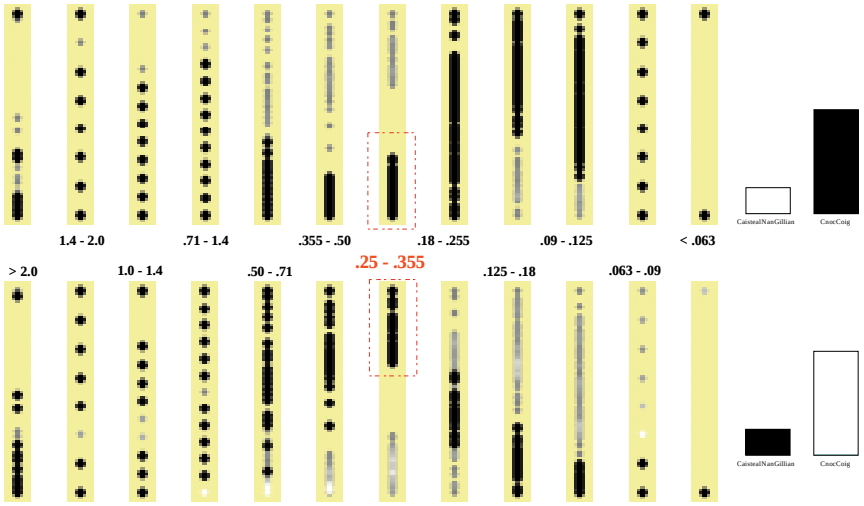


Figure 19: Clear classification between sites *Cnoc Coig* (top row) and *Caisteal nan Gillean* (bottom row) when selecting clusters of variable '[0.25, 0.355) mm'. Also the variables '[0.355, 0.5) mm', '[0.125, 0.18) mm', and '[0.09, 0.125) mm' allow a clear separation between *Cnoc Coig* and *Caisteal nan Gillean*.

we see that all highlighted points are *Caisteal nan Gillean* samples. Selecting the points from the bottom cluster in variable '0.25 – 0.355 mm' in Figure 19 (top row) highlights all *Cnoc Coig* samples.

In Figure 19 we also see that the particles of the *Cnoc Coig* samples tend to be much smaller than those at the *Caisteal nan Gillean* site since the latter have high weights for bigger particle sizes, whereas the former have high weights for small particle sizes. This is the same observation that has also been made in the ExplorN section. The cross–over takes place in variable '0.18 – 0.25 mm' and explains why this variable does not show such a clear clustering. The clearest separation is obtained for the [0.25, 0.355) mm particle size and a plausible rule to classify between sites can be based just on this variable. By interrogating the dotplot we can easily get the values for the cluster boundaries, which yields the rule "classify as '*Caisteal nan Gillean*' if values for variable '0.25 – 0.355 mm' are greater than 16g and classify as '*Cnoc Coig*' if values are smaller than 8g" (by rounding to the closest integer). Similar rules can be established using one of the variables '[0.355, 0.5) mm', '[0.125, 0.18) mm', and '[0.09, 0.125) mm' as well. For example, for particles sizes '[0.355, 0.5) mm' we would use "classify as '*Caisteal nan Gillean*' if values are greater than 3g and classify as '*Cnoc Coig*' if values are smaller than or equal 2g". Using the variable with the greatest gap between the clusters leaves a large range of possible values unclassified (all the values in the interval (8, 16) in contrast to the values between (2, 3) when using variable '[0.355, 0.5)

mm').

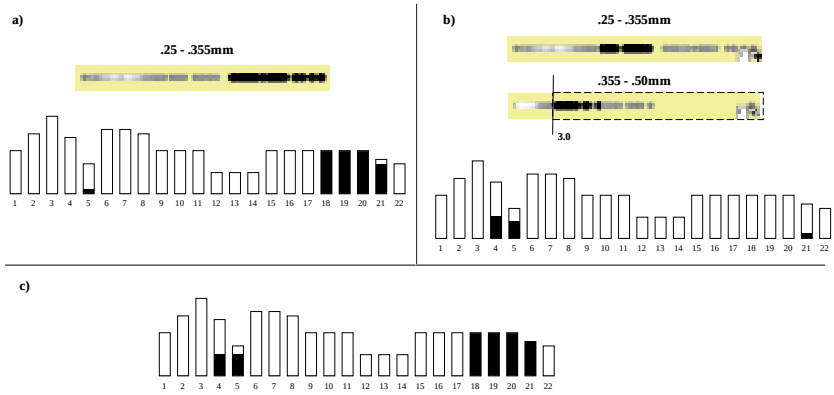


Figure 20: We apply the classification rule of training data to the entire data set of 226 samples. (a) Selecting the right–hand cluster in variable .25 – .355mm highlights all training samples at *Caisteal nan Gillean*, one sample of test group 5 and also all but one of group 21. (b) However, 16 points fall between the previously established clusters. Those points are classified by using the classification rule based on the training data for variable .355 – .50mm. (c) The resulting classification is correct for groups 18 to 21, but misses two samples in group 5 and misclassifies five samples in group 4.

In the next step we used this classification rule for the complete 226 samples. Now the observations for variable '0.25 – 0.355 mm' fall into four clusters (see Figure 20(a)): two big ones at the higher and the lower end respectively, each of which contains one of the clusters found for the known samples, and two small ones in the middle. At this point the data tells us to adjust the previously established classification rule to the new cluster boundaries and we therefore use "classify as '*Caisteal nan Gillean*' if values for variable '0.25 – 0.355 mm' are greater than 16g and classify as '*Cnoc Coig*' if values are smaller than 9g" (instead of 8g as with the training data). Unfortunately, 16 observations (mainly from groups 4 and 5) lie between the two great clusters, see Figure 20(b)). Two options are possible: we either allow the two additional clusters in the center to be from one or two additional sites or we force a classification of the center clusters to the two known sites. If we insist on a classification to the two known sites, we can add information from other variables to our classification rule. If we additionally use the above mentioned rule for variable '0.355 – 0.5 mm' (see Figure 20(c)) we come up with the classification rule visualized in the classification tree in Figure 21.

When comparing the so–obtained results with the classification as given in Table 1, we see that we have misclassified two *Caisteal nan Gillean* midden samples (out of group 5) and five *Cnoc Coig* samples (out of group 4).

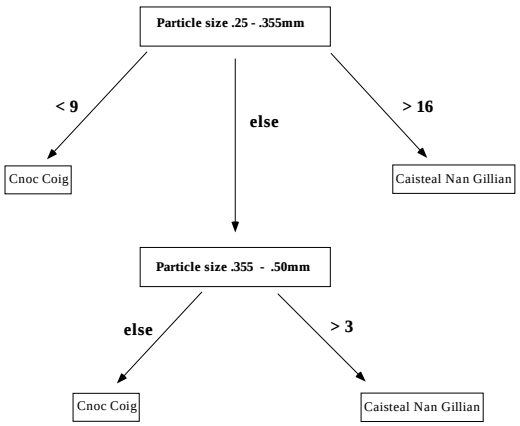


Figure 21: Final classification tree for separating all 226 sand samples by site.

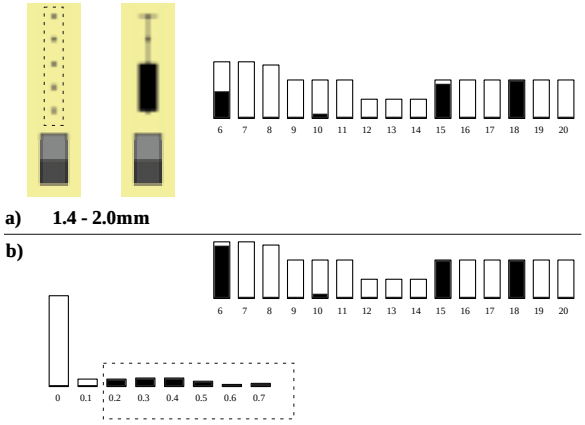


Figure 22: (a) Outliers for particle size '[1.4, 2.0)mm' fall mainly in groups 15 and 18. (b) Enlarging the selected group to all samples with values greater than 0.1g for variable 1.4 – 2.0mm highlights all but one sample of group 6, all samples of groups 15 and 18, and one sample of group 10 (misclassified).

Although we did not recognize a good separation between dune and beach in general, there can be found some structure without separating both sites. When looking at the outliers for the large particle sizes, one realizes that they all stem from lower beach samples. The largest outlier group is obtained from particle sizes '[1.4, 2.0)mm' (all samples with more than 0.2g) (Figure 22(a)). If we even enlarge the selected group to all samples with more than 0.1g, we obtain a nice description of some of the lower beach samples, i. e., groups 6, 15, and 18 — with one sample of group 6 not highlighted and one mid–beach sample of group 10 misclassified (Figure 22(b)).

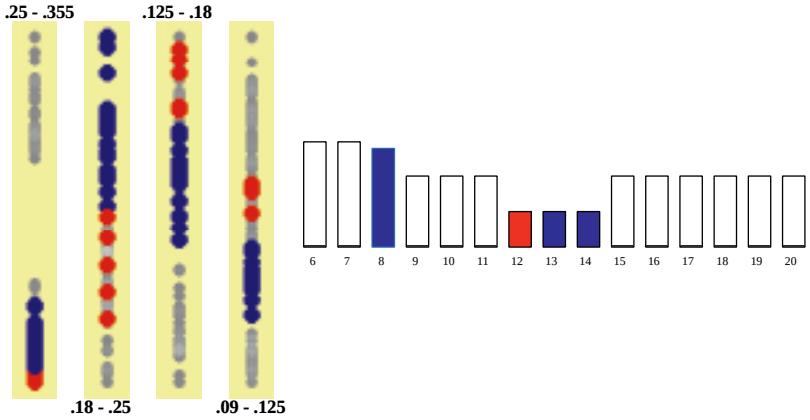


Figure 23: The dune samples at *Cnoc Coig* split into two groups for variables .18 – .25mm and .125 – .18mm. Group 12 (marked red, i.e. bigger light dots) is different from groups 8, 13, and 14 (all marked blue, i.e. bigger dark dots). From the position of the two groups it can be concluded that dune sands of group 12 are much finer since these samples have heavier weights from the finer sieves and lighter weights from the coarser sieves.

In the next step we looked at the dune samples at *Cnoc Coig* (Figure 23). Selecting groups 8, 12, 13, and 14 in the bar chart immediately explains the bad classification results. For particle sizes '[.25, .355)mm', '[.18, .25)mm', and '[.125, .18)mm' the dune samples already cover the whole range. Moreover, they split into two groups: group 12 shows a different behavior than the other three dune groups 8, 13, and 14, which also has been noticed in the previous section on ExplorN. The base of dune sands (group 12) are considerably finer than the upper dune, face of dune and top of dune sands (groups 8, 13, and 14). Group 12 developed heavy weights with the fine sieves and light weights with the coarser sieves, whereas the other groups had light weights with the finer sieves and heavy weights with the coarser sieves.

From our original 12 variables we already excluded 6 at the very beginning due to their quantized structure. The measurements for particle sizes .50 – .71mm and .355 – .50mm are also quite discretized and should only be used with care. Three other variables now have shown to be inappropriate to be used as a first step towards classification between dune and beach. We therefore are now left with one variable: '[0.09, 0.125)mm'. Looking at a histogram for particle size '[0.09, 0.125)mm' indicates a distribution with about four modes. The left–hand group contains the samples from the *Caisteal nan Gillean* sites. Checking the other modes results in three clusters within the *Cnoc Coig* samples: the first cluster shown in Figure 24(a) is formed by the groups 7, 8, 11, 13, and 14; the second cluster (Figure 24(b)) consists of groups 6, 10, 12, and 17, and the third cluster (Figure 24(c)) is

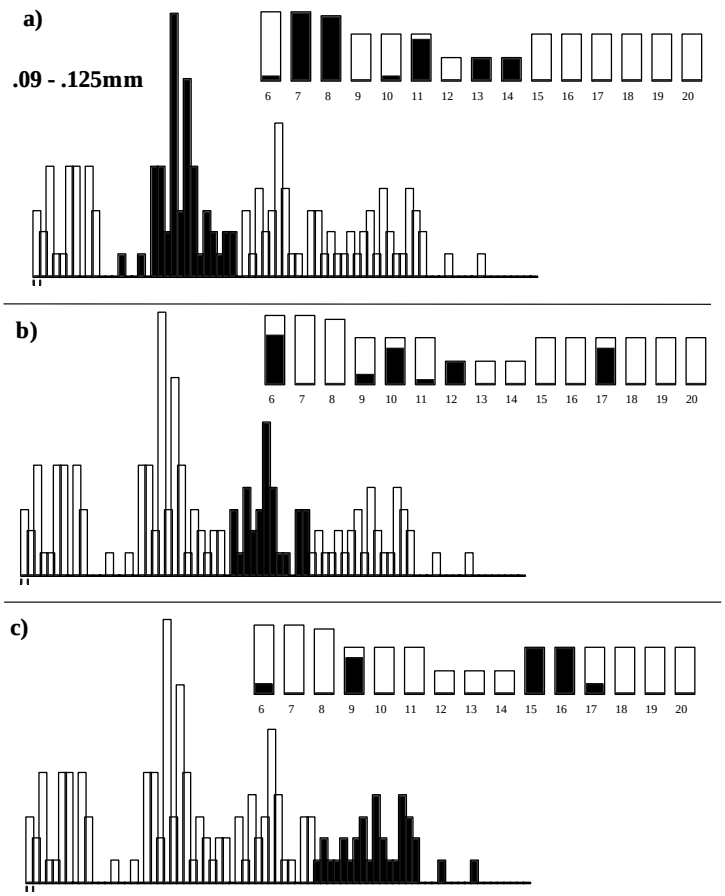


Figure 24: The distribution for particle size '[0.09, 0.125)mm' seems to be a mixture of four individual distributions: one for the *Caisteal nan Gillean* samples, one for groups 7, 8, 11, 13, and 14 (in a), one for groups 6, 10, 12, and 17 (in b), and one for groups 9, 15, and 16 (in c).

built by groups 9, 15, and 16. In two clusters, dune samples are mixed with beach samples.

We can further clean these clusters by using information of some of the other variables: we obtain groups 7, 8, 11, 13, and 14 with one misclassification out of group 10 when intersecting the above cluster with the lower cluster in the highlighting of particle size '[.355, .50mm)'. Selecting the groups individually in the bar chart, we can see in the dotplots that the two upper beach groups 7 and 11 constitute a rather homogeneous cluster but that the dune sites have a higher variability and cover for all variables the range of the upper beach samples.

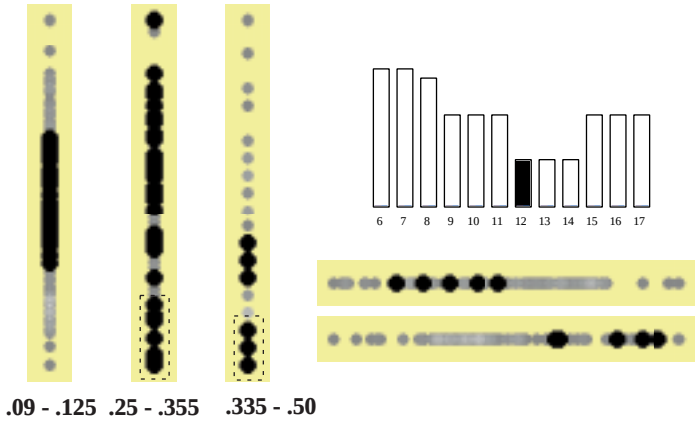


Figure 25: Individual separation of group 12 by sequentially selecting subclusters (the dashed areas) in the respective highlighting of particle sizes '[0.09, 0.125)mm', '[0.25, 0.355)mm', and '[0.335, 0.50)mm'. No further clustering can be found in the dotplots of the other variables .125 – .18mm and .18 – .25mm (right under bar chart for group).

Group 12 can be easily separated from its cluster above (Figure 25). After having selected the appropriate cluster in variable .09 – .125mm, we select the lowest cluster in variable .25 – .355mm. The resulting selections split into two groups for variable .355 – .50mm. In the other "inner" variables (.125 – .18mm and .18 - .25mm no significant further clustering can then be found.

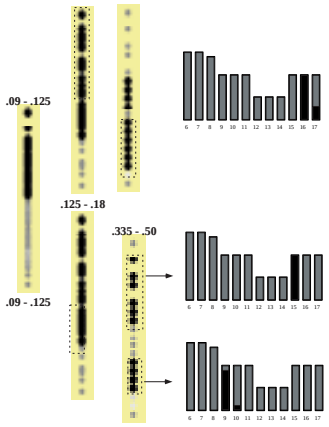


Figure 26: Individual separation of groups 9, 15 and 16 by sequentially selecting subclusters in the respective highlighting of particle sizes '[0.09, 0.125)mm', '[0.125, 0.18)mm', and '[0.335, 0.50)mm'.

Using similar selection sequences, we can individually separate each of the

groups 9, 15, and 16 from their above grouping (Figure 26). All our classification for the *Cnoc Coig* samples start with the natural clustering of particle size '[0.09, 0.125)mm'.

In summary, we end up with a separation in six groups: two "dune–like" groups (groups 7, 8, 11, 13, 14, and group 12) and four "beach–like" groups (each of the groups 9, 15, and 16 and the rest, i. e., groups 6, 10, and 17). In the "dune–like" cluster we have one additional sample from group 10, in the "beach–like" samples we miss one sample of groups 9 and 10, respectively, and misclassify 3 observations from group 17 into group 16. So, altogether, we end up with the same six misclassifications as in the previous section with ExplorN.

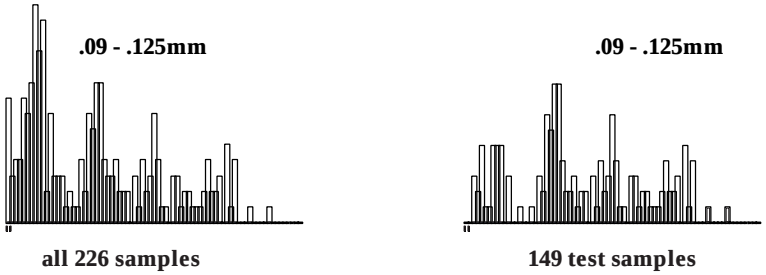


Figure 27: Histograms of the root variable 'particle sizes [0.09, 0.125)mm' for modern samples and entire data set.

When turning towards the 77 archaeological samples, we go to the root of our classification tree and draw a histogram for particle size '[0.09, 0.125)mm' using the same bin width as for the training data (Figure 27). Again, we conclude four–modality of the distribution. We immediately recognize that almost all of the archaeological samples fall into the lowest cluster, i. e., the *Caisteal nan Gillean* cluster. However, from the other variables there is enough evidence not to mix the archaeological samples with the modern *Caisteal nan Gillean* samples. Preferably, we would define a new group since the archaeological samples do not have much in common with the other *Cnoc Coig* samples. According to the root variable in our hierarchical classification, these samples are closer to the "dune–like" samples than to the beach samples. From those 178 samples that have been classified as *Cnoc Coig* in the first step, only three (one of group 2 and both misclassified samples of group 5) do not fall into the lowest cluster of particle size '[0.09, 0.125)mm'. These three samples directly fall into the "dune–like" cluster.

In the next step, we looked at the 30 known *Caisteal nan Gillean* samples (groups 18 to 20). Toggling between dune and beach in the bar chart gives a clear picture for the separation: all the beach samples yield high measurements for particle sizes greater than .355mm or lower than .18mm and low values for particle sizes between .18 and .355mm. The dune samples resulted

in complementary measurements. One dune sample shows a contrary behavior, low values for variable '0.25 – 0.355 mm' and high values for variable '0.125 – 0.18 mm'. The structure is clear when looking at all variables but it is much more difficult to assess clear separation rules since the two groups overlap in the center of each marginal distribution.

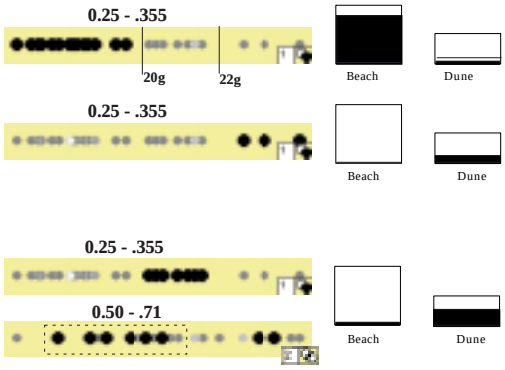


Figure 28: Classification of 30 known sand samples at *Caisteal nan Gillean*.

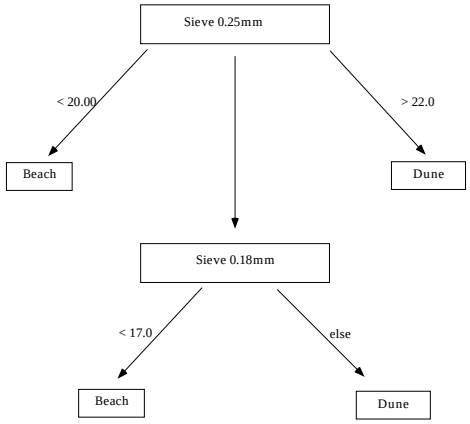


Figure 29: Classification tree based on 30 known samples for separating sand types at *Caisteal nan Gillean*.

The best univariate separation is obtained for variable '0.25 – 0.355 mm' where all samples with values smaller than 20g are beach and those larger than 22g are dune. Selecting the intermediate group (see Figure 28) shows a split into one homogeneous group and two outliers for variable '0.50 – 0.71 mm', the group being in majority dune, the outliers beach. The resulting

classification rule which misclassifies one dune sample and one beach sample is shown in Figure 29.

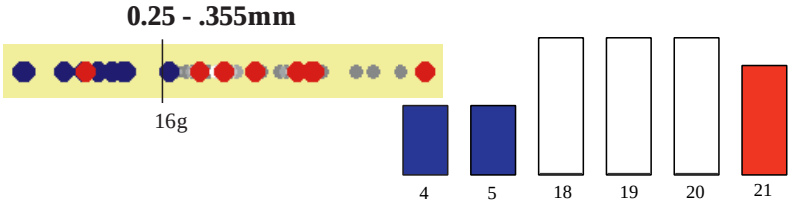


Figure 30: Classification of the test data at *Caisteal nan Gillean*. Groups 4 and 5 build a separate cluster that would be classified as neither dune nor beach. Half of the samples in group 21 fall into the beach and the dune clusters.

Applying this rule to all the 48 samples that have been classified as *Caisteal nan Gillean* in the first step of our analysis (Figure 20 (c)) yields three clusters. Four samples (3 of group 21 and 1 of group 5) clearly fall into the beach cluster. Actually, the one sample out of group 5 that is "beach–like" is the sample with the smallest sum (51.8g), so this should be taken with care. As well, four samples (all from group 21) fall into the dune cluster. But 10 samples (all of group 4, four out of five in group 5 and one of group 21) constitute a separate cluster with values way below the beach samples (varying from about 10g to 14g, in contrast to 16g to 20g for the beach samples) (Figure 30).

So again, if we had a third choice, we would have classified groups 4 and 5 as neither dune nor beach. Forced to make a classification, our judgement is that they are closer to beach than to dune, a contradiction to the results obtained by ExplorN. It is worth noting that the lower beach samples constitute a homogeneous subcluster and differ much more from the dune samples than the upper beach samples do. Of course we know that group 4 is CC Soil Pit and thus it is not surprising that it has nothing in common with modern *Caisteal nan Gillean* samples. But the environmental conditions a few thousand years ago are clearly unknown to us. In some sense, this step could be considered a graphical hypothesis test where the null hypothesis is that CC Soil Pit samples are related to modern *Caisteal nan Gillean* samples. We graphically could reject this hypothesis.

To summarize our work with Manet, which features of linked graphics have we primarily used for this analysis? First of all, we needed linked views to toggle between the type selections made in a bar chart. Dynamically toggling between the selections was superior to using different colors or plot symbols in almost all plots where a lot of overplotting was present. The overplotting could have been partially avoided by using higher–dimensional views but only one–dimensional dotplots and boxplots use little enough screen space so that 12 of them can be drawn at the same time. The next essential feature

was that we have been able to easily combine various selections. Having the selection sequences tool in MANET gives a short and easy way to switch from one branch in the classification tree to another. To set up classification rules, we need to be able to interrogate the plot to get the boundary values, so that we can transform the visual clusters into implementable rules. The linking of different plot types — bar charts, dotplots, histograms, and boxplots — enhances the interpretation by providing different aspects of the data in an easy and efficient way.

7 Conclusion

While it took about 10 years — the timespan between Fieller et al. (1984) and Flenley & Olbricht (1993) — before it was clearly stated that one cannot simply discriminate between beach and dune sand but one has to consider the site as first discrimination factor, this result becomes immediately obvious in plots such as in Figures 3(b), 9, and 19 within seconds when using the visual approaches presented within this paper. One does not even need to know about this result previously or spend any efforts to search for it — it shows up almost by itself.

One should carefully think about the implications of the classification of the archaeological *Caisteal nan Gillean* samples (groups 5 and 21). If we had permitted a third option neither "dune–like" nor "beach–like", we probably would have selected this classification for both groups based on Figures 7(b) and 30. Alternatively, our three visual classification approaches ended in three different classifications of these samples, suggesting that the archaeological *Caisteal nan Gillean* samples are not simply comparable with modern dune or beach samples.

Likewise, the analysis of *Cnoc Coig*, especially the parallel coordinate analysis (see Figures 12 through 15), suggests that if one were forced to draw a conclusion, the sand below the midden at *Cnoc Coig* behaved more like dune sand rather than beach sand. Indeed all of the sand in the midden area as well as the soil pit appears to be closest to dune sand. However, the midden area sand is distinctly different from dune sand and from beach sand and would be classified by the visual classification techniques as neither "dune–like" nor "beach–like". The conclusions about the seaward shift of the beach–dune interface appears to be unwarranted based on the particle size data. This does not preclude support for this conjecture based on chemical evidence or evidence based on particle shape.

XGOBI's Figure 6(b), the individability of groups 7, 8, 11, 13, 14 (and 10) that has been found in EXPLORN and MANET, and the rating in Andrews et al. (1987) where the CC and the TG transects fall into a region with wind

exposure rated as "Exposed" are a very good explanation for the misclassification of entire upper beach subclusters that occur in the classical literature on this topic. However, the explanation is easy. We assume that the sand at these locations most likely is dune sand, blown away by the wind.

Of course, there are some restrictions of graphical (and similar) approaches when analyzing particle size data as pointed out in Fieller et al. (1992). The visual approaches require that all samples are sieved through the same set of sieves to obtain measurements for the same set of variables. The statistical analyses based on parameter estimates for particle size distributions will also work if different sets of sieves have been used. However, we are strongly convinced that our visual approaches will also be very useful in case that only parameter estimates are available for the particle size data, in particular, since Fieller et al. (1984), page 650, state that "*this case study has concentrated on plots of α versus μ. In other cases, our unpublished work indicates the values of inspecting three–dimensional plots of α against β against μ. Single plots of β against μ have also provided adequate environmental discrimination in other situations*". Our graphical approaches allow such plots and many additional features in a dynamic environment.

Anyway, the main goal of this paper was not a full reevaluation of the Oronsay particle size data through each of our three approaches but a comparison and evaluation how well these packages and paradigms can be used for visual clustering and classification. All three approaches found the known clusters in the Oronsay data set. Methodologically, the linked views approach differs from the other two in respect of the user's role. The grand tour — in parallel coordinates as well as in standard Euclidean coordinates — puts the user in the role of a spectator who has to examine a series of projections until an interesting pattern shows up. For linked views the user has to choose the conditioning sets which determine the complete analysis. At this point, the user's experience as well as additional subject information will strongly influence the results. At least theoretically the grand tour will pass through all possible projections and will therefore show all possible clusters. In the current environments for linked low–dimensional views no guidance is given to the user to perform a systematic search through all possible conditioning sets. Thus, it is very likely that the user will just see a fraction of all possible views. However, linked low–dimensional views enhance the interpretation of clusters detected. Since they rely only on the canonical projections of a high–dimensional point cloud to the coordinate space determined by the individual variables, it is straightforward to set up classification rules that describe the visually found clusters. Another strength of linked low–dimensional views is that for mixed data sets with discrete and continuous variables each variable type can be visualized by the best suited graph.

Using the grand tour in Euclidean coordinate space seems to be the most familiar approach. The parallel coordinates have the advantage that a fairly

big number of variables can be shown on the screen at the same time. Obviously, XGobi and ExplorN share many features since they both support high–dimensional rotating scatterplots and parallel coordinate plots. However, there are many features that are available in only one of the packages, e. g., XGobi's projection–pursuit–guided grand tour or ExplorN's parallel coordinate grand tour. But these features have been proven to be very effective for the visual clustering and classification. Variations of the Brush–Tour strategy have been used in both XGobi and ExplorN to detect clusters.

Our main conclusion of our research is, that there is nothing such as a "best" dynamic graphics program for the visual clustering and classification. All three programs — XGobi, ExplorN, and Manet — and the related paradigms have unique advantages and disadvantages that highly relate to the user's experience and personal preferences. For the Oronsay particle size data, we did not notice that any of the programs provided better results — however, it should be noted that the programs were operated by well–trained individuals for each software package. Overall, we recommend that future research and work should focus on the development of an integrated software product that offers combined features of all three of the approaches presented here. Additional features, e. g., a dynamic graphics approach for cluster analysis based on dendrograms, constellation plots, and flipped empirical distribution function (FEDF) scatterplots as presented in (Lee, Kim, Huh & Jeong 1997), might be worth to be included as well into such a dynamic graphical software system.

Acknowledgments

We would like to thank Nick Fieller for providing us with the Oronsay particle size data set and additional background information. Thanks are also due to Walter Olbricht for his additional comments and to Qiang Luo who assisted with the preparation of data and the analysis. The work of all three authors was supported in part by the NSF with a Group Infrastructure Grant DMS–9631351. In addition, the work of Edward Wegman was supported in part by the Army Research Office under grant DAAH04–94–G–0267. This work was initiated when Adalbert Wilhelm was visiting the Center for Computational Statistics as a Habilitanden–Stipendiat of the Deutsche Forschungsgemeinschaft under contract Wi 1584/1–1.

References

Andrews, M. V., Gilbertson, D. D. & Kent, M. (1987), Storm Frequencies along the Mesolithic Coastline, *in* P. A. Mellars, ed., 'Excavations on

Oronsay: Prehistoric Human Ecology on a Small Island', Edinburgh University Press, Edinburgh, UK, pp. 108–114.

Asimov, D. (1985), 'The Grand Tour: A Tool for Viewing Multidimensional Data', *SIAM Journal on Scientific and Statistical Computing* **6**(1), 128–143.

Becker, R., Chambers, J. & Wilks, A. (1988), *The New S Language — A Programming Environment for Data Analysis and Graphics*, Wadsworth and Brooks/Cole, Pacific Grove, CA.

Buja, A. & Asimov, D. (1986), Grand Tour Methods: An Outline, *in* D. M. Allen, ed., 'Proceedings of the 17th Symposium on the Interface between Computer Science and Statistics, Lexington, KY', Elsevier, pp. 63–67.

Buja, A., Cook, D. & Swayne, D. F. (1996), 'Interactive High–Dimensional Data Visualization', *Journal of Computational and Graphical Statistics* **5**(1), 78–99.

Carr, D. B., Wegman, E. J. & Luo, Q. (1997), ExplorN: Design Considerations Past and Present, Technical Report 137, Center for Computational Statistics, George Mason University, Fairfax, VA.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Wadsworth, Monterey, CA.

Cleveland, W. S. & McGill, M. E., eds (1988), *Dynamic Graphics for Statistics*, Wadsworth & Brooks/Cole, Belmont, CA.

Cook, D., Buja, A. & Cabrera, J. (1993), 'Projection Pursuit Indices Based on Expansions with Orthonormal Functions', *Journal of Computational and Graphical Statistics* **2**(3), 225–250.

Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995), 'Grand Tour and Projection Pursuit', *Journal of Computational and Graphical Statistics* **4**(3), 155–172.

Cook, D., Majure, J. J., Symanzik, J. & Cressie, N. (1996), 'Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data Using Linked Software', *Computational Statistics: Special Issue on Computeraided Analysis of Spatial Data* **11**(4), 467–480.

Cook, D., Symanzik, J., Majure, J. J. & Cressie, N. (1997), 'Dynamic Graphics in a GIS: More Examples Using Linked Software', *Computers and Geosciences: Special Issue on Exploratory Cartographic Visualization* **23**(4), 371–385. Paper, CD, and `http://www.elsevier.nl/locate/cgvis`.

Eick, S. G. & Wills, G. J. (1995), 'High Interaction Graphics', *European Journal of Operational Research* **84**, 445–459.

Fieller, N. R. J., Flenley, E. C. & Olbricht, W. (1992), 'Statistics of Particle Size Data', *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **41**(1), 127–146.

Fieller, N. R. J., Gilbertson, D. D. & Olbricht, W. (1984), 'A new Method for Environmental Analysis of Particle Size Distribution Data from Shoreline Sediments', *Nature* **311**, 648–651.

Fieller, N. R. J., Gilbertson, D. D., Olbricht, W. & Timmins, D. A. Y. (1983), 'A Computer–Compatible Archive of Sedimentological Data from Oronsay, Inner Hebrides', Archive No. 2, Department of Prehistory and Archaeology, University of Sheffield, Sheffield, UK.

Fieller, N. R. J., Gilbertson, D. D. & Timmins, D. A. Y. (1987), Sedimentological Analyses of the Shell–Midden Sites, *in* P. A. Mellars, ed., 'Excavations on Oronsay: Prehistoric Human Ecology on a Small Island', Edinburgh University Press, Edinburgh, UK, pp. 78–90.

Flenley, E. C. & Olbricht, W. (1993), Classification of Archaeological Sands by Particle Size Analysis, *in* O. Opitz, B. Lausen & R. Klar, eds, 'Information and Classification. Concepts, Methods and Applications. Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e. V."', Springer, Berlin, Heidelberg, pp. 478–489.

Friedman, J. H. & Tukey, J. W. (1974), 'A Projection Pursuit Algorithm for Exploratory Data Analysis', *IEEE Transactions on Computers* **C–23**, 881–889.

Huber, P. J. (1985), 'Projection Pursuit (with Discussion)', *Annals of Statistics* **13**, 435–525.

Inselberg, A. (1985), 'The Plane with Parallel Coordinates', *The Visual Computer* **1**, 69–91.

Klinke, S. & Cook, D. (1997), 'Binning of Kernel–based Projection Pursuit Indices in XGobi', *Computational Statistics & Data Analysis* **27**(3), 363–369.

Koschat, M. A. & Swayne, D. F. (1996), 'Interactive Graphical Methods in the Analysis of Customer Panel Data (with Discussion)', *Journal of Business and Economic Statistics* **14**(1), 113–132.

Kruskal, J. B. (1969), Toward a Practical Method which Helps Uncover the Structure of a Set of Observations by Finding the Line Transformation which Optimizes a New "Index of Condensation", *in* R. C. Milton & J. A. Nelder, eds, 'Statistical Computation', Academic Press, New York, NY, pp. 427–440.

Lee, K. M., Kim, K. Y., Huh, M. Y. & Jeong, N. C. (1997), 'Dynamic Graphics Approach for Cluster Analysis', *Computing Science and Statistics* **28**, 385–388.

Olbricht, W. (1982), 'Modern Statistical Analysis of Ancient Sand'. MSc Thesis, University of Sheffield, Sheffield, UK.

Swayne, D. F., Cook, D. & Buja, A. (1998), 'XGobi: Interactive Dynamic Graphics in the X Window System', *Journal of Computational and Graphical Statistics* **7**(1), 113–130.

Symanzik, J., Majure, J. J. & Cook, D. (1996), 'Dynamic Graphics in a GIS: A Bidirectional Link between ArcView 2.0 and XGobi', *Computing Science and Statistics* **27**, 299–303.

Theus, M., Hofmann, H. & Wilhelm, A. (1998), 'Selection Sequences — Interactive Analysis of Massive Data Sets', *Computing Science and Statistics* **29**(1), 439–444.

Timmins, D. A. Y. (1981), 'Study of Sediment in Mesolithic Middens on Oronsay'. MA Thesis, University of Sheffield, Sheffield, UK.

Unwin, A. R., Hawkins, G., Hofmann, H. & Siegl, B. (1996), 'MANET — Missings Are Now Equally Treated', *Journal of Computational and Graphical Statistics* **5**, 113–122.

Wegman, E. J. (1990), 'Hyperdimensional Data Analysis Using Parallel Coordinates', *Journal of the American Statistical Association* **85**, 664–675.

Wegman, E. J. (1991), 'The Grand Tour in k–Dimensions', *Computing Science and Statistics* **22**, 127–136.

Wegman, E. J. (1995), 'Huge Data Sets and the Frontiers of Computational Feasibility', *Journal of Computational and Graphical Statistics* **4**(4), 281–295.

Wegman, E. J. & Carr, D. B. (1993), Statistical Graphics and Visualization, *in* C. R. Rao, ed., 'Handbook of Statistics, Vol. 9', Elsevier Science Publishers, Amsterdam, pp. 857–958.

Wegman, E. J. & Luo, Q. (1997*a*), 'High Dimensional Clustering using Parallel Coordinates and the Grand Tour', *Computing Science and Statistics* **28**, 361–368.

Wegman, E. J. & Luo, Q. (1997*b*), High Dimensional Clustering using Parallel Coordinates and the Grand Tour, *in* R. Klar & O. Opitz, eds, 'Classification and Knowledge Organization', Springer, pp. 93–101.

Wilhelm, A., Unwin, A. R. & Theus, M. (1996), Software for Interactive Statistical Graphics — A Review, *in* F. Faulbaum & W. Bandilla, eds, 'SoftStat '95 — Advances in Statistical Software 5', Lucius & Lucius, Stuttgart, pp. 3–12.