

Selection Sequences in MANET

Heike Hofmann, Martin Theus

University of Augsburg, 86135 Augsburg, Germany
e-mail: hofmann@mathpool.Uni-Augsburg.DE

Summary

Flexible and easy to use tools are of great importance in exploratory data analysis. They enable statisticians to detect structures in data. Often this is done by conditioning the data onto a subset to reduce both the number and in some sense the dimension. Successive selections are a very simple way for conditioning to concentrate on more relevant observations. The program MANET offers an easy and user-friendly method for working with such selections by so called *selection sequences*.

Keywords: MANET, exploratory data analysis, interactive statistical graphics, selection, *selection sequence*

1 Motivation

Generally selections of subgroups form one of the most important tools of exploratory data analysis (Theus, Martin (1996), Cleveland (1994)). Selections in particular are the key feature in an interactive graphical environment based on the paradigm of global linking and highlighting. Successive selections allow a stepwise reduction of data to the observations of interest. Because of the sequential procedure this method models the way analyses proceed and is therefore intuitive to use. When working with successive selection one soon realizes that most effort is not spent on studying the results but on keeping track of the actions necessary to achieve them. With categorical data, for example, limits are reached very soon: Selecting subgroups by intersecting only

three or four barcharts will usually overtax the user, as it is hard to monitor both the different combinations of levels and the corresponding results. On the other hand these analyses are important for exploring the data. Therefore we need methods which do the purely technical work for us and let us concentrate on the essential issues. In the example of the barley-dataset several ideas for improvement are obvious. In this dataset from the years 1931 and 1932 the yields of barley were measured at different sites. An interesting feature is an error made when recording the data. Figure 1 shows this error. The yield is displayed in boxplots classified by sites. Yields from

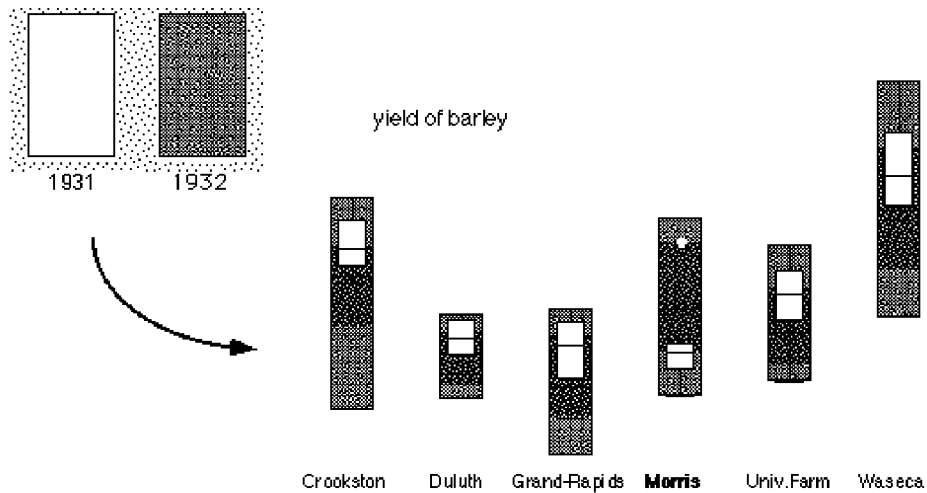


Figure 1: *The yield of barley is displayed in boxplots according to different sites. Highlighted are the values from 1931. Different behaviour of the ‘Morris’ site is obvious.*

the year 1931 are selected. In contrast to the other sites, the yields of Morris were worse than in 1932. This suggests that the years have mistakenly been swapped. Here it would be useful to get an impression of these plots if the results for Morris were switched for the two years. Figure 2 shows a two-step selection that gives this view. In the first step again the data of the year 1931 is selected (as shown in figure 1) but now in a second step the Morris site is selected additionally in XOR-mode in the barchart of sites. The visible highlighting therefore shows the yields from 1932 in the case of Morris, and the yield from 1931 for the other sites, since the second selection swaps the highlighted data of Morris. To indicate which subgroup is selected in which mode black dots mark the selected bar and icons the used mode. Figure 2 is one two-step selection out of $2 \cdot 6 = 12$ possible (given the modes above). To set up all 12 selections one would have to select at least $2 \cdot 12$ bars. In

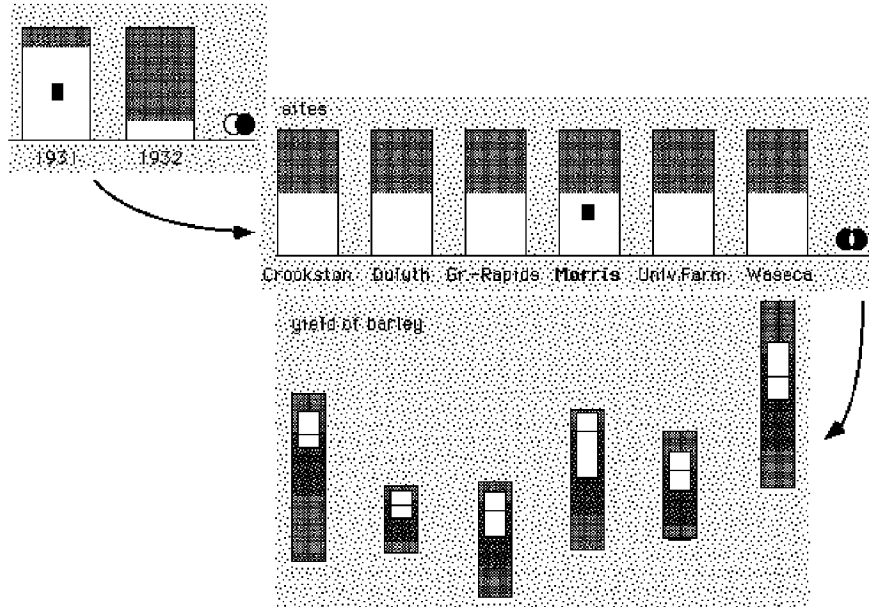


Figure 2: A two-step selection sequence fixes the error from Morris, thus the highlighting in side by side boxplots displays the correct yields.

general a redefinition of k selections in an n step selection ($k < n$) obviously needs a newly selection of $n - k$ unchanged steps besides k necessary ones. There are several possibilities to make working with successive selections more efficient:

- permit UNDO on the last selection
This means, mechanical errors can be corrected, which otherwise would invalidate a whole sequence and make it necessary to start all over again.
- provide a tool for changing the last selection in a sequence
This allows a quick overview of the different combinations in the last selection.

Given that much, why not go further and change selections at earlier stages of the sequence?

Attention therefore focus on ways for changing existing selections. Going back to the example of the barley data, in the setting of figure 2 a changing

between the years would be possible. The correcting selection in XOR mode allows a quick and efficient comparison of the yields.

MANET (Theus, Martin 1996) offers all these advanced selection techniques in a tool called *selection sequence*.

2 Implementation in MANET

A *selection sequence* in MANET is realized by a list of quadruples of the form:

(plot, mode, area, tool)

in which *mode* is one of the logical operators AND, OR, NOT (or an operator derived from these) and *tool* usually is a pointer, a drag-box or a lasso.

The idea with MANET was to make this list fully interactive to provide the desired flexibility.

To meet the user requirements several ways of accessing the tuple's elements have been implemented besides the common handling of lists.

In addition to deleting, adding, inserting and swapping list elements there are ways to reset any of the quadruple's elements like choosing another mode or relocating the area of selection.

2.1 Limiting factors of the implementation

The open access to the sequence and its elements is based on a 1-1 relationship between plot and selection. This makes a multiple selection impossible in one plot because any further selection is interpreted as a resetting of the previous one. Although this does not allow a selection of disjoint areas in a single plot, complex selections in one plot would be a contradiction to an easy redefinition of a selection.

(On a smaller scale the problem of an disjoint highlighting area can be solved by opening another copy of the plot.)

Dealing with more than one sequence at a time is not implemented either. This would mean excluding plots from global linking - something that should be avoided in an interactive system - at least until a sufficient user interface is provided for distinguishing between different sequences.

2.2 Interface

There are several ways of putting the concept of open access to selections into practice:

2.2.1 User performed action

The user explicitly starts a sequence by picking the corresponding tool, setting the right mode or something similar.

Figure 3 and figure 4 show two examples of control panels.

This method turned out to be not very practical as it was often irritatingly

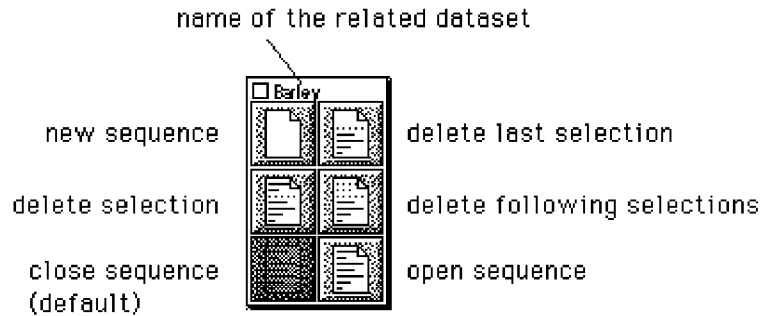


Figure 3: *Script-like first approach of a toolbox for controlling selection sequences*

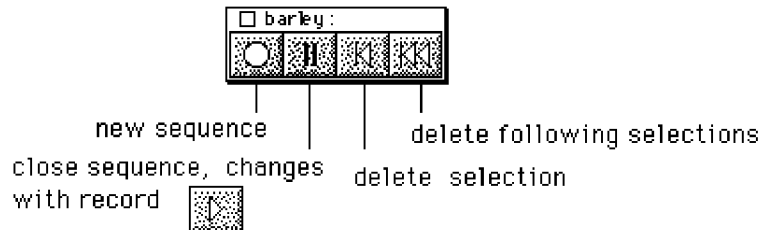


Figure 4: *Toolbox for controlling selection sequences with the appearance and functionality of a movie or cassette player*

obvious that recording should have been started earlier. This is not what one expects of an intelligent and supportive system.

2.2.2 Automatic action

The automatic recognition of a *selection sequence* is based on the fact that every selection carried out in the default mode starts a new sequence. Once the sequence has more than one element an icon appears in the lower right corner of each involved plot (see figure 5). The icon contains information about the order of selection as well as the local selection mode.



Figure 5: *Icon displayed in the lower right corner of each plot in a selection sequence. From left to right icons of default-, XOR-, intersection- and add-mode are shown.*

2.3 Changes in an existing selection

Changing the area and the tool of selection can simply be done by a new selection. Picking up a new selection mode in the toolbox changes the selection mode of the active plot.



The icon on the lower right in a plot has an interactive function as well as an informative one. By clicking on the index number a popup menu appears displaying the functions for handling list elements.(See figure 6)



Figure 6: *Popup menu with commands for working on list elements in a selection sequence*

2.4 Getting further information on the actual sequence

An important principle in developing the MANET software has been to maintain a consistent interface and this also applies to sequences.

- option-click (provides plot specific information, cued in by ) displays the area of selection by inverting it (see figure 7).
- option-shift-click (for further or more general information; visualised by ) shows the corresponding SQL-query in another window (see figure 8)

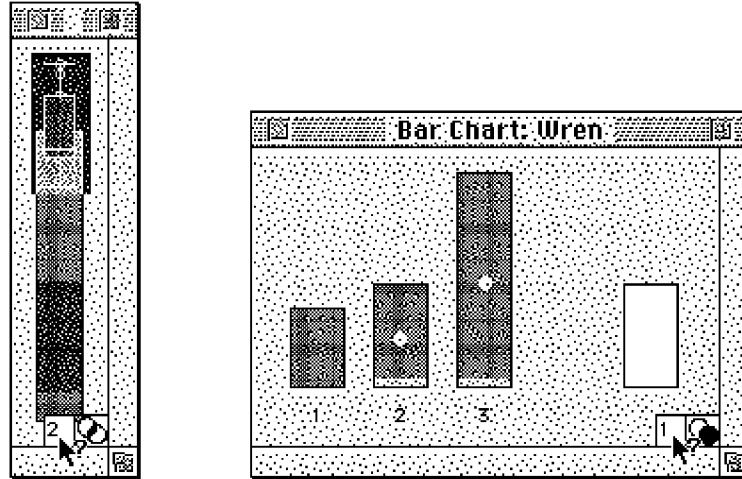


Figure 7: Option-click on the sequence icon displays the selection area in a plot

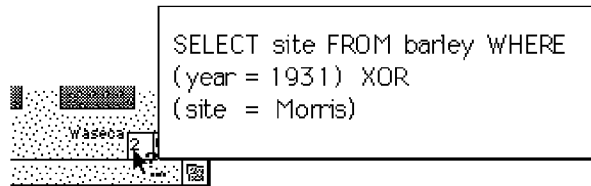


Figure 8: *SQL-query from the barley example above*

3 Applications

3.1 Neighbourhoods

Searching for objects that are similar to one another is often of interest in data analysis.

In figure 9 and 10 a dataset of Augsburg from the time of the Thirty Years War (1616-1646) including socioeconomic and geographic data is shown. Figure 9 shows all 11 "neighbours" of the tax district in which the Fugger family of bankers lived in medieval Augsburg with the neighbour-relation defined by:

- being an outlier in respect to the taxes paid in 1618 (1) and 1646 (2),
- being an outlier in respect to percentage of patricians (3) and merchants (4) living in the district and
- no weaver (5) living in the same district.

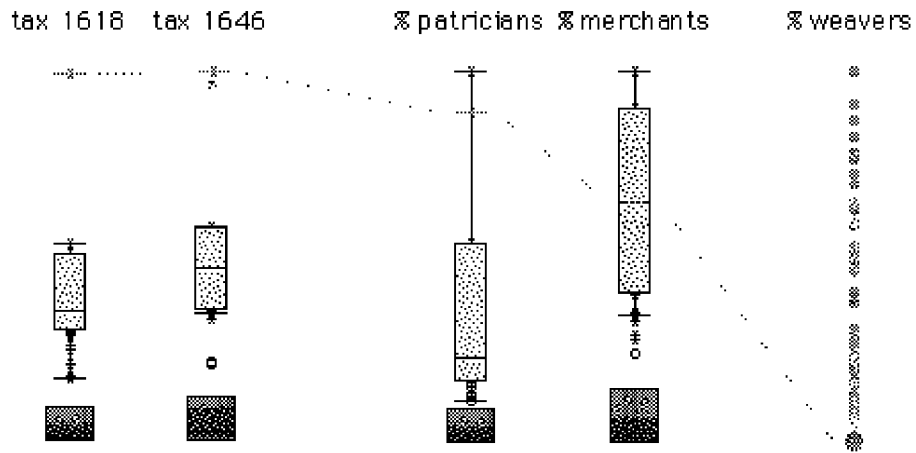


Figure 9: selection sequence *displaying the neighbour-relation defined above. Highlighted are all 'neighbours' to the Fuggers' district. For better comparison the point for the Fuggers' district is marked by a line separately.*

Figure 10 reveals the interesting but not surprising fact that they are spatial neighbours as well, along the important Maximilian Strasse.

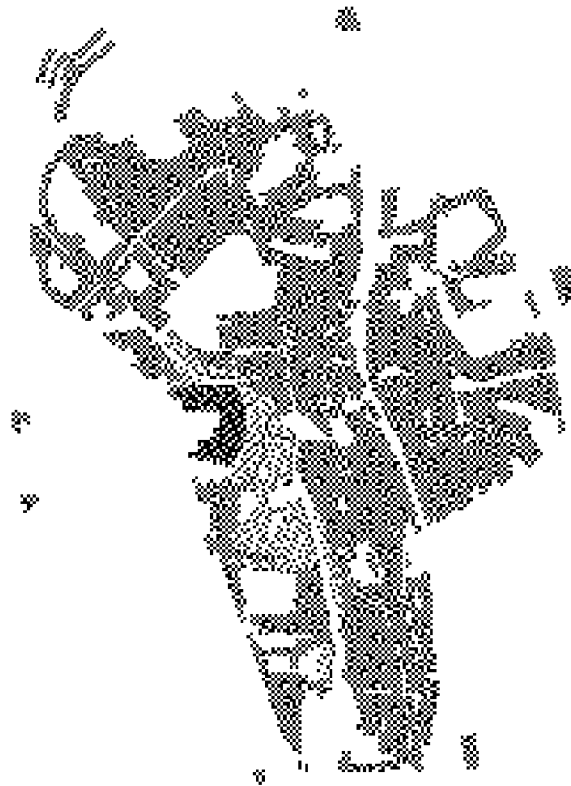


Figure 10: *Map of Augsburg's tax district from the year 1626, highlighted are the neighbours to the Fuggers' district defined above.*

3.2 Comparison of combinations in a result variable

Figure 11 shows the example of the Munich dataset which records the presence of the bronchitis disease depending on dust concentration and exposure time. 'Bronchitis' (0 = no/ 1 = yes) is the result variable; for better comparison of the highlighted proportions the display as a spine plot (Hummel, Jrgen 1995) is chosen. In a spine plot the height of all bars is the same, but the width is proportional to the number of cases falling into this category. Different combinations of 'dust' and 'duration' are selected. It is obvious to see that high dust concentration going along with a high duration of exposure causes a higher bronchitis risk.

Using *selection sequences* these results can be reviewed for any arbitrary interval of exposure time. Since MANET offers not only ststic selections but brushing as well, the bronchitis risk can be monitored while brushing over the whole exposure (or concentration) interval.

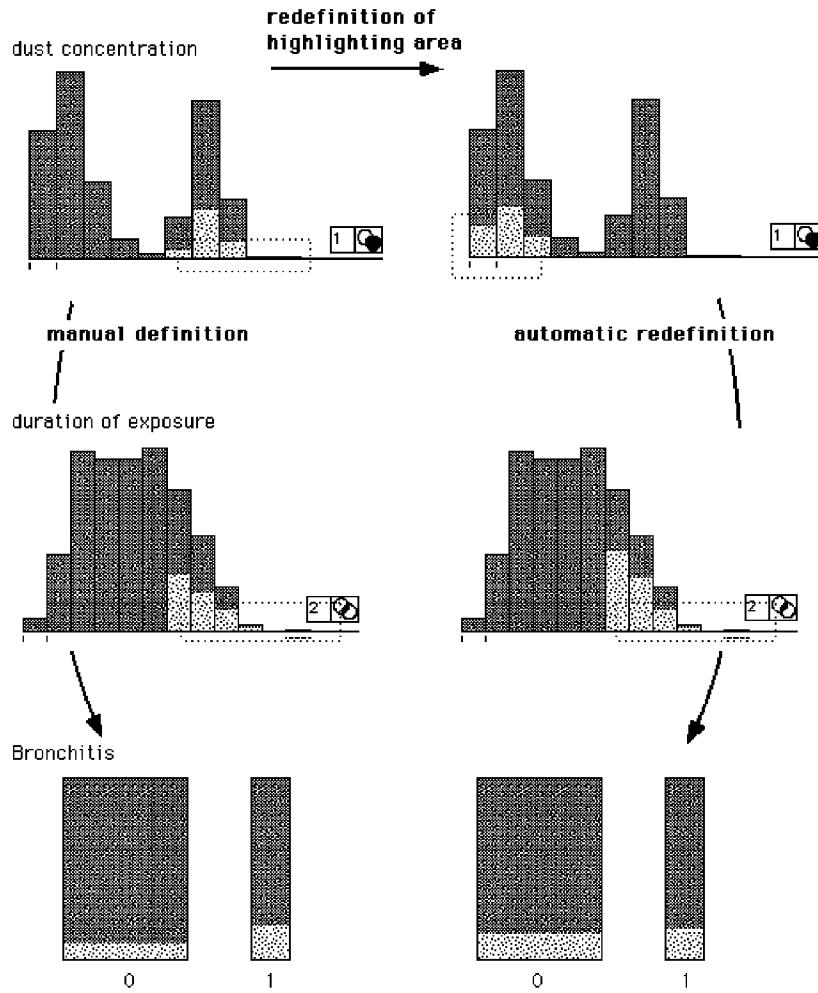


Figure 11: *Editing a selection sequence. On the left we have selected those exposed to a high dust concentration for a long time and linked to whether they have bronchitis or not. On the right we have switched the initial selection from high dust concentration to low, resulting in an update of the highlighting in B and C.*

4 Conclusion

Both trellis displays and selections sequences enable a reduction of dimension by projecting the data onto a subset (at least in case of categorical data). While trellis displays are a more static approach for this, selection sequences offer the full range of interactive methods and provide a very powerful tool for exploring data.

Additionally, *selection sequences* form a starting-point for further developments. Fuzzy highlighting now becomes interesting, that means that not only the selected points are highlighted, which is equivalent to absolute highlighting, but also points are highlighted, that would ‘almost’ have been selected. To distinguish these points from the others different colours or different shading of highlighting will become necessary. Beyond the approach of the programme ‘Visualization Construction Kit’ (Dawkes H., Tweedie L. and Spence R. 1996), which is doing this already in a more restricted way, even a weighting of the fuzzy highlighting should be possible, to offer a possibility for a subjective or data-based measure for ‘almost’ selected points.

References

- Cleveland, William S. (1994) *Visualizing Data* Hobart Press, Summit NJ.
- Dawkes H., Tweedie L. and Spence R. (1996),
VICKI - The Visualisation Construction Kit,
<http://www.ee.ic.ac.uk/research/information/www/huwlrd/vicki/VICKI.html>
- Hummel, J^orgen (1995) Linked Bar Charts: Analysing Categorical Data Geographically, *Computational Statistics*, Vol.11 Issue 1
- Theus, Martin (1996) *Theorie und Anwendung Interaktiver Statistischer Graphik*, Wißner Verlag, Augsburg.
- Theus, Martin (1996), *MANET*
<http://www1.math.uni-augsburg.de/~theus/Manet/MANET-new.html>
- Wilhelm, Adalbert, Unwin Antony R. & Theus, Martin (1996) Software for Interactive Statistical Graphics A Review in *Advances in Statistical Software* 5, *Softstat* 95, ed. Frank Faulbaum, Gustav Fischer Verlag, Stuttgart