Please cite as: Bürgin, R. and G. Ritschard (2014). "A decorated parallel coordinate plot for categorical longitudinal data" *The American Statistician*, **68**(2), 98–103. DOI 10.1080/00031305.2014.887591

A decorated parallel coordinate plot for categorical longitudinal data

Reto Bürgin*[a] and Gilbert Ritschard[a]

[a]Institute for Demographic and Life Course Studies, University of Geneva

January 2, 2014

*Reto Bürgin is LIVES PhD student in statistics and Gilbert Ritschard is professor in statistics, both at the Institute for Demographic and Life Course Studies, University of Geneva, Uni-Mail, 1211 Geneva 4, Switzerland (email: <reto.buergin, gilbert.ritschard>@unige.ch). This publication results from research work executed within the framework of the Swiss National Centre of Competence in Research LIVES (IP14), which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support. They also warmly acknowledge the many constructive comments of associate editor Ronald Christensen and the four anonymous referees.

Abstract

This article proposes a decorated parallel coordinate plot for longitudinal categorical data, featuring a jitter mechanism revealing the diversity of observed longitudinal patterns and allowing the tracking of each individual pattern, variable point and line widths reflecting weighted pattern frequencies, the rendering of simultaneous events, and different filter options for highlighting typical patterns. The proposed visual display has been developed for describing and exploring the order of occurence of events, but it can be equally applied to other types of longitudinal categorical data. Alongside the description of the principle of the plot, we demonstrate the scope of the plot with a real data set.

Key Words: Visualization; Event sequences; State sequences, Longitudinal categorical data; Exploratory data analysis; Sequence analysis; Graphical statistics; Multiple time series plot

1 Introduction

The article introduces an original way of plotting a set of event sequences such as the successions of life events describing professional careers or family trajectories. The plot is intended for identifying the typical order of occurrences of the events in the considered sequences while rendering at the same time the diversity of the observed sequencing patterns. Although the plot can be used for any kind of categorical sequences, it is specifically designed for rendering events that, unlike states for example, do not have durations, can simultaneously occur at a same time point, and whose position in the sequence does not convey other time information than the order of occurrence.

As a first illustration of the proposed plot, Figure 1 renders the sequencings of family life events up to 45 years old. The plotted sequences come from the European Social Survey (2006) and concern 487 Scandinavians born between 1930 and 1939. In the left panel, events are aligned on their order of occurrence in the sequence. Each line represents a unique observed order pattern and the line width reflects the frequency of the pattern. Looking at the thickest line, we learn that the most frequent pattern is to first experience Leaving Home (rank 1), later First Union and First Marriage a same year (rank 2), and later again First Childbirth (rank 3). The lines are jittered to avoid overlapping and, to help identifying typical patterns, only patterns with a minimal support—here 5%—are colored. The diversity of all observed patterns is rendered through the remaining bleached out patterns. To facilitate the tracking of distinct patterns, there are gray arrangement zones at the intersection of the x—rank order of occurrence—and y—event label—coordinates, and the events in an order pattern are represented by solid squares occupying a same position inside the successively crossed arrangement zones. Simultaneous events, i.e., events occurring during a same year of age as 'First Union' and 'First Marriage' in the most frequent pattern, share the same rank and are connected by a vertical line.

The left panel only accounts for the order of the events. Therefore, each pattern may represent people who experience the events in the corresponding order but not necessarily at

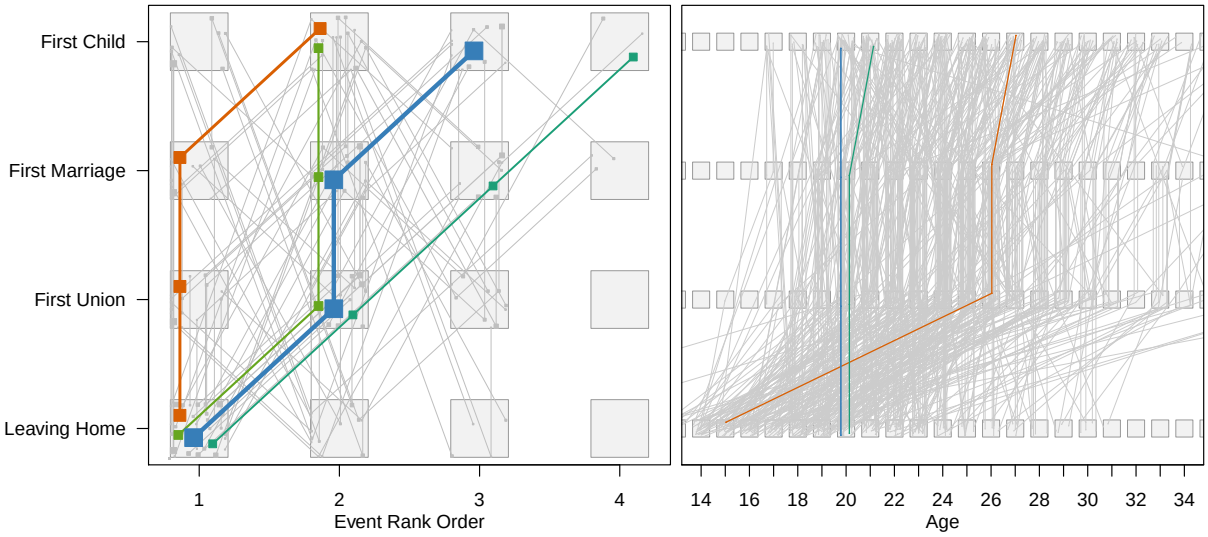

Figure 1: Parallel coordinate plot of Scandinavian family life events of the 1930-39 birth cohort. Left panel, alignment on rank orders of occurrence of the events. Right panel, alignment on event time stamps. Patterns with frequency below the minimum support of 5% in the left panel and 1% in the right panel are grayed out.

the same ages. As shown in the right panel of Figure 1, accounting for the timing information dramatically increases the pattern diversity and makes it more difficult to identify typical patterns. The three highlighted patterns are the only ones reaching a 1% support.

The proposed graphical method has been developed to respond three main objectives: (i) identification of standard patterns with possible simultaneous events; (ii) ability to render the whole diversity of the observed patterns; and (iii) suitability for group comparisons.

The literature proposes several methods for rendering categorical sequences. Bar, mosaic or association plots (Hartigan and Kleiner; 1984; Friendly; 2000) are helpful to render distributions of categorical data and highlight the association between pairs of categorical variables. By cross tabulating event occurrences with their rank order, such plots can visualize how events are distributed at and across the successive positions. Alternatively, by considering the event occurring at each successive position as a categorical variable, a set of sequences can be seen as a series of categorical variables and the successions of events at the successive positions rendered by means of parallel coordinate plots. Examples of categorical parallel coordinate plots are hammock plots (Schonlau; 2003) and parallel sets (Kosara et

al.; 2006), and an explicit example of parallel coordinate for sequential pattern can be found in Yang (2003). The plot consists in reporting the position in the sequence (or time point) on the x-axis and assigning a vertical coordinate to each event-category. Each unique sequence pattern is then visualized as a polyline connecting the successive events in the order they appear in the sequence. Varying line widths can be used to visualize the support of each event-to-event segment. Among plots specifically designed for sequence data, there are various plots for state sequences (Brzinsky-Fay et al.; 2006; Gabadinho et al.; 2011) These plots essentially render the duration of the states and do not apply for sequences made of elements such as events that do not have durations. Alongside the already mentioned parallel coordinate plot, there are two further types of graphics that can potentially be applied to any kind of categorical sequences including event sequences. Graphics of the first type, known as *life lines* or *calendar plots*, arrange color-coded event symbols along horizontal lines (Wang et al.; 2010; Wongsuphasawat et al.; 2011). The second type of plots are *directed graphs* (Hébrail and Cadalen; 2000; Huzurbazar; 2004), such as the graphical representation of a flowgraph, that connect event nodes with directed line segments along the event order.

The decorated parallel coordinate plot proposed in this article extends the parallel coordinate principle with the following main features: (i) algorithmically controlled *jittering*; (ii) possibility to merge *embeddable* sequences; and (iii) filter instruments and criteria to improve the exploratory power of the plot. The plot can also render weighted frequencies of the sequence patterns and cases experiencing no event.

The article is organized as follows. In the upcoming section, we provide additional details on the algorithmically controlled jittering and discuss options for improving plot readability. Subsequently, we extend the family life event example to illustrate the plot capacities including its suitability for comparison purposes. Finally, we address practical issues regarding the plot usage, its scope and limits and conclude by summarizing our findings.

2 Jittering, embedding and filtering mechanisms

The basic principle of the proposed plot has been explained in the introduction. This section gives additional details regarding the jittering arrangement, embeddable sequence patterns and filtering criteria.

Jittering arrangement The jittering arrangement is defined within the light gray rectangular *arrangement zone* replicated at each grid point. A distinct location in this zone is assigned to each sequence pattern. For example, the thickest line in the left panel of Figure 1 goes through solid squares located at the bottom-center in each crossed arrangement zone. The placing procedure first assigns a solid square of size proportional to the (possibly weighted) sample frequency to each order pattern. Next, a random location is successively assigned to the squares. Location is allocated in decreasing order of the size of the squares and so that squares do not touch each other. In case the remaining space is insufficient, the size of all solid squares are proportionally reduced to make them all fit in the zone. The plot is then finalized by drawing connecting lines between the successive squares belonging to a same patterns. The widths of the line segments are adjusted to the pattern frequency but are slightly thinner than the event-squares for readability. Simultaneous events appear as vertical segments. To maintain the line-continuity in these cases, we connect the precedent event with the lowest event of the vertical segment and the subsequent event with the highest one (or optionally conversely). In the exceptional case where a same event would occur several times at a same position the multiple occurrences would be reflected by a 'sunflower' inscribed in the concerned square. Finally, *zero-event sequences*, i.e., empty sequences corresponding to cases that do not experience any event, are reflected by a square outside the bottom-left arrangement zone.

Full-scaled real data sets will most often include a great number of distinct patterns and additional tricks may be necessary to distinguish patterns of interest in the plot. We propose two such adjustments.

Emphasizing interesting patterns The first option is to bleach out less interesting patterns and lay them in the background. The level of interest will typically be measured by the frequency of the pattern, but could as well be, for example, the inverse frequency if we are interested in atypical patterns, or some measure of the strength of association between the pattern and a target variable such as the sex, birth year or income of the concerned individuals. In Figure 1 patterns with support of 5% or higher are colored and all others are bleached out. Instead of the minimal support, we can also chose to highlight the minimum number of patterns such that their cumulated frequency reaches a given threshold. The latter would however only make sense for summable interest measures, and would not make sense for example for association measures.

Plotting only non-embeddable sequence patterns The second option allows to reduce the number of plotted lines without losing information and consists in drawing only *non-embeddable* sequence patterns. A sequence pattern S_1 is *embeddable* into a pattern S_2 if S_2 can be transformed into the exact form of S_1 by cutting an ending—or starting—substring from the sequence S_2. The *non-embeddable* patterns are those unique event order patterns which cannot be embedded into any other one.

The embedding is visualized by adjusting the line widths of shared partial line segments. For instance, in the left panel of Figure 3 where non-embeddable sequencing patterns are plotted, we observe that the ending segment and square of the thick ascending diagonal line are slightly thinner than those at the start of the line, meaning that the line also represents shorter embedded patterns. Compared with the right panel in Figure 2 that plots the same data, we see that embedding shorter patterns in longer ones permits to reduce the number of drawn lines from 55 to 30.

The embedding trick raises two difficulties: first, the trick implies a technical ambiguity. Short event order patterns can often be embedded into more than one *non-embeddable* event order candidates. We suggest in that case to embed the patterns into the most frequent

pattern among the available candidates. Doing so, instead of distributing them evenly over all candidates for example, will emphasize the commonness of the shared segments. Second, the interpretation becomes ambiguous when two or more event orders with both different start and end positions are embedded in the same non-embeddable event order pattern. For example, the three sequences A-B-B-*, *-B-B-C and A-B-B-C, where a '*' indicates an empty position, can be merged into the single non-embeddable sequence A-B-B-C with a weight of 2 for the paths A-B and B-C, and a weight of 3 for the path B-B. The same non-embeddable sequence results from the three sequences A-B-B-C, A-B-B-C and *-B-B-* and it is thus not possible to univocally retrieve the original sequences from the non-embedded sequence; hence the ambiguity. We recommend to use the embedding adjustment only with either left-aligned or right-aligned sequences.

Combining both adjustments Both tricks above can be applied together on a same plot. In that case, when one or more patterns have been embedded in a longer one, the whole non-embeddable event order pattern is highlighted whenever its most frequent segment fulfills the highlighting condition. As a consequence, some non-embeddable patterns which do not themselves reach the minimum interest level may be highlighted just because some other patterns were embedded in them.

3 An application: Family life event histories

In order to illustrate the practical scope of the proposed plot and especially its suitability for group comparison, we consider again the 487 Scandinavian family life trajectories of the 1930-39 birth cohort rendered in Figure 1 and compare them with the 885 trajectories collected for the 1950-59 cohort. All data come from the 2006 European Social Survey Round 3.

When analyzing life events, a question of interest is whether typical sequencing patterns change or remain the same across age groups and an answer to this question is obtained by plotting side-side the trajectories of the different groups as in Figure 2. To facilitate

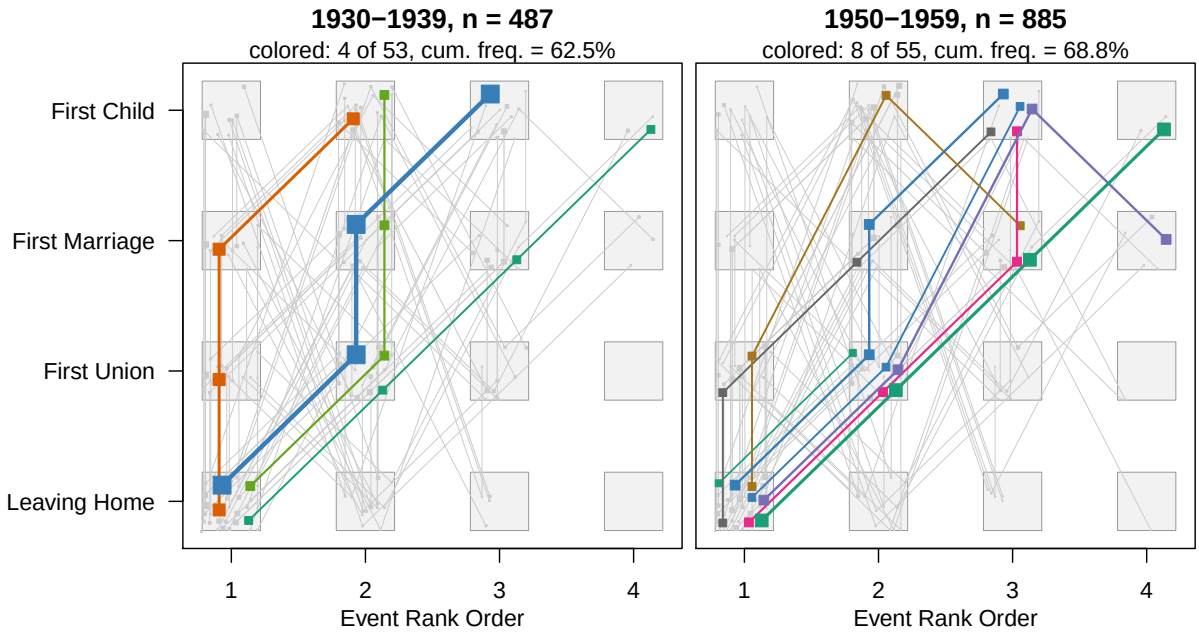

Figure 2: Cohort comparison of Scandinavian family life event orders. Highlighted lines describe order patterns with weighted frequency above 5%. No embedding.

the comparison, the same highlighting color and location in the arrangement zone are used in each of the groups when a pattern is present in several groups. For example, the pattern (Leaving Home) – (First Union, First Marriage) – (First Child) is displayed in blue and jittered up-left in both panels[1] of Figure 2.

The two plots in Figure 2 widely differ. The number of highlighted event orders with at least 5% support increases from four to eight and there are only two common highlighted patterns. The most typical pattern for cohort 1930-39 becomes much less frequent for the 1950-59 cohort where the most frequent pattern is the diagonal line, i.e., (Leaving Home) – (First Union) – (First Marriage) – (First Child). The cohort 1950-59 appears to be much less standardized. Even though the number of frequent patterns increases from four to eight, the cumulative frequency of the frequent patterns decreases from 62.5% to 45.8%. For the youngest cohort, there also are frequent patterns with 'First Child' without marriage or before 'First Marriage'. In summary, the plot clearly exhibits how norms in the organization

[1]Due to the random factor in location and color assignments, the location and colors in Figure 2 differ from those in Figure 1.

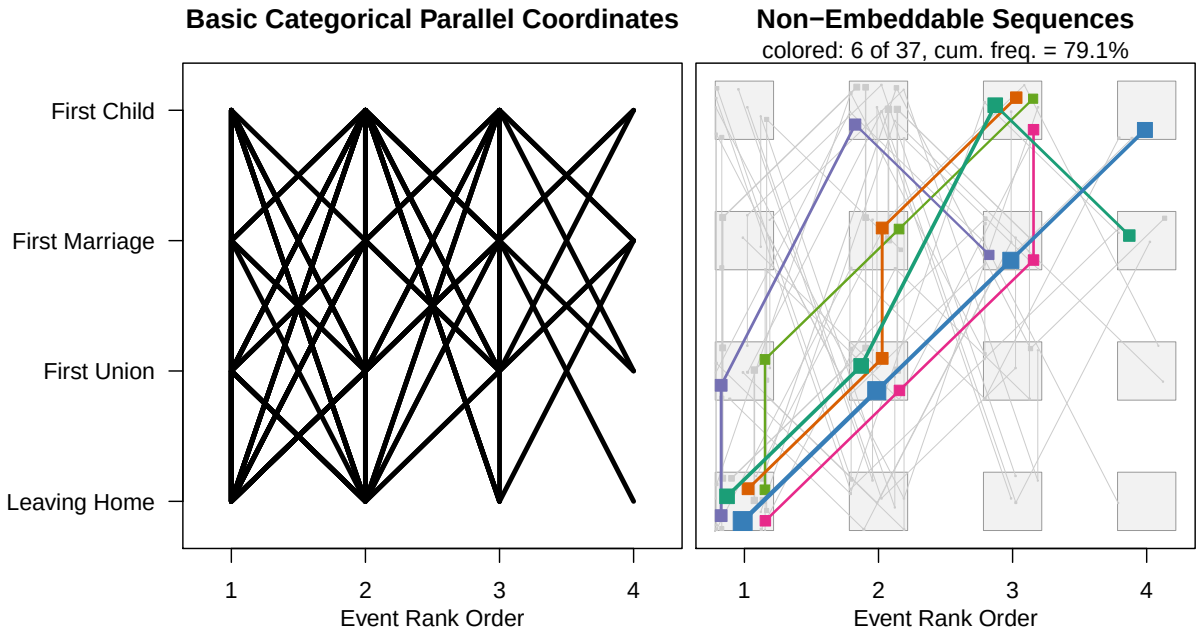

Figure 3: Alternative plots of the 1950-59 cohort. Left panel: basic parallel coordinate plot. Right panel: non-embeddable event order patterns.

of life trajectories changed across cohorts.

The superiority of our proposition over the basic parallel coordinate plot appears clearly when comparing the basic plot for the 1950-59 cohort shown in the left panel of Figure 3 with the plot in the right panel of Figure 2. In the basic plot, the plotted lines overlap, which makes it impossible to track single patterns. Even worse, basic parallel coordinates could be misleading regarding patterns actually not observed. For example, the pattern (First Union, First Child) − (Leaving Home, First Marriage) is not present in the data set while the plotted line segments may suggest it is. This problem does not occur with our proposition because, as can be seen in Figure 2, the distinct sequence patterns are jittered and can be univocally tracked by following up the corresponding event-squares similarly located in the arrangement zones.

The plot for the Scandinavian 1950-59 cohort can be slightly simplified with the embedding trick. The resulting plot is shown in the right panel of Figure 3. In that plot, the pattern (Leaving Home)−(First Union), for example, has been embedded into the pattern (Leaving

Home)−(First Union)−(First Child)−(First Marriage), and both patterns are visualized by the same single line. The method reduces the total number of lines from 55 to 37 and the number of highlighted patterns from 8 to 6. Due to these changes, the square points within the gray zones have been rearranged, the widths of the event-squares and line segments adjusted, and colors newly reassigned. All these characteristics are therefore different from those in Figure 2.

4 About the plot usage

The plot has been implemented in the TraMineR R package (Gabadinho et al.; 2011). The `seqpcplot` function producing the plot offers a series of arguments for controlling, among others, the widths of the square-points and lines as well as their coloring, the filtering thresholds and position versus time alignment. The complete list of arguments is documented in the online help file of the `seqpcplot` function where the user also finds several examples.

The coordinate assignment for the event categories is basically arbitrary and could be for instance the alphabetical order. The readability of the solution will, however, most often depend on this coordinate order and could be improved by a suitable ordering. A meaningful solution is for example to arrange the event categories in their most frequently observed order of occurrences as in Section 3.

The default representation is obtained by aligning the successive elements in the sequences on their rank order of occurrence. A possible alternative is to align the states/events on their time of occurrence. By using time alignment we can render transition times. Practically, however, when the number of time positions increases the resulting graphic may become very cluttered because of the variability in the timing of similarly sequenced events. The right panel in Figure 1, for example, gives the time aligned representation of the Scandinavian family life event sequences of cohort 1930-39. The time-aligned plot exhibits a high diversity—essentially a timing diversity—of the trajectories which contrasts with the rela-

tively low sequencing diversity shown in the left panel. We learn from the time-aligned plot that leaving home starts at about 14 years old, and that events First Union, First Marriage and First Child occur since age 17 but become much more frequent after 20 years old. Nevertheless, the plot looks cluttered and other plots such as survival curves or life and calendar lines (Wang et al.; 2010; Wongsuphasawat et al.; 2011) could be more appropriate for rendering the timing. By transforming event sequences into state sequences—as explained in Ritschard et al. (2009) for example—we could also resort to plots for state sequences (Gabadinho et al.; 2011) that explicitly render timing and durations.

Although there are no technical limitations to the scalability of the plot, increasing the number and/or length of the sequences or the alphabet size may impair the plot interest. The limitation is not that of the total number of sequences but that of the number of unique sequences. The number of unique sequences is intimately linked with the sequence length and the size of the alphabet, i.e., the number of distinct events or states. The larger the alphabet, the less chances we have to find out a significant proportion of sequences sharing a common pattern. The same is true for the sequence length: the longer the sequence, the lower the chances of two sequences following a common pattern. The solution to find out regularities in case of a large alphabet would be to merge close elements of the alphabet. In case of long sequences, the solution could be to use a rougher time granularity which would transform the different sequencings of events occurring in a given laps of time into a unique set of simultaneous events. To give an order of magnitude, the alphabet should not exceed about 10. Likewise, the plot may become hard to read when sequences contain more than 10 distinct successive elements. With shorter sequences we could afford a larger alphabet and reciprocally with a small alphabet we could afford longer sequences.

5 Conclusion

The decorated parallel coordinate plot proposed in this article and provided by the TraMineR R package (Gabadinho et al.; 2011) is a powerful tool for exploring how elements are typically ordered in a set of sequences. The filtering mechanisms that dim out less interesting patterns together with the embedding trick, permit to clearly highlight the most frequent patterns while still rendering the entire diversity of the observed patterns. Moreover, replicated arrangement zones facilitate the tracking of individual jittered patterns. Although the plot is primarily designed for event sequences where only the rank order of occurrence of the events matters, the plot can also render time aligned events and be used with other types of categorical longitudinal data such as categorical panel data for example.

References

Brzinsky-Fay, C., Kohler, U. and Luniak, M. (2006). Sequence Analysis with Stata, *The Stata Journal* **6**(4): 435–460.

European Social Survey (2006). ESS Round 3, Data File Edition *3.4*, Norwegian Social Science Data Services, Norway – Data Archive and Distributor of ESS Data.

Friendly, M. (2000). *Visualizing Categorical Data*, SAS Institute, Cary, USA.

Gabadinho, A., Ritschard, G., Müller, N. S. and Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR, *Journal of Statistical Software* **40**(4): 1–37.

Hartigan, J. A. and Kleiner, B. (1984). A Mosaic of Television Ratings, *The American Statistician* **38**(1): 32–35.

Hébrail, G. and Cadalen, H. (2000). Visualisation et Classification Automatique de Parcours Professionnels, *Actes des XXXIIe Journées de statistique, Fès, Maroc*, pp. 458–462.

Huzurbazar, A. V. (2004). *Flowgraph Models for Multistate Time-to-Event Data*, John Wiley & Sons, New Jersey, USA.

Kosara, R., Bendix, F. and Hauser, H. (2006). Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data, *IEEE Transactions on Visualization and Computer Graphics* **12**(4): 558 –568.

Ritschard, G., Gabadinho, A., Studer, M. and Müller, N. S. (2009). Converting Between Various Sequence Representations, *in* Z. Ras and A. Dardzinska (eds), *Advances in Data Management*, Vol. 223 of *Studies in Computational Intelligence*, Springer-Verlag, Berlin, Germany, pp. 155–175.

Schonlau, M. (2003). Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots, *Proceedings of the Section on Statistical Graphics, American Statistical Association; 2003, CD-ROM.*

Wang, T. D., Plaisant, C. and Shneiderman, B. (2010). Temporal Pattern Discovery Using Lifelines2, *IEEE VisWeek 2010*, Salt Lake City, USA.

Wongsuphasawat, K., Gómez, J. A. G., Plaisant, C., Wang, T. D., Taieb-Maimon, M. and Shneiderman, B. (2011). LifeFlow: Visualizing an Overview of Event Sequences, *Proceedings of the 2011 annual conference on Human Factors in Computing Systems (CHI), Vancouver, Canada, May 7-12, 2011*, ACM, New York, pp. 1747–1756.

Yang, L. (2003). Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates, *in* V. Kumar, M. Gavrilova, C. Tan and P. L'Ecuyer (eds), *Computational Science and Its Applications - ICCSA 2003*, Vol. 2668 of *LNCS*, Springer-Verlag, Berlin, Germany, pp. 21–30.