SpringView: Cooperation of Radviz and Parallel Coordinates for View Optimization and Clutter Reduction

Enrico Bertini, Luigi Dell'Aquila, Giuseppe Santucci
Dipartimento di Informatica e Sistemistica - Università di Roma "La Sapienza"
Via Salaria, 113 - 00198 Roma, Italy - {bertini, dellaq, santucci}@dis.uniroma1.it

Abstract

In this paper we integrate radviz and parallel coordinates, two methods able to handle multidimensional datasets, exploiting their contrasting characteristics. From on side radviz offers good direct data manipulation (i.e., brushing) techniques and low cluttering but it fails in providing visualization of quantitative information; conversely, parallel coordinates clearly shows the values of data attributes and their ranges but suffers from high cluttering also on small datasets and presents tedious manipulation techniques. We developed a prototype, called SpringView, that allows for simultaneously viewing both radviz and parallel coordinates and implements several useful techniques to manipulate the data, both interactively and, more interestingly, automatically. We challenged our approach against two well know multidimensional datasets, proving its effectiveness.

1 Introduction

Visualizing a data set containing large amounts of data and/or attributes likely produces a cluttered image. The user suffers from a strong sense of mess that rises from both the intrinsic limits of visual devices and adopted techniques. As the number of items increases, almost any kind of visual technique fails to convey detailed information; a lot of graphical elements overlap, losing useful pieces of information. However, clutter can obscure the structure present in the data even in small dataset, making it hard for the user to find patterns and reveal relationships.

In this paper we focus on parallel coordinates, a very useful and widely adopted visualization technique, risen during 1980's and allowing for dealing with several multidimensional analysis problems. This method is able to handle data sets with a large number of attributes, where each dimension corresponds to an axis. The axes are arranged as uniformly spaced vertical or horizontal lines. Each data element is represented through a polyline, crossing the axes according to its dimensions' values. Unfortunately, as the number of displayed elements or dimensions increase the overlapping of lines produces a non readable image.

Several proposals have been presented to reduce the cluttering, addressing this issue either reordering the involved dimension or reducing the number of of displayed elements, through clustering or brushing techniques. The latter approaches require the user to interact with the system, interaction that is time consuming and boring. As an example, assume that the user wants to zoom in on a subset of the displayed polylines; in such a case s/he is required to specify in some way N range values, one for each dimension. Moreover, some clustering techniques are based on the presence of hierarchies among the data, not always available.

In this paper we attack the problem using a different approach: we present the user with two different visualizations, the parallel coordinates and the radviz representation, a 2D visualization in which data elements are drawn on a normalized circle that presents on its circumference the data dimensions uniformly spaced. The position of a data point depends on the relationships among its attributes values: a data element presenting a high value on dimension d_i and low values for the other dimensions is drawn very far from the center, close to the d_i dimension, while a data element showing for all the dimension the same percentage values (e.g., the values are all 20% of the dimensions's ranges) is drawn exactly in the center.

The presence of a double visualization provide a novel way for clustering and brushing the data. As a first advantage, the user can quickly select 2D zones on the radviz representation getting the corresponding elements highlighted in the parallel coordinates view. Moreover, the user can exploit two automatic techniques that allow for reducing the parallel coordinates cluttering. The first one, that does not alter the number of displayed elements, is based on a color coding automatically computed on the radviz view; such a coding is exported on the parallel coordinate view, producing a clear vision of similar data elements together with their actual values. A second technique, allows for automatically clustering the parallel coordinates polylines, exploiting their similarity and non their distance.

In order to show the potentials of our approach we developed a prototype, called SpringView, that implements these ideas. The paper presents some case studies with two well known multidimensional data sets: “cars”(a 7-dimensional set of data about cars) and “out5d” (a set of observations on a region of Western Australia.), providing indications on the effectiveness of our method.

Summarizing the contribution of the paper is the following:

- it presents a novel system based on the visual cooperation of two *multidimensional* methods that is, to the best of our knowledge, a quite uncommon approach. Moreover one of the two methods (i.e., radviz) plays a double role: it can be used per se or to drive the parallel coordinates view: user operations on radviz are used as input to modify the parallel coordinates view;

- it presents a simplified interactive method to deal with multidimensional datasets;

- it provides two novel automatic interaction mechanisms that allow for showing similarities/clustering on the data belonging to the parallel coordinates view exploiting their arrangement on the radviz view.

The paper is structured as follows: Section 2 analyzes related works, Section 3 details our approach, Section 4 presents two case studies, and, finally, Section 5 presents some conclusions and future work.

2 Related Work

Multiple and coordinates views in Information Visualization have always been considered a powerful method to interact with data visualization to explore and discover correlations, query the data from multiple points of view and, strengthen the power of single visualizations allowing for synergic coordination and comparison of different views. Only recently, however, multiple views have been analyzed and employed more systematically. There is a need in fact to understand what are their advantages and disadvantages, when they represent an appropriate solution, and what are the attributes that make their employment useful and effective.

Baldonado et al. provide a set of guidelines on when and how multiple views should be used [11]. The authors argue that multiple views must be used with parsimony, considering carefully the cost-benefit tradeoffs. Three main requirements should be satisfied in order to justify their employment: the views must show different aspects of the underlying data, they must produce correlations and/or disparities, the problem should lend itself to the decomposition into smaller parts. Similarly, Roberts provides in [9] some additional guidelines. In order to encourage their use, the systems should group similar information/tasks together, provide automatic generation of view from data, maintain consistency between different visualizations, and support correlation between the different views.

SpringView conforms to these requirements. In particular, while the radviz visualization lends itself to the detection of aggregates, parallel coordinates permit to explore finer details inside subsets, thus offering different data views. Moreover, parallel coordinates permit to reveal trends that radviz cannot easily detect, thus they form a synergy to detect data disparities. The fact that subsets of data selected in the radviz can be visualized into parallel coordinated also permits to decompose the problem of the detection of potentially interesting trends into two phases, one performed at a lower grain in radviz, another with finer details in parallel coordinates. Finally, SpringView uses some coupled coloring strategies (details in Section 5.1) that permit to detect correlations offering a natural way to correlate the results across multiple views.

Parallel coordinated have been extensively used as a method to visualize multi-dimensional data [7]. The basic idea is to map the attributes of a n-dimensional dataset to vertical axes so that each data point is represented by a polyline that crosses each axe at the corresponding attribute value. Parallel coordinates permit to visualize in a compact format datasets on the order of thousands of items and ten different data attributes, allowing for detection of correlations, fast perception of aggregations and data distributions among multiple dimensions.

Radviz [6] is a radial spring-based visualization, similar in spirit to parallel coordinates that permits the visualization of n-dimensional datasets. Data attributes are arranged around a circle and equally spaced, each data item is virtually connected to a spring that starts for the circle perimeter and ends on the data item, and each pulls the item with a force proportional to the item’s attribute value. The position of any data item is where the sum of the spring forces is equal to zero. The interesting features is that it places data items with similar values close together so that data aggregations can be discovered.

Multiple views have been successfully employed together with parallel coordinates. The inherent complexity of parallel coordinated and their limited interaction capabilities make the use of additional coordinated views appropriate. As an example, the selection of subsets of data in a coordinated view, permits to easily highlight interesting subsets in the parallel coordinates. In [10] an interesting combination of parallel coordinates and reorderable matrix is presented. The author comments on the benefits of coupling two different n-dimensional visualizations showing that they can be complementary and thus play a synergic role. While, for instance, in parallel coordinated decomposition is naturally applied (e.g., brushing, axes manipulation), reorderable matrix does not offer any easy decomposition mechanisms. At the same time, reorderable matrix allows for fast and easy ordering of data items according to some attribute; the same cannot easily be done with parallel coordinates. Andrienko et al. offer in [1] and interesting overview of multiple ways to look at multidimensional data, discussing on the constant tradeoff between detailed representations, where each data item is mapped to one visual mark, and higher level visualizations (i.e., aggregates). They present a modified version of the parallel coordinate visualization that provides in a single view both detailed and aggregate data, allowing for the detection of interesting subsets and their selection for further analysis.

The problem of clutter, especially in parallel coordinates, has recently drawn major attention. All the visualizations that try to visualize large datasets with details, that map each data item to a single visual mark, and that permit data to overlap, suffer from the same problem of cluttering. For parallel coordinates a variety of solutions have been proposed. Hierarchical parallel coordinates is a technique that uses hierarchical clustering as a way to reduce the overall density. A hierarchical clustering algorithm is previously applied to data and each cluster is mapped to a single visual mark with a surrounding halo which depicts its extension. Then, using a custom interactive tool, the user can drill down to one or more clusters selectively and explore its content in detail [4][5]. In [3] the authors propose a density based technique to map data density to the intensity of parallel coordinates poly-lines. This, together with the constraint that more intense lines cannot be drawn under less intense ones, creates an effective visualization where n-dimensional clusters can be easily detected. Another interesting approach is the use of dimension reordering [2] to decrease clutter in multidimensional visualizations. In [8] the authors take into account a set of multi-dimensional visual techniques (e.g., parallel coordinates, scatter plot matrices, star glyphs) and for each one they define a clutter measure. The idea is to find, looking in the space of possible reordering combinations, the ones that produce the less cluttered images. In SpringView we explore the idea of coping with clutter by means of multiple views using both interactive and automatic techniques. Similarly to the proposals described above, we employ some clustering mechanism to reduce visual density, but in our case the clustering method is not based on any preliminary pure algorithmic calculations. We fully exploit the tight coupling of the two views to aggregate data, with the assumption that close data items in radviz already provides a way to form coherent subsets.

3 SpringView

Our intention is to integrate radviz and parallel coordinates views exploiting their contrasting characteristics. Radviz representation is obtained as follows: we arrange in circle N anchors, one per each data dimension and we suppose each data point attracted by the i_{th} anchor by a force proportional to distance from the anchor and to value of the i_{th} dimension; the balance of forces determines the final position of each data point. In parallel coordinates we have N parallel axes, one per dimension; each data item is represented as a polyline intersecting axes at values of dimensions. Both the systems are N-dimensional and very scalable on number of dimensions and show complementary advantages. From one side radviz offers good direct data manipulation (i.e., brushing) techniques and low cluttering but it fails in providing visualization of quantitative information; conversely, parallel coordinates clearly shows the values of data attributes and their ranges but suffers from high cluttering also on small datasets and presents tedious manipulation techniques. To show the advantages of integrating these representations we developed a small application, called SpringView, that implements the following functions:

- simultaneous representation of radviz and parallel coordinates;
- brushing on radviz through mouse click and drag and exporting the selection on the parallel coordinates;
- data coloring based on radviz, applied to parallel coordinates to facilitate pattern detection and brushing;
- data clustering based on radviz and representation of clusters in parallel coordinates to reduce clutter.

In order to demonstrate the potentials of our approach we will use two well known data sets: “cars” (a 7-dimensional set of data about cars; variables are: miles per gallon, number of cylinders, horsepower, vehicle weight, time to accelerate from 0 to 60 mph (sec.), model year, origin of car) and “out5d” (a set of observations on a region of Western Australia; attributes are: Spot, magnetics and three bands of radiometrics - potassium, thorium and uranium). These datasets are quite simple and low dimensional, allowing us to describe functionality of the system in a clear way. Anyway SpringView shows good results on larger and higher dimensional datasets as well.

4 Radviz and parallel coordinates: a case study

After loading the “cars” dataset (figure 1) the system shows the classic, apparently uncorrelated, radviz and parallel coordinates representations. The user can only detect patterns on radviz, manually brush them and see their exhaustive description on parallel coordinates graphics. Let us assume we want to select data laying near the “horsepower” dimension on radviz (we suppose that these data points are there because of strong attraction from the “horsepower” anchor, so they should have high value for that dimension), to do this we just have to drag the mouse on radviz and the system will show the results of brushing on both representations. Figure 2 shows the result of this operation: we isolated a cluster of cars with high fuel consumption (low MPG) and high number of cylinders and we can see that they are all produced in USA.

4.1 Coloring and brushing on parallel coordinates

On the previous representation users are able to get quick projections from radviz to parallel coordinates. To make the selections on parallel coordinates we use an automatic approach that is different from brushing directly on parallel coordinates: supposing we want to brush data items from USA with low value of MPG, high number of cylinders etc.; with a classic parallel coordinates based approach we should have to make N different range selections, one per dimension, directly on dimension axes or on a slider; to simplify this operation our system creates an association of colors between points in radviz and polylines on parallel coordinates (in figure 3, previously described data are orange polylines on parallel coordinates and correspond to orange points on radviz), allowing the user to use radviz as a brushing tool for parallel coordinates. The system associates a 2D color-map to a rectangular board (in our examples we will use a simple RGB map with increasing values of blue from left to right and increasing values of red from top to bottom), then the radviz algorithm disposes data points on that board, associating to each data point the color of the corresponding point of the board. This coloring is then transferred on parallel coordinates representation (figure 3). Users who want to make brushing on a section of parallel coordinates just have to make an association between colored patterns on parallel coordinates and colored points on radviz, performing brushing directly on that. This approach reduces the effort of brushing at $1/N$ (one mouse click to select one pixel, or click-drag operation to select a region on radviz versus N operations on parallel coordinates). Moreover, many applications that implement brushing based on parallel coordinates allow only to select adjacent polylines; with our approach we can brush different and disjointed clusters and isolate them (e.g., we can compare the previously selected cluster with a new one composed by cars with high value of mpg and low value of horsepower etc, represented respectively as orange and blue points in figure 4 and labelled on radviz respectively with ’a’ and ’b’). In this step, the real connection between parallel coordinates and radviz is the human perceptual system that allows to find patterns in data and associate polylines on parallel coordinates to points on radviz comparing their color. In most cases polylines with similar colors lay near one another, giving the visual perception of a cluster, as is possible to see comparing figures 1 and 3. This can help users to follow polylines and detect patterns in data: different structures of data appear as *automatically brushed* in different colors. After detecting the structures to select, a user just have to locate the same color on radviz and complete brushing (if needed) adding or deleting some points in the neighborhood.

4.2 Clustering for clutter reduction

Sometimes simple coloring is not sufficient to reduce clutter on parallel coordinates view. Figure 5 shows the Out5d dataset: on radviz we can clearly detect some clusters, but parallel coordinates do not give enough information about some of them (i.e., the pattern of green points visible in the left upper part of radviz) because of cluttering. To face this problem our system tries to group data items into clusters based on the radviz representation. We use a simple kind of clustering which is computed splitting the radviz representation on a grid whose dimension can be set by the user, and grouping together the data that reside in each cell. The color of each cluster is computed considering that data elements in a cluster are neighbors on radviz, so we assign to each cluster the mean color of the data elements that fall inside it. Each polyline represents one cluster and the attribute values are obtained using the mean of each data attribute. To convey the information about cluster dimension, we enhance the coloring strategy by adding the alpha channel to the color of polylines: higher for big clusters, lower for smaller ones. In this way we can easily distinguish small clusters (more transparent) from larger ones. Comparing figures 5 and 6 it is evident the reduction of clutter: figure 5 shows the out5d dataset, using colored parallel coordinates, but without clustering; we can distinguish a purple cluster of data having low values of potassium, thorium and uranium ad high magnetic values, but we have not information about other data in the dataset. In figure 6 (obtained applying clustering and transparency) we can clearly distinguish, in addition to the purple cluster, other ones, for example green points (data with high value of uranium and potassium and very low magnetic values) whose pattern was hidden. With a minimum effort (note that here we have only used automatic clustering and coloring strategies based on radviz) we have found some relations between values of different dimensions (i.e., inverse proportionality between magnetics and uranium and direct relation between uranium and potassium). Starting from these results we can easily isolate and analyze in detail interesting clusters using the simple brushing technique described in 4.1, thanks to the existing relationship between clusters and colors maintained across parallel coordinates and radviz. Moreover, incrementing or decrementing clusters' dimension (using various sizes of the grid), the user has a trade-off between accuracy of representation and cluttering. If we use small clusters almost every point is represented separately, therefore we have many details on single items but potentially high clutter. If we use large clusters we reduce clutter, but this may cause grouping of very different data thus reducing the accuracy of the representation.

5 Conclusion and future work

In this paper we presented a novel multiple view collaboration approach using radviz and parallel coordinates methods. The most innovative elements are that (a) both representations handle multidimensional data, (b) one representation (i.e., radviz) is used at the same time to show the data and to drive the appearance of the parallel coordinates view, and (c) the system uses automatic visualization mechanisms that allow for showing similarities/clustering on the data belonging to the parallel coordinates view exploiting their arrangement on the radviz view. Our ideas have been tested through an ad hoc prototype, SpringView, in order to prove their effectiveness. Actually we are extending the capabilities of our environment through several metrics able to characterize in a formal way the quality of the obtained images and to drive in a more precise way our clustering algorithms.

References

[1] G. Andrienko and N. Andrienko. Parallel coordinates for exploring properties of subsets. In *CMV '04: Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*, pages 93–104. IEEE Computer Society, 2004.

[2] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, page 52. IEEE Computer Society, 1998.

[3] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 81–88. IEEE Computer Society, 2004.

[4] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *VIS '99: Proceedings of the conference on Visualization '99*, pages 43–50. IEEE Computer Society Press, 1999.

[5] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Navigating hierarchies with structure-based brushes. In *INFOVIS '99: Proceedings of the 1999 IEEE Symposium on Information Visualization*, page 58. IEEE Computer Society, 1999.

[6] P. Hoffman, G. Grinstein, and D. Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *NPIVM '99: Proc. Workshop on New Paradigms in Information Visualization and Manipulation*, pages 9–16. ACM Press, 1999.

[7] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, pages 361–378. IEEE Computer Society Press, 1990.

[8] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 89–96. IEEE Computer Society, 2004.

[9] J. C. Roberts. On Encouraging Multiple Views for Visualization. In E. Banissi, F. Khosrowshahi, and M. Sarfraz, editors, *IV'98 – Proceedings International Conference on Information Visualization*, pages 8–14. IEEE Computer Society, July 1998.

[10] H. Siirtola. Combining Parallel Coordinates with the Reorderable Matrix. July.

[11] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119, 2000.

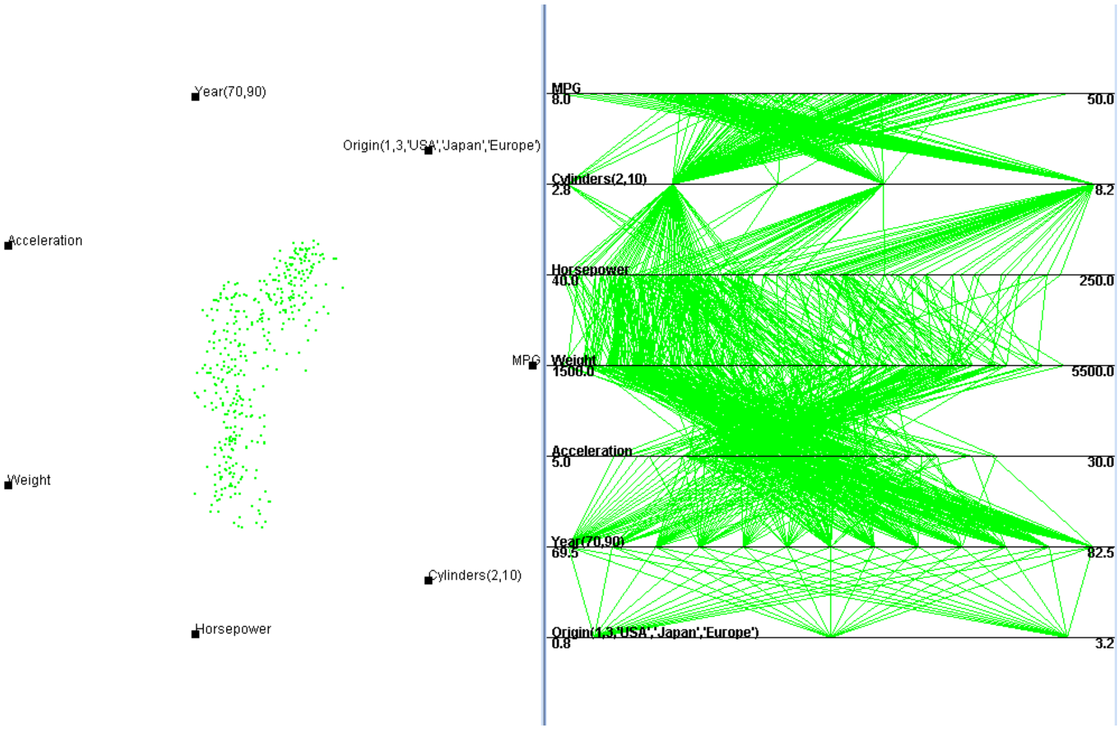

Figure 1. Initial view of SpringView: simple radviz and parallel coordinates views

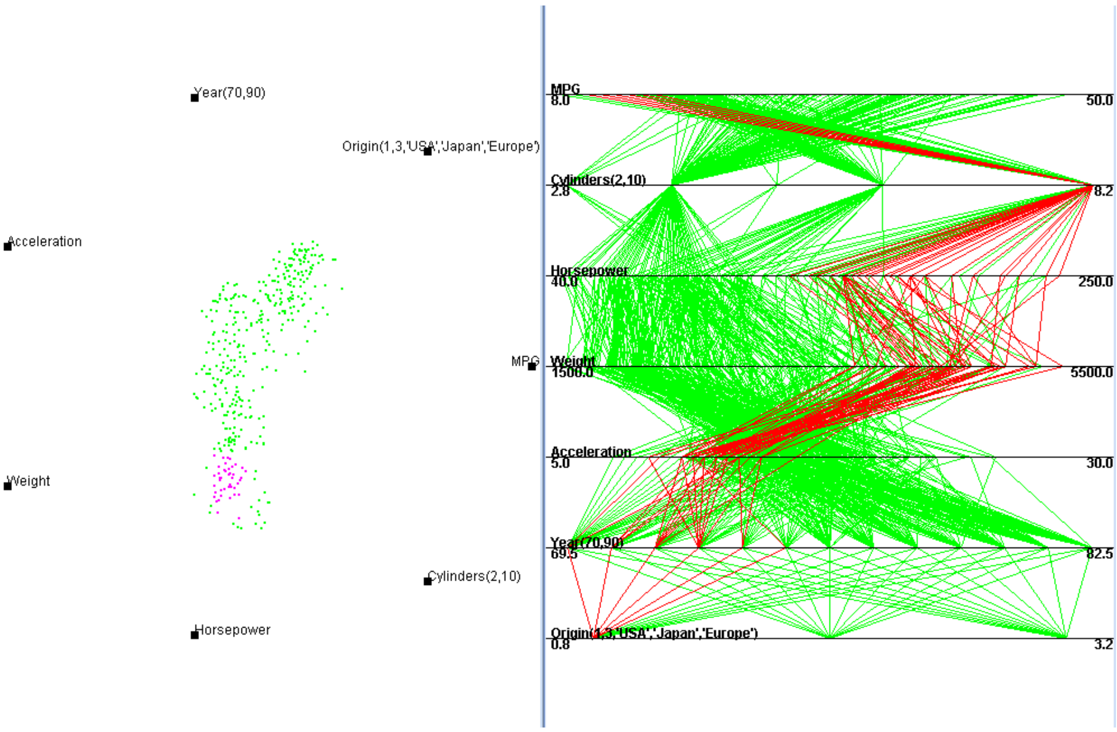

Figure 2. Selection on radviz and description of selected data on parallel coordinates

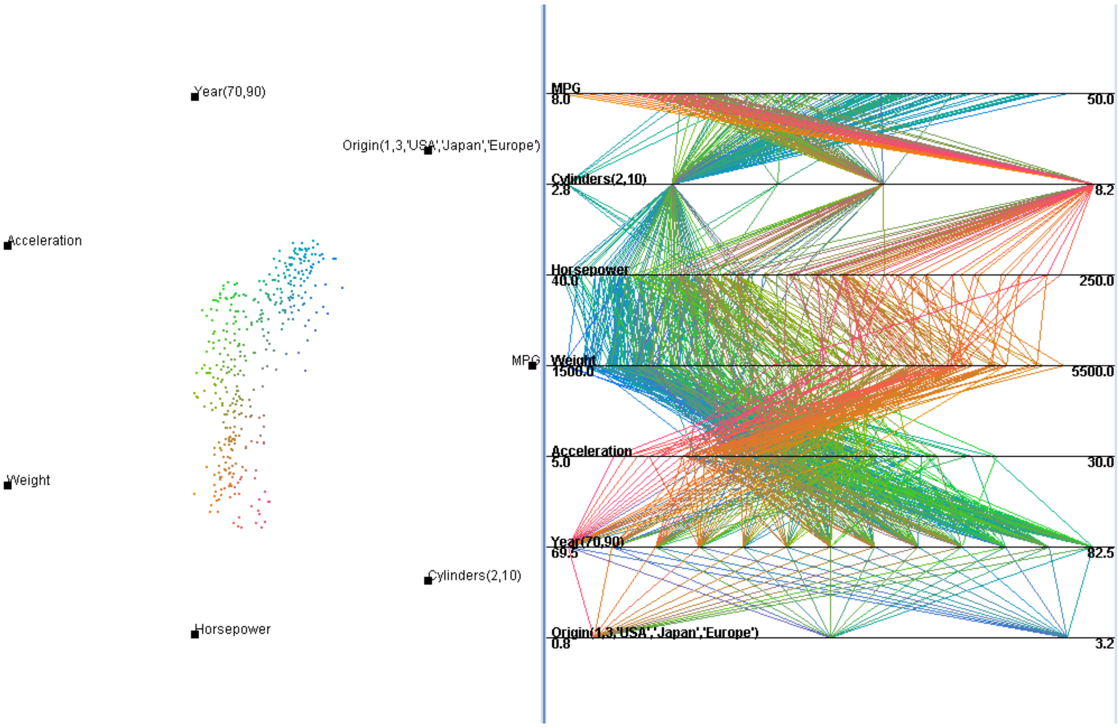

Figure 3. Color map based on radviz presentation, applied to parallel coordinates

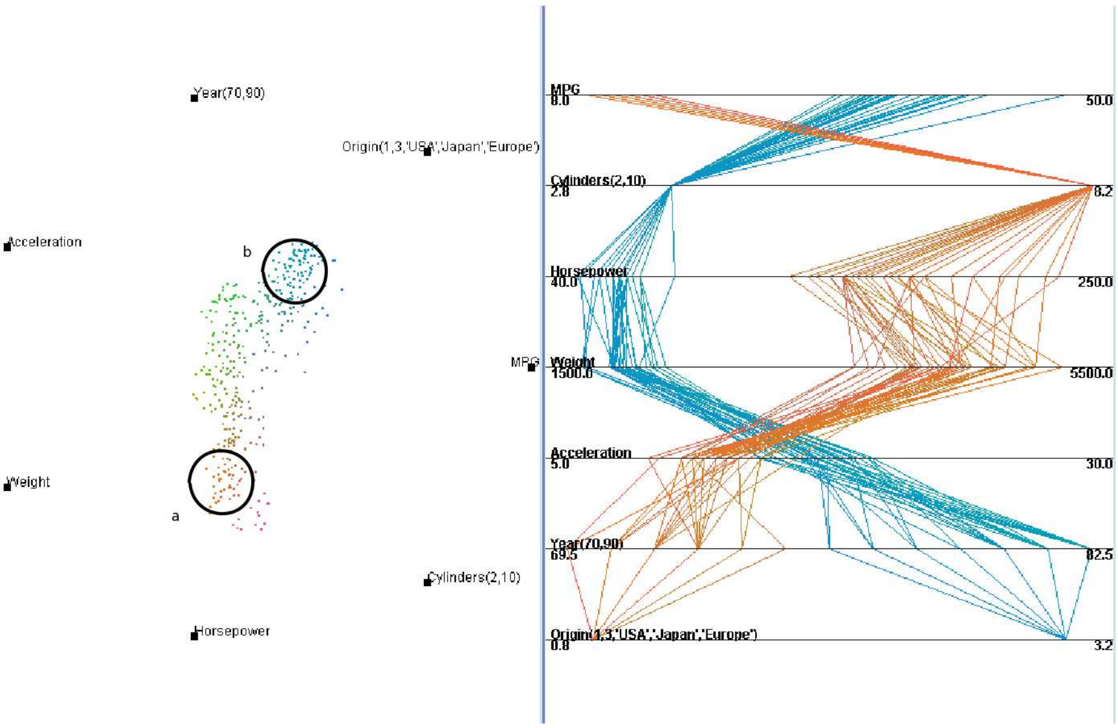

Figure 4. Selection of distinct clusters using coloring. a) powerful cars b) low power cars

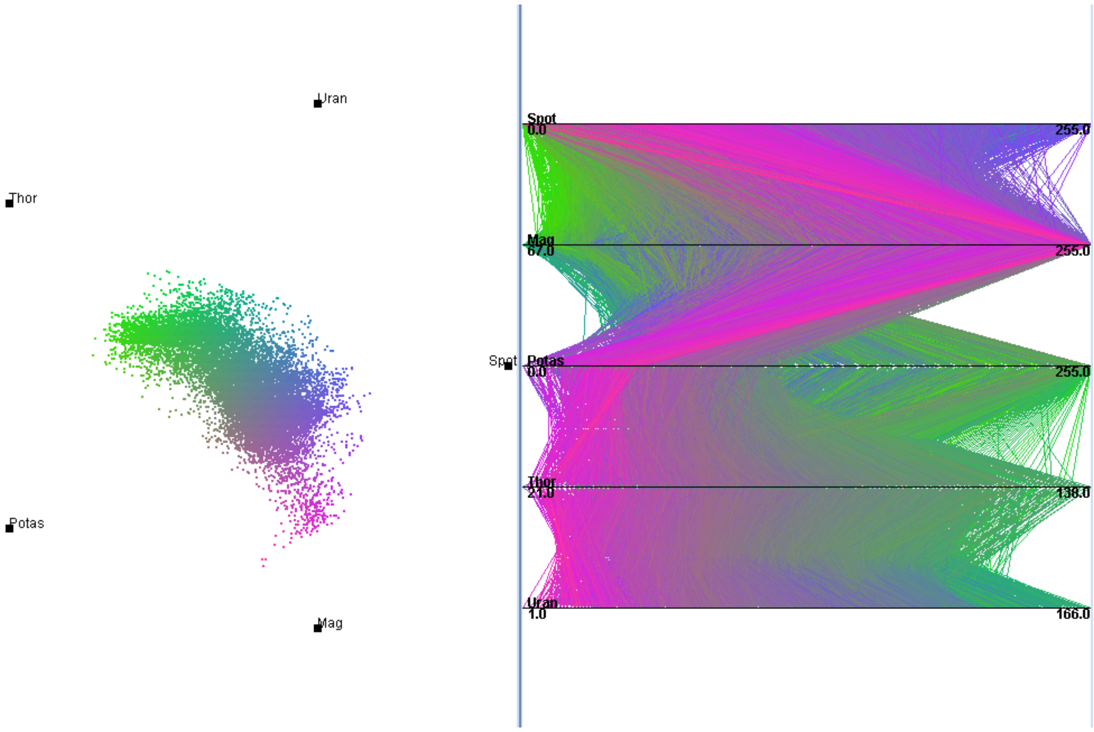

Figure 5. Out5d dataset (high clutter on parallel coordinates representation)

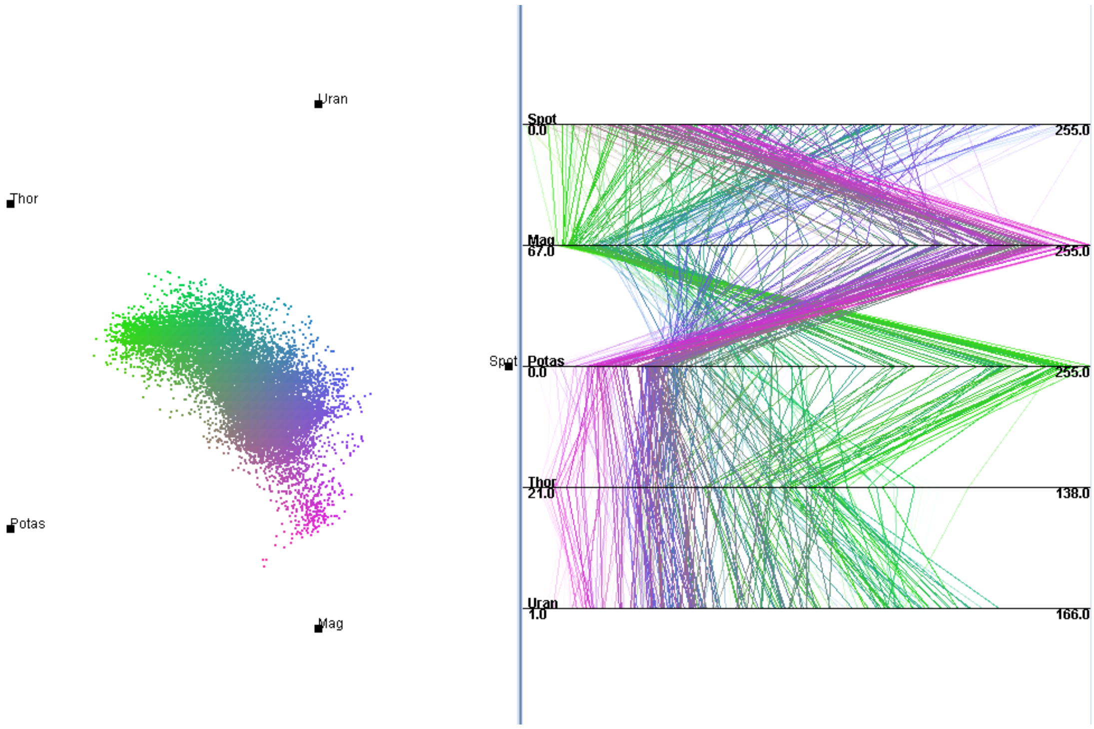

Figure 6. Out5d dataset with clustered parallel coordinates representation