

Visualization of n-dimensional data

Andrii Zakharchenko

April 27, 2020

1 Problem 1

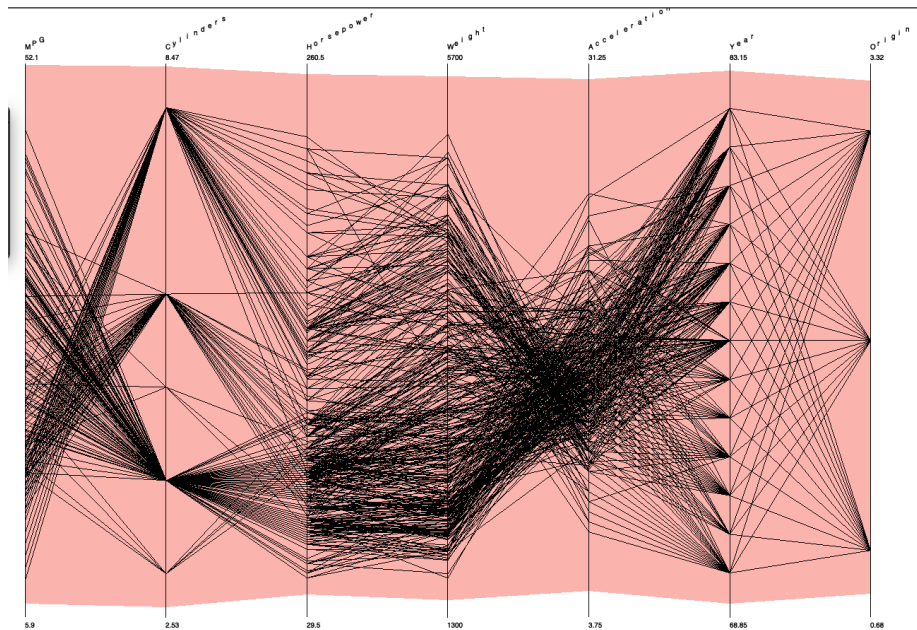


Figure 1: Parallel coordinates

In the figure above it's hard to see correlations between all variables at once. However, we can see that there is negative linear correlation between weight and acceleration, then it looks like there is a positive correlation between horsepower and weight and also it seems that cars with higher MPG tend to have fewer cylinders.

Now let's change a color mapping and color lines by their order in dataset.

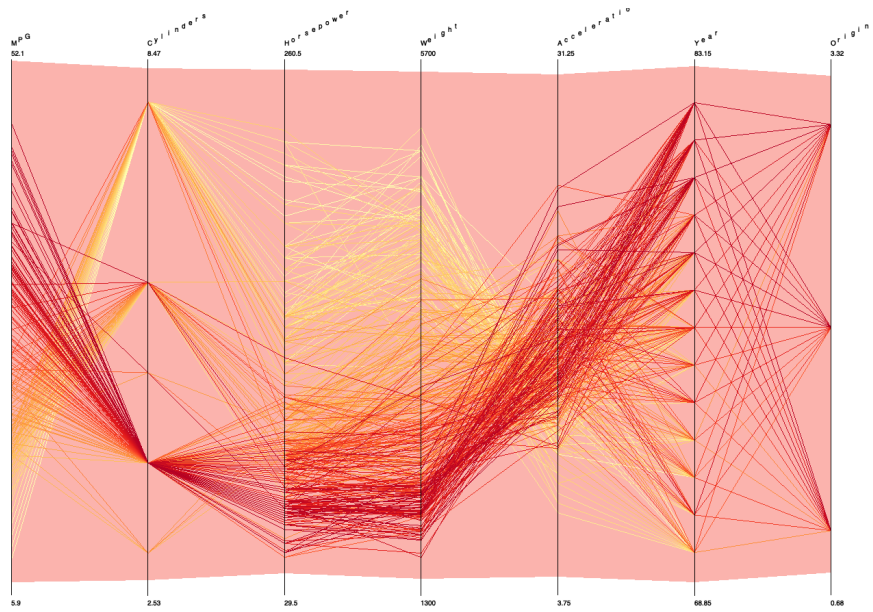


Figure 2: Parallel coordinates. Color by order

With this coloring it's easier to see correlation between multiple variables. For example, in the figure above we can now see more clearly that cars with lower MPG also have more cylinders and more horsepower and higher weight and smaller acceleration.

We can as well color lines by the value of some certain attribute.

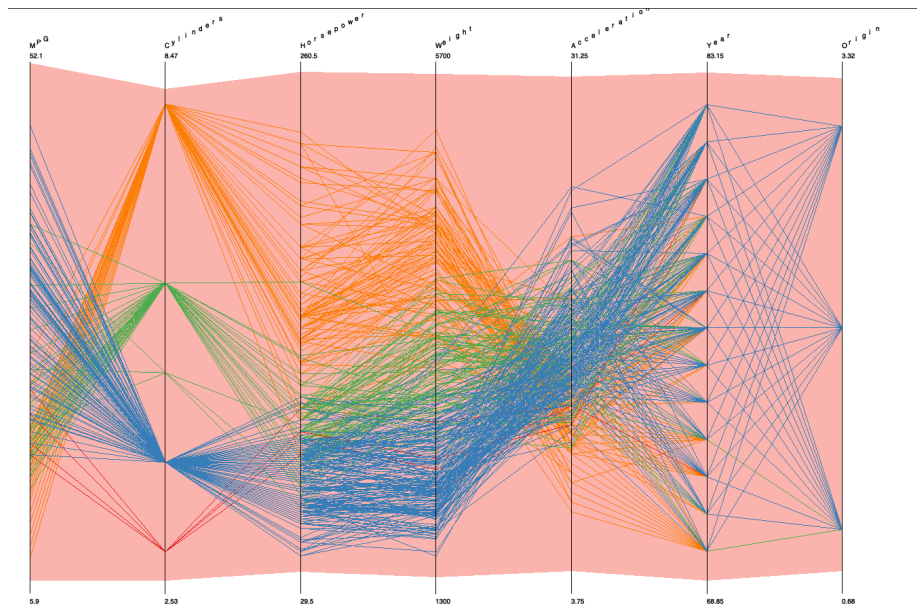


Figure 3: Parallel coordinates. Color by number of cylinders

For example, in the figure above we colored lines by the number of cylinders, which helps us to visualise correlation between this attribute and others. Here, for example, we can see a positive correlation between number of cylinders and weight more clearly. This data set contains several categorical attributes and coloring lines by these attributes can improve the visualisation.

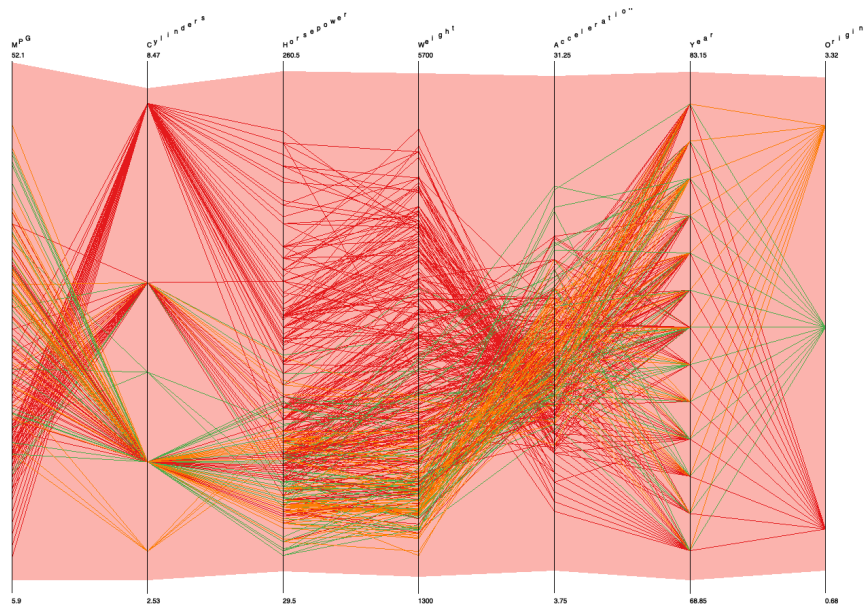


Figure 4: Parallel coordinates. Color by origin

In the figure above lines are colored by the Origin attribute, this allows us to see that cars from "red" country tend to have more horsepower, they are heavier, have more cylinders, slower acceleration and smaller MPG. While cars from 2 other countries have more or less the same characteristics and differ from cars from "red" country.

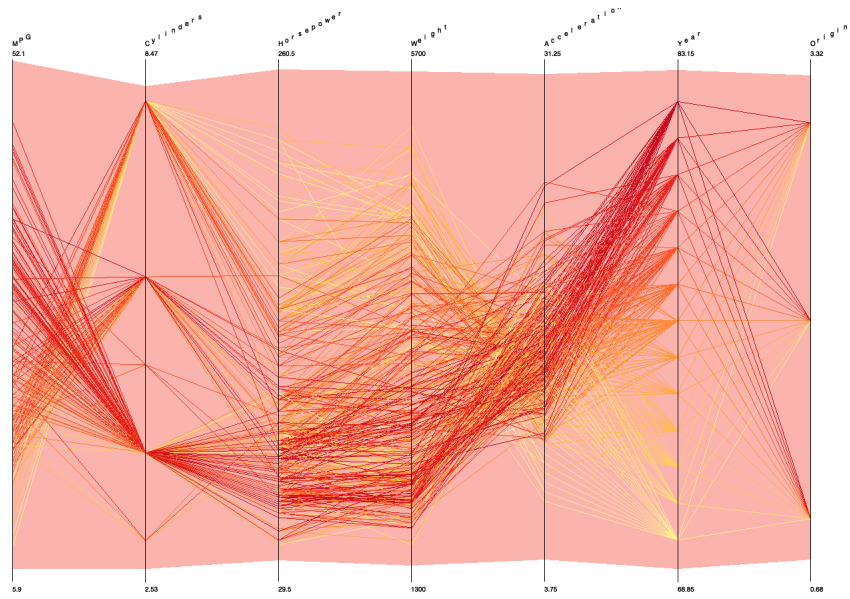


Figure 5: Parallel coordinates. Color by year

In the figure above lines are colored by year. It seems like the newer the car the less horsepower and higher MPG it has.

Finally, let's plot a scatter-plot matrix and have a look at correlations between variables.



Figure 6: Scatter-plot matrix. Color by order

Here we can see a correlation between each pairs of variables. Indeed with some appropriate color mappings we were able to see this correlation using parallel coordinates plot.

Next, let's use brushing to highlight cars with 4 cylinders and high acceleration.

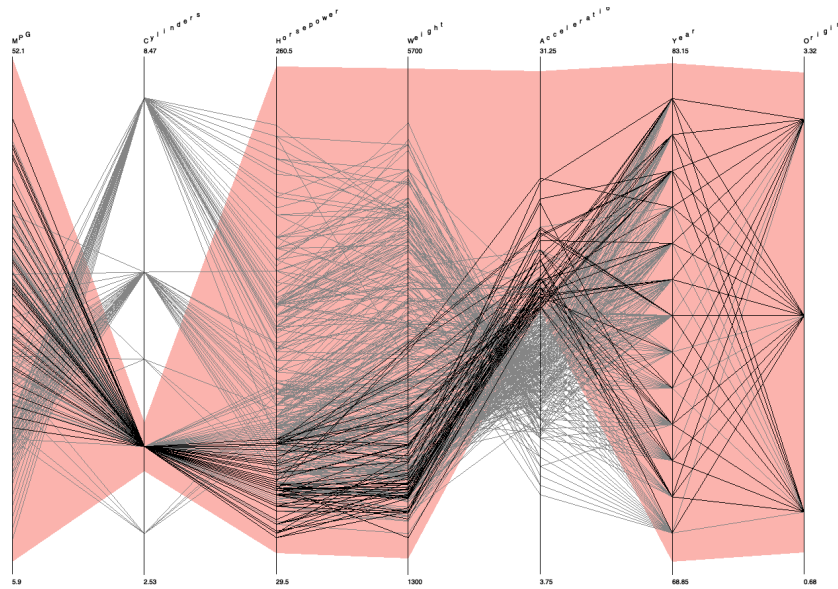


Figure 7

We can see that these cars also have lower horsepower and lower weight.

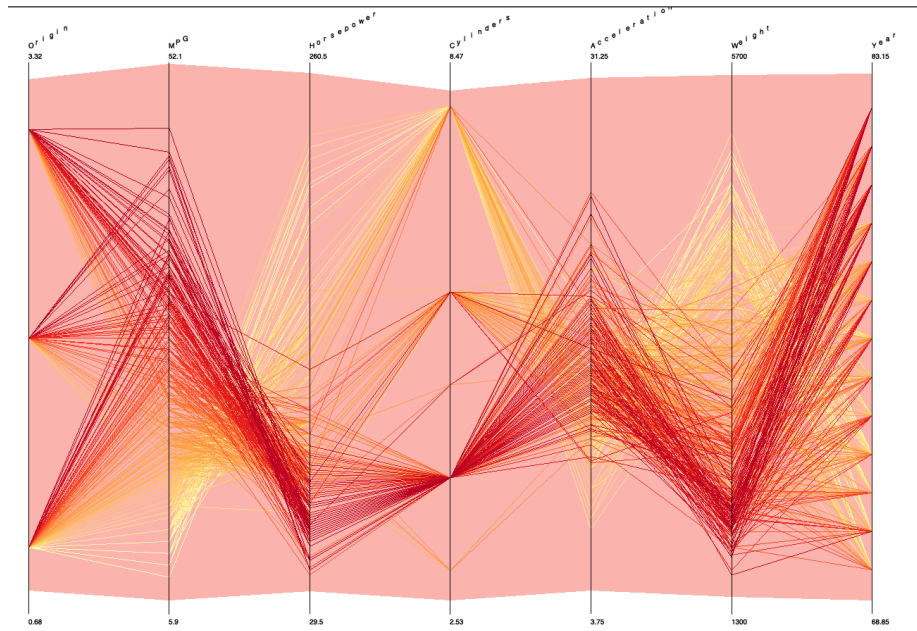


Figure 8: Parallel coordinates with reordered axes

Reordering parallel axes can improve visualisation, I would order axes to emphasize a correlation between attributes and also I would try to place axes with categorical attributes next to continuous attributes so there will be more information. Example of this is shown above.

2 Problem 2

The best correlation between rating and other variables can be seen from scatter-plot.

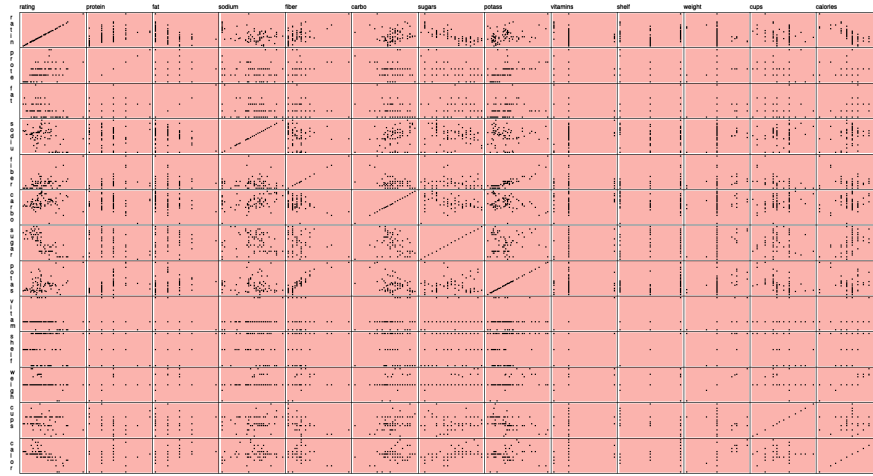


Figure 9

From there we see that there is a strong correlation of rating with sugar and calories.

We can see that cereals with high rating have less sugar and less calories, while cereals with lower rating have more sugar and calories.

3 Problem 3

3.1 Lze některou techniku použít k vizualizaci kategorických n-rozměrných dat? Pomůže nám u některých technik brushing?

Due to categorical nature of this data for this task the most useful technique among parallel coordinates, scatter-plot and glyphs is probably glyphs. Glyphs allow us to estimate number of people of certain categories. For example, as shown in figure 10, we can use brushing to select children that perished. However, from these glyphs it's hard to understand other attributes such as class and gender, maybe some legend would improve this situation.

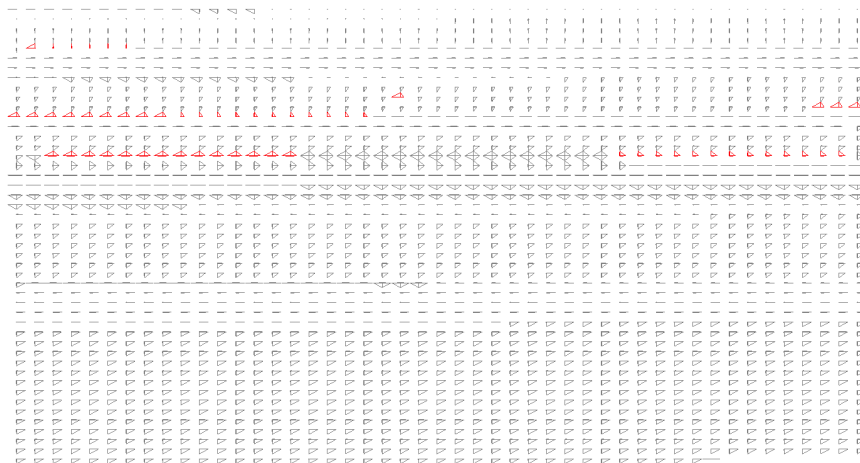


Figure 10

3.2 Zahynuly na Titaniku nějaké děti? V jakých třídách cestovali?

Yes, 52 children has perished on titanic and all of them traveled in 3rd class.

3.3 Existuje mezi třídami nepoměr mezi zahynulými muži?

I would say that there is no **significant** disproportion for different classes among perished men. In the first class there were 180 men and 118 of them died, which accounts for 65%, in second class were 179 men and 154 of them died, which is 86%, in 3rd class 82% of men died, and among crew 77% of men died.

3.4 Existuje mezi třídami nepoměr mezi zahynulými ženami?

It seems like there is a disproportion of perished women among classes. In 1st class not a single woman has died, in 2nd class 12% of women died, among female crew members 3 out of 23 women died and in 3rd class 54% of women died.