Visualization of the performance of baseball players. The STAR report

Andrii Zakharchenko

April 23, 2020

1 Introduction

In this work we are going to visualise data of baseball players. Particularly, we want to be able to visualise batting, pitching and fielding performance of individual players as well as we want to be able to compare several players with each other. In order to pick a correct visualisation techniques we firstly need to determine what kind of data we have. After that we will make an overview of existing state-of-the-art visualisation techniques with an emphasis on the goals stated above. And in the final part of this report we will make a summary of the visualisation techniques and we will state how they can be used in the visualisation of baseball players' performance.

2 Data overview

The documentation of Sean Lahman's Baseball Database could be found here [4]. From there we can see that the database consist of multiple different table, however we are only interested in Batting, Fielding and Pitching tables. The first thing that we can notice is that the data that we are going to visualise is **n-dimensional**. Each player can have multiple different statistics for his performance in batting, pitching or fielding. Another thing that we can notice is that statistics are recorded per year, so apart from being multidimensional the data is also **time-oriented**.

According to data categorization in [1] we can describe our data as multivariate, abstract (by abstract data we mean data that have been collected in a non-spatial context [1]), time-oriented with linear time (Linear time corresponds to our natural perception of time as being a (totally or partially) ordered collection of temporal primitives [1]). With this in mind we can proceed to overview of visualisation techniques.

3 Overview of state-of-the-art visualisation techniques

There exists a variety of different methods that are used to visualise data of different types. In this overview we will mention only those methods that we think can be used to fulfill the goals stated in the Introduction, namely visualise batting, pitching and fielding performance of individual players as well as we want to be able to compare several players with each other.

3.1 Multivariate data visualisation

3.1.1 Parallel Coordinates

Parallel coordinates is a well-know technique where attributes are represented by parallel vertical axes linearly scaled within their data range. Each data item is represented by a polygonal line that intersects each axis at respective attribute data value[3], as shown in 1.

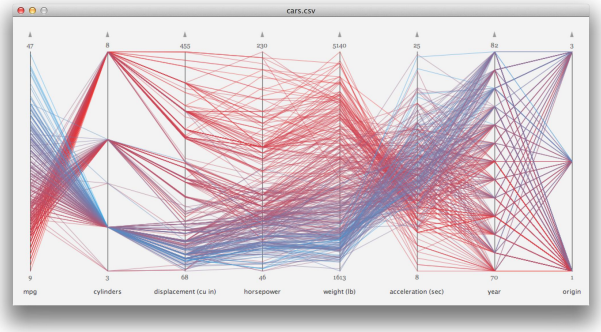

Figure 1: Classical parallel coordinates

Parallel coordinates can be used to study the correlations among attributes by spotting the locations of the intersection points[3]. However, for our purpose of visualising a player's performance another type of parallel coordinated can be used, namely **radial parallel coordinates**[3] shown in 2.

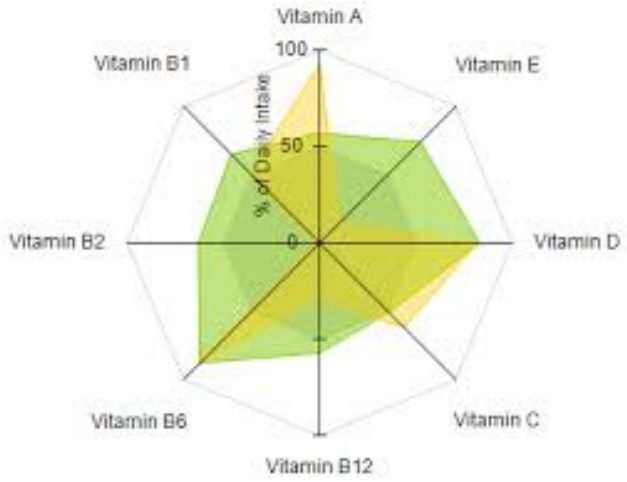

Figure 2: Radial parallel coordinates

Circular Parallel Coordinates is one of the variations of parallel coordinates adopting a radial arrangement of the axis [3]. This visualisation can be useful to our purpose since it allows to better capture overall player's performance according to multiple attributes.

3.1.2 Radial Coordinates Visualization

Radial Coordinates Visualization is similar to parallel coordinates in spirit, in which n lines emanate radically from the center of the circle and terminate at the perimeter, as shown in 3. Each line is associated with one attribute; spring constants attached to the data attribute values define the positions of the data points along the lines. Points with approximately equal or similar dimensional values lie closer to the center [3].

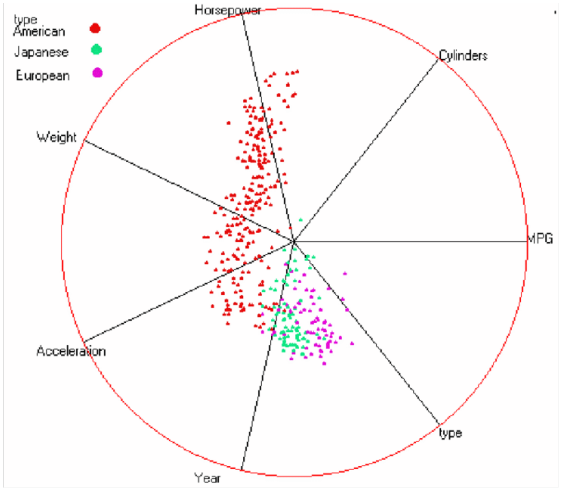

Figure 3: Radial parallel coordinates

This type of visualisation can be useful for comparing multiple players at once and discovering clusters of players.

3.1.3 Chernoff Faces

Chernoff face visualization is an example of icon-based visualisation technique. Two attributes are mapped to the 2D position of a face and remaining attributes are mapped to its properties of the face, for instance, the shape of nose, mouth, eyes and that of the face itself. One of the shortcomings is that different visual features are not quite comparable to each other.

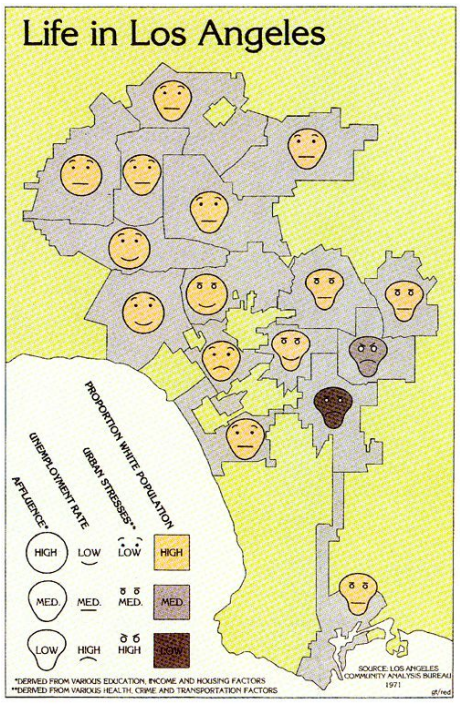

Figure 4: Chernoff faces visualisation

This type of visualisation can be useful to compare players by mapping their statistics to some icons (e.g. player with better batting will be mapped to a bigger bat) this is also can help to plot player and their performance against time axes.

3.1.4 Star Glyph

There are many variants in the glyph family for displaying multidimensional data; star plot is one of the most widely used glyphs. The dimensions are represented as equal angular axes radiating from the center of a circle, with an outer line connecting the data value points on each axis. Each data item is presented by one star glyph. They are helpful for multivariate datasets of moderate size, but their primary weakness is that the display becomes overwhelming when the number of data items increases [3].

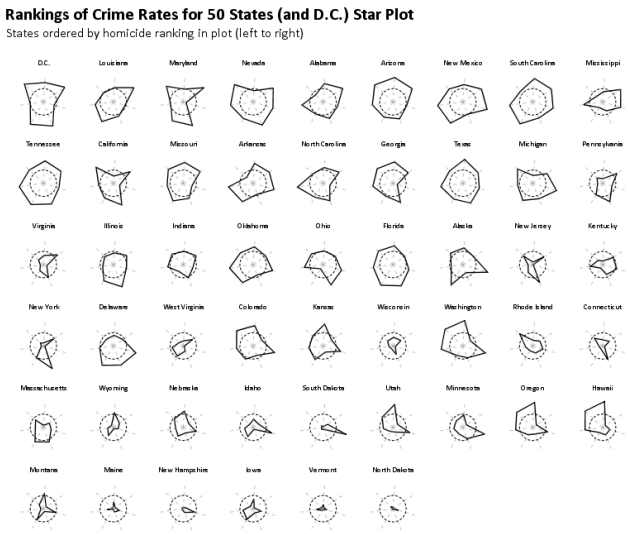

Figure 5: Star glyphs

We can use star glyphs to compare performance of several players by drawing glyphs against time axis, for example. Star glyphs are similar to Radial parallel coordinates in a sense that they allow to capture overall performance better.

3.2 Multivariate time-oriented data visualisation

3.2.1 TimeWheel

The TimeWheel consists of a single time axis and multiple data axes for the data variables. The time axis is placed in the center of the display to emphasize the temporal character of the data. The data axes are associated with individual colors and are arranged circularly around the time axis. In order to visualize data, lines emanate from the time axis to each of the data axes to estab- lish a visual connection between points in time and associated data values [2].

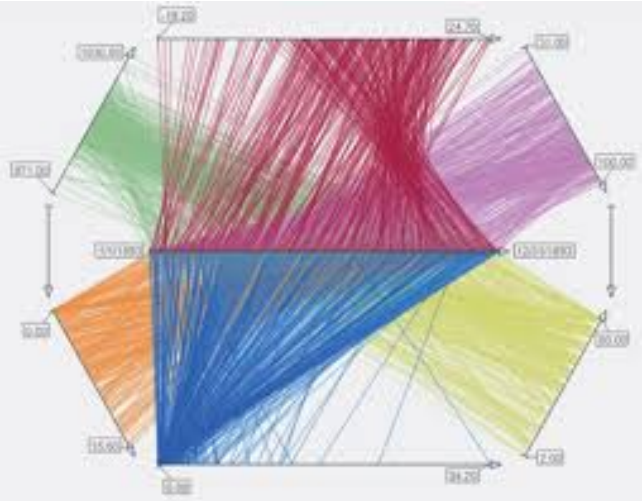

Figure 6: Time Wheel

This type of visualisation can allow us to visualise player's performance accroding to different attribute throught the time.

3.2.2 MultiComb

The rationale behind the MultiComb visualization is to utilize this expressiveness for representing multiple time-dependent variables. Tominski et al. (2004) describe the MultiComb as a visual representation that consists of multiple radially arranged line plots. Two alternative designs exist: time axes are arranged around the display center or time axes extend outwards from the MultiComb's center. In the latter case, optional mirror plots duplicate plots of neighbor variables to ease visual comparison. To maintain a certain aspect ratio for the separate plots, the axes do not start in the very center of the MultiComb. The screen space in the center can therefore be used to provide additional views: a spike glyph can be shown to allow a detailed comparison of data values for a selected time point, or an aggregated view might display the history of a temporal data stream in an aggregated fashion [2].

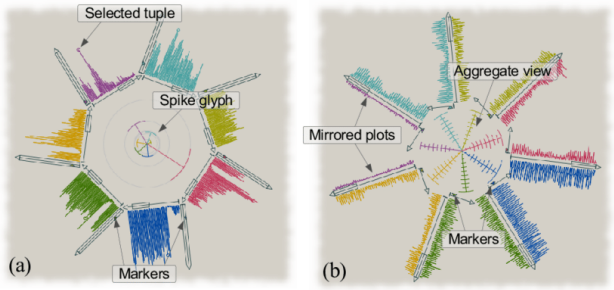

Figure 7: MultiComb

This type of visualisation is somewhat similar to TimeWheel, but here we can see development of certain attribute throughout the time more clearly, but on the other side it is more complicated to perceive overall performance of a player.

3.2.3 CircleView

Keim et al. (2004) developed CircleView for visualizing multivariate streaming data as well as static historical data. Its basic idea is to divide a circle into a number of segments, each representing one variable. The segments are further divided into slots covering periods of time, and color shows the (aggregated) data value for the corresponding interval. Thus, time is mapped linearly along the segments. Keim and Schneidewind (2005) also presented a multi-resolution approach on top of CircleView, where time slots for coarser granularities are shown besides detail values [2].

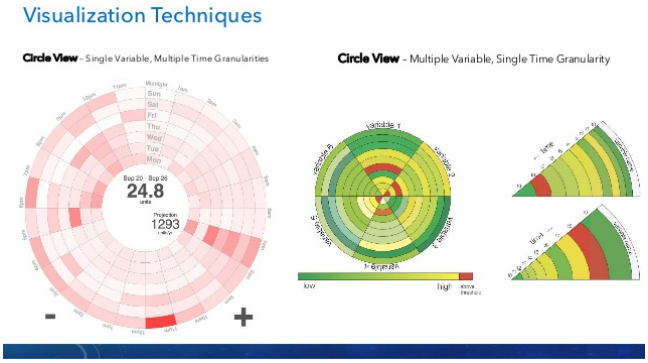

Figure 8: Circle View

This type of visualisation can also help to show development of player's statistics throughout the time.

4 Summary

In this work we gathered several visualisation techniques that can help us visualise performance of baseball players.

The techniques in Multivariate data visualisation section can help us visualise the performance of players according to multiple different attributes. We can also expand these techniques to be able navigate time dimension. One option would be that we can make visualisations interactive, for example, using radial parallel coordinates we can introduce some slider that will allow to change the year. Also, as was stated in star glyph section we can plot the glyphs against some axis, for example axes Y can be categorical and represent a certain player, while axes X will represent a year, thus enabling us to compare performance of multiple players throughout time domain.

And visualisations described in section Multivariate time-oriented data visualisation can help us visualise development of performance of certain players throughout the time. And of course if we want to compare a limited amount of players we can plot this visualisations side-by-side.

Visualisation techniques that we gathered in this report allow us to fulfill the goals of visualising performance of individual players as well as visualising comparison of several players with each other.

References

[1] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, H. Schumann, and Christian Tominski. Visualizing time-oriented data – a systematic view. *Computers & Graphics*, 31:401–409, 06 2007. doi:10.1016/j.cag.2007.01.030.

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[3] W Chan. A survey on multivariate data visualization. 01 2006.

[4] Sean Lahman. Sean lahman's baseball database. URL: http://lahman.r-forge.r-project.org/doc/.