

Analyzing Company Data with Interactive Statistical Graphics

Annerose Zeis^{*}, Sergej Potapov[†], Martin Theus[‡], Antony Unwin[§]

Department of Computational Statistics and Data Analysis, Augsburg University, Germany

1 DATA CLEANING

Big datasets always have problems with quality. The particular issues arising with this dataset are:

- **Removal of all "non-companies". 0 Sales and 0 Employees for all years**
There were 12,324 companies, which neither had a single employee nor any sales over the whole period.
- **Recorded over all years**
Only 7,660 companies have data for all years, i.e. can be analyzed over the complete 15 years.
- **Only one year in business**
2,737 companies were in business for only one year.
- **Missing data during operation**
1,801 companies have missing data within the time of operation, which can be regarded as recording errors. (1 company has data for 1989 and 2003 and no data in between)
- **Identical values in successive years**
Looking at the data more closely, it turns out that many companies state identical sales values in successive years. This not only holds true for small values, e.g. 1 or 5, but even for values like 35,472.7. For the years 1989 and 1990 39.5% of all sales values bigger than 1 in both years are identical. For the years 2002 and 2003 this value is even 41.5%!

2 MAJOR FINDINGS AND VISUALIZATIONS

- 2.1 WAL*MART in Benton, AR is an extreme outlier
- 2.2 There are striking irregularities in sales in 1993 for a few companies
- 2.3 Industry Type "NON" subsumes much information, especially in recent years, probably due to large conglomerates which are not primarily technical firms
- 2.4 Industry Types and States can easily be mapped to show where industries are located and which States are heavily dependent on particular industries
- 2.5 Companies, which move, usually move to neighboring States
- 2.6 Startups vary greatly by Industry Type over time
- 2.7 New York County profited most from the dotcom boom

^{*}e-mail: annerose.zeis@student.uni-augsburg.de

[†]e-mail: sergej.potapov@student.uni-augsburg.de

[‡]e-mail: martin.theus@math.uni-augsburg.de

[§]e-mail: antony.unwin@math.uni-augsburg.de

Maps

For geographically referenced data, it is valuable to look at maps. Figure 1 shows a map of the US States colored according to 2003

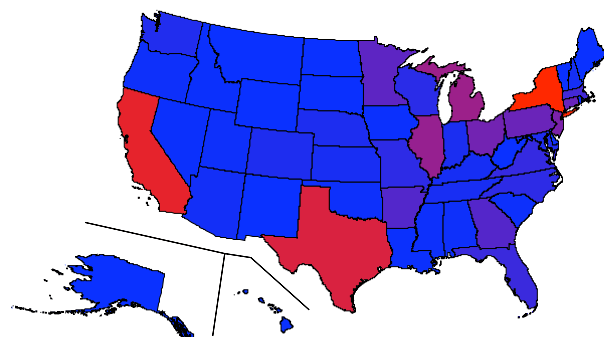


Figure 1: A Map of the US States colored according to 2003 sales.

sales figures. CA, NY and TX clearly stand out. Maps on County level reveal consistent patterns for the New England area as well as for the Bay Area and Los Angeles.

Parallel Coordinate Plots

Parallel Coordinate Plots are the ideal tool to analyze highly multivariate continuous data (cf [6]). They can also be used to display longitudinal data like time series. Figure 2 shows a PCP of the sales data for all years from '89 to '03. A company — which can be iden-

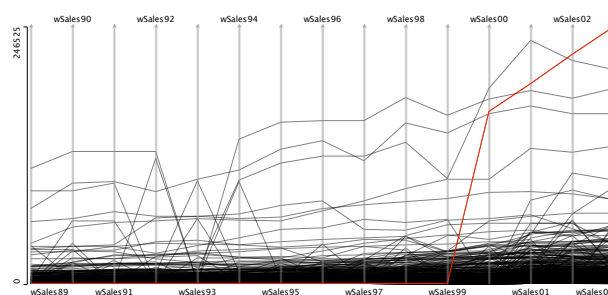


Figure 2: Parallel coordinate plot for all 72,148 companies with non-zero sales.

tified to be WAL*MART — has been selected, that has 0 sales in 1999 and the most sales in 2003.

Mosaic Plots and Fluctuation Diagrams

Mosaic Plots are designed to display multivariate categorical data [2]. Extending the basic definition of Mosaic Plots leads to many variations, which can be used to show different properties of the data [3]. These variations are particularly useful when dealing with only two variables, each with many categories, as we find in this dataset. Figure 4 is a Same Binsize Mosaic Plot, which can be used

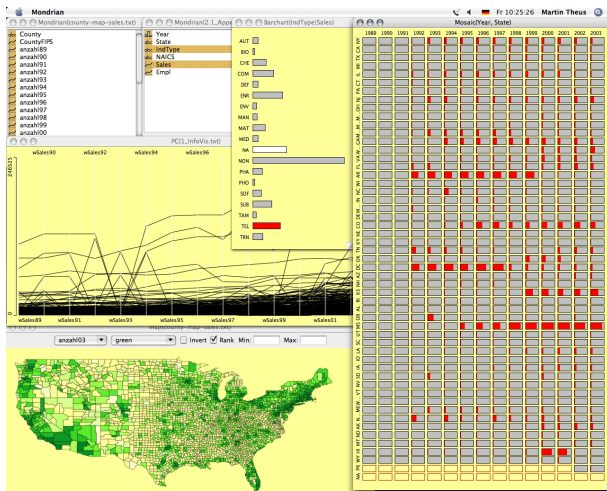


Figure 3: A sample screen shot of a Mondrian session.

to judge how dominant Industry Types are in different States over the years. Industry "CHE" is selected, and you can see how dominant this industry is in DE and how its importance is decreasing in WI.

3 SOFTWARE

3.1 mySQL

A database system is still the most efficient tool for handling bigger datasets, especially for relating tables, which was one important step in matching Counties to ZIP-codes. Although this is a trivial task for the database, aggregating over data which change over time is not.

3.2 R

The statistical computing environment R (<http://www.r-project.org>) makes it very easy to restructure data. Most of the derived datasets needed for specific analyses were generated within R. R is not efficient at this, so that if the problem were larger by a factor of 10, mySQL would be our choice to handle the data.

3.3 Mondrian

To visualize the data, the Mondrian data analysis and visualization software [4] was used. Amongst other features, Mondrian is designed for handling big datasets, in terms of both cases and variables. Plots for multivariate categorical and continuous data are available as well as special plotting techniques like α -blending to deal with overplotting.

4 THE ROLE OF INTERACTIVITY

There is no single best visualization for a complex dataset like the one we looked at here. With highly interactive tools ([5]), it is possible to look at the data from many different angles. Linking between views adds dimensionality and can give deeper insights. Common scaling (i.e. ensuring that equivalent scales were used inside plots for different, but comparable variables and across plots for the same variable) is often valuable. Having a range of flexible viewing options for multivariate plots, such as fluctuation plots and other alternatives for mosaic plots (and not forgetting using various weighting variables and zooming) encourages effective exploratory

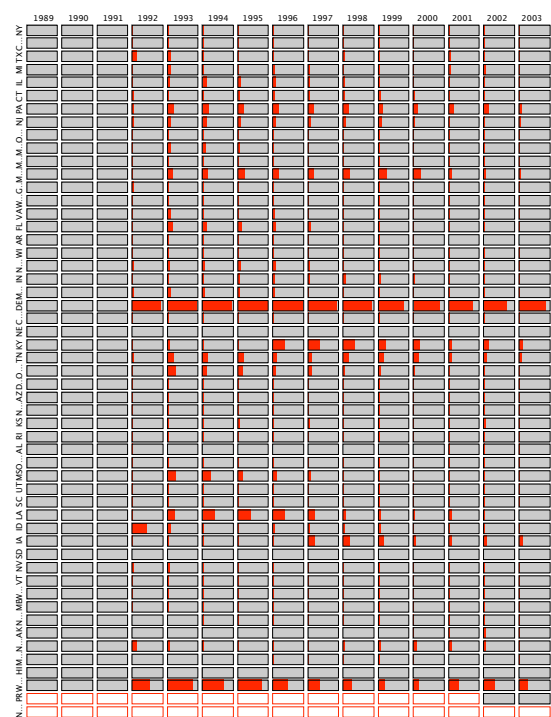


Figure 4: A Mosaic Plot in the 'Same Binsize' view.

analysis. Once interesting patterns are identified, static views can be used to display specific properties of the data.

5 VISUALIZATION LESSONS LEARNED FROM THE DATASET

As always, visualization tools are excellent for investigating data quality and tracking down data problems. It is important that the tools needed for analysis of the company data, both basic displays like barcharts, boxplots, histograms and scatterplots and more sophisticated multivariate displays like Parallel Coordinate Plots, (weighted) Mosaic Plots and Maps, are available in an integrated interactive software. What would additionally be of great assistance would be methods for seamlessly switching between different levels of aggregation, for instance between displays for individual companies and displays for aggregations by Industry Type, by NAICS codes or by States. Naturally, a proper linking structure would be needed between these different levels. Interactivity is key to the successful exploration of large datasets, the more flexible and powerful the interactivity the better.

REFERENCES

- [1] J Dykes, A MacEachren, and M-J Kraak. *Exploring Geovisualization*. Elsevier, 2005.
- [2] Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994a.
- [3] Heike Hofmann. Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1):11–26, 2000.
- [4] M. Theus. Interactive Data Visualization using Mondrian. *Journal of Statistical Software*, 7(11), 2002.
- [5] Antony R. Unwin. Requirements for interactive graphics software for exploratory data analysis. *Journal of Computational Statistics*, 1:7–22, 1999.
- [6] E.J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675, 1990.