Top-down data Analysis with Treemaps

Martijn Tennekes Edwin de Jonge

March 1, 2011



Outline

- Introduction
- Methods
 - Comparison treemaps
 - Density treemaps
 - Confidence intervals
- Implementation in R
- Conclusions and future research

A National Statistical Institute (NSI) produces statistics on:

- Economic growth
- Consumer pricing
- Income of persons and households
- Count of population
- Unemployment
- . . .

Typical production process:

• Data collection

- Survey data
- Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Typical production process:

• Data collection

- Survey data
- Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Typical production process:

- Data collection
 - Survey data
 - Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Typical production process:

- Data collection
 - Survey data
 - Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Typical production process:

- Data collection
 - Survey data
 - Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Typical production process:

- Data collection
 - Survey data
 - Administrative data
- Data editing
 - Automated editing
 - Interactive editing
- Data analysis (aggregated level)
- Publication

Interactive data editing:

Traditional approach

- Analysis of data at record level
- Tabular format only

Top-down approach

- Analysis of aggregated data In case of unexpected outcome: zoom in
- Tabluar format and visualizations

Interactive data editing:

Traditional approach

- Analysis of data at record level
- Tabular format only

Top-down approach

- Analysis of aggregated data In case of unexpected outcome: zoom in
- Tabluar format and visualizations

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - Automated data editing
 - Thusesauds error. Respondents are requested to fill in the value in thousands, but they often don't:
 - Wrong sign: Respondents often fill in a negative number for variables such as expenditures
 - Obvious typos
 - Interactive data editing and analysis
 - Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - Automated data editing
 - Thousands error Respondents are requested to fill in the value in thousands, but they often don't
 - variables such as expenditures
 - Obvious typos
 - Interactive data editing and analysis
 - Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- O Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- ④ Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- O Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- Output Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- O Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- Publication

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos



Interactive data editing and analysis

Structural Business Statistics (SBS)

- Large business survey
- Production process:
 - Survey of circa 50,000 responding enterprices
 - 2 Automated data editing

Thousands error Respondents are requested to fill in the value in thousands, but they often don't

Wrong sign Respondents often fill in a negative number for variables such as expenditures

Obvious typos

- Interactive data editing and analysis
- O Publication

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

• Space-filling visualization method

- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Treemaps

- Space-filling visualization method
- Hierarchically structured data
- Applications
 - Hard drive storage
 - Stock market analysis
 - Statistical data editing/analysis

Comparison treemaps

Goal

Detect disruptive or unextected changed in time

> Sizes Aggregated variable y at period t Colors Growth of y w.r.t. period t - 1

Chemicals. Food, beverages, Transportation, storage, communication chemical products and tobacco Manufacturing Transport Coke, petroleum equipment n.e.c. equipment products, and Real estate, renting, nuclear fuel Construction business activities Pulp, paper, Rubber Other and publishing, plastic minera products and printing products Metals and Electrical and metal products N.e.c. optical equipment Mining and quarrying Electricity, gas, persona Wholesale and retail trade water supply Hotels and estaurants socia

20% 30% 40% 50% 60%

Growth w.r.t. last year

Total value added

Top-down data Analysis with Treemaps

-20% -10% 0%

Density treemaps

Goal

Analyze the relationship between two variables

Sizes Aggregated variable y Colors Density parameter x/y

N.e.c.	Machinery and equipment n.e.c.	l Pulp, paper, publishing, and printing	Real estate, renting, business activities	
Food, beverage Man and tobacco	ufacturing	Electrical and optical equipment		
Metals and metal products	Transport pro equipment Other n minera	ber and wood and wood and wood products and textile and textile and textile products and text		
Wholesale and retail trade			Transportation, storage, communication	
			Hotels and restaurants	Health and social work

1 2 3 4 5 Turnover (in millions) per person employed

Top-down data Analysis with Treemaps

Number of persons employed

Density treemaps

Goal

Analyze the relationship

Sizes Aggregated

Colors Density

variable v

parameter x/y

between two variables

Turnover



2 4 6 8 10 12 14 16 18 20 Number of persons employed per one million euro turnover

Top-down data Analysis with Treemaps

22

Confidence intervals

Goal

Visualize the confidence interval along with the corresponding estimation of a parameter



Implementation in R

- Package treemap
- Available on CRAN
- Main function

```
\label{eq:mplot_tmplot_tmplot} \begin{array}{l} tmPlot\ (myDataFrame, \\ index = myIndexVariables, \\ vSize = mySizeVariable, \\ vColor = myColorVariable, \\ \ldots) \end{array}
```

• Used algorithm: ordered treemap (pivot-by-size)

Conclusion

- Method of visualizing hierarchically structured data
- Top-down data analysis

Future research

- Communication between R visualizations and MacroView (tool developed at Statistics Netherlands for top-down analysis)
- Evaluation of treemaps by data analysts

Conclusion

- Method of visualizing hierarchically structured data
- Top-down data analysis

Future research

- Communication between R visualizations and MacroView (tool developed at Statistics Netherlands for top-down analysis)
- Evaluation of treemaps by data analysts

Conclusion

- Method of visualizing hierarchically structured data
- Top-down data analysis

Future research

- Communication between R visualizations and MacroView (tool developed at Statistics Netherlands for top-down analysis)
- Evaluation of treemaps by data analysts

Conclusion

- Method of visualizing hierarchically structured data
- Top-down data analysis

Future research

- Communication between R visualizations and MacroView (tool developed at Statistics Netherlands for top-down analysis)
- Evaluation of treemaps by data analysts