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

Introduction

A National Statistical Institute (NSI) produces statistics on:

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

Introduction

Typical production process:

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

Introduction

Typical production process:

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

Introduction

Typical production process:

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

Introduction

Typical production process:

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

Introduction

Typical production process:

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

Introduction

Typical production process:

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

Introduction

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

Introduction

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

Introduction

Structural Business Statistics (SBS)

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

Introduction

Structural Business Statistics (SBS)

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

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Introduction

Structural Business Statistics (SBS)

- Large business survey
- Production process:
  1. 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
  3. Interactive data editing and analysis
  4. Publication

Methods

Treemaps

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

Methods

Treemaps

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

Methods

Treemaps

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

Methods

Treemaps

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

Methods

Treemaps

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

Methods

Treemaps

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

Methods

Treemaps

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

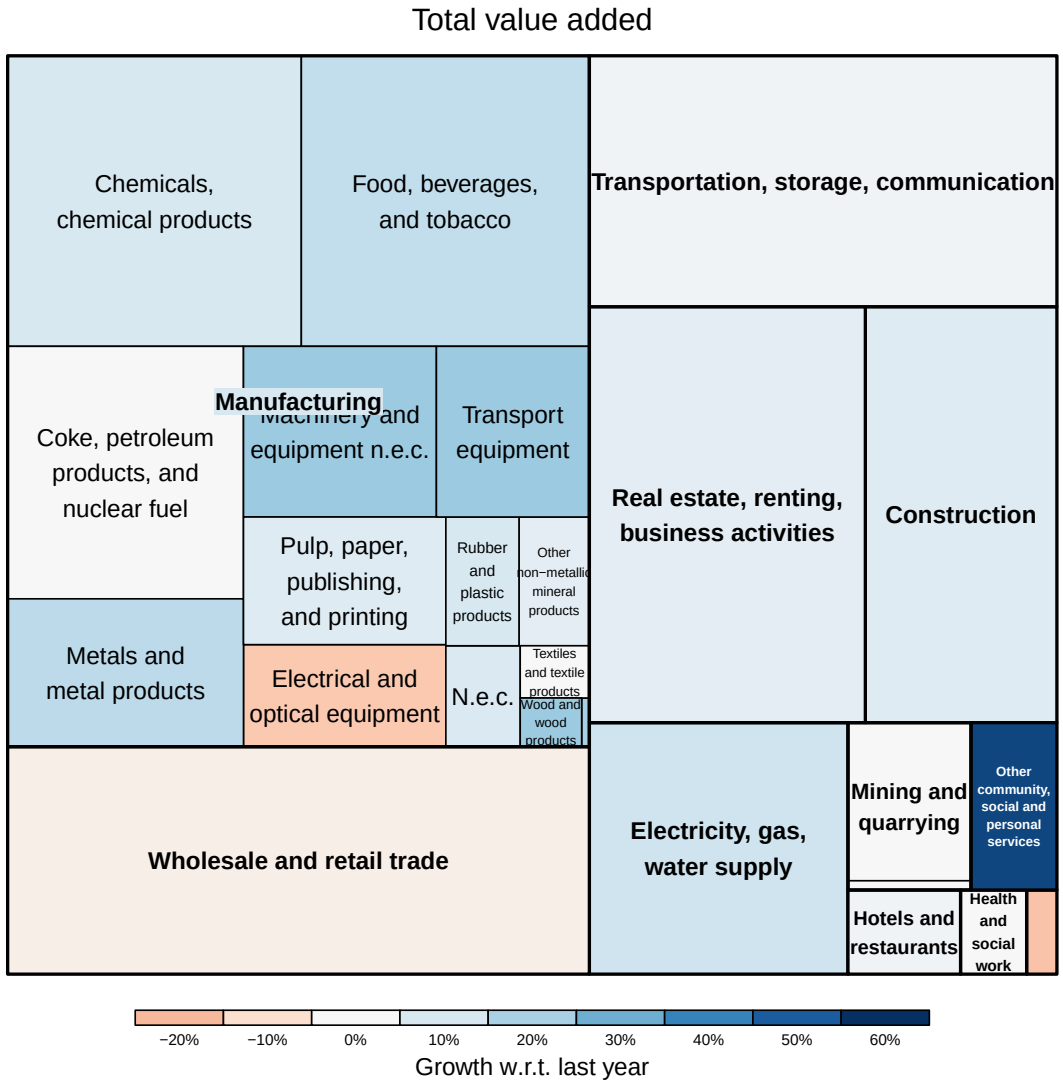
Comparison treemaps

Goal

Detect disruptive or unextected changed in time

Sizes Aggregated variable y at period t

Colors Growth of y w.r.t. period $t-1$



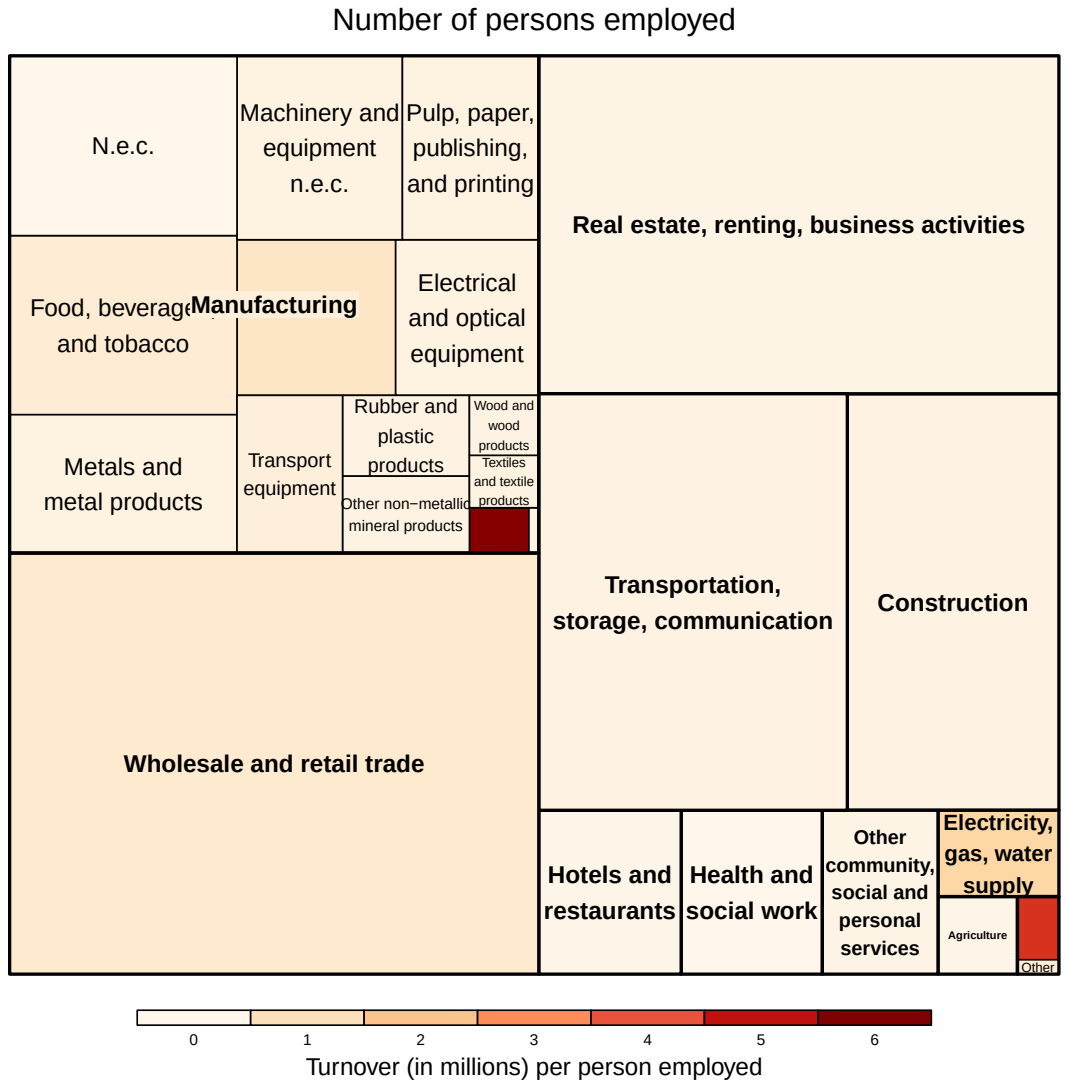
Density treemaps

Goal

Analyze the relationship between two variables

Sizes Aggregated variable y

Colors Density parameter x/y



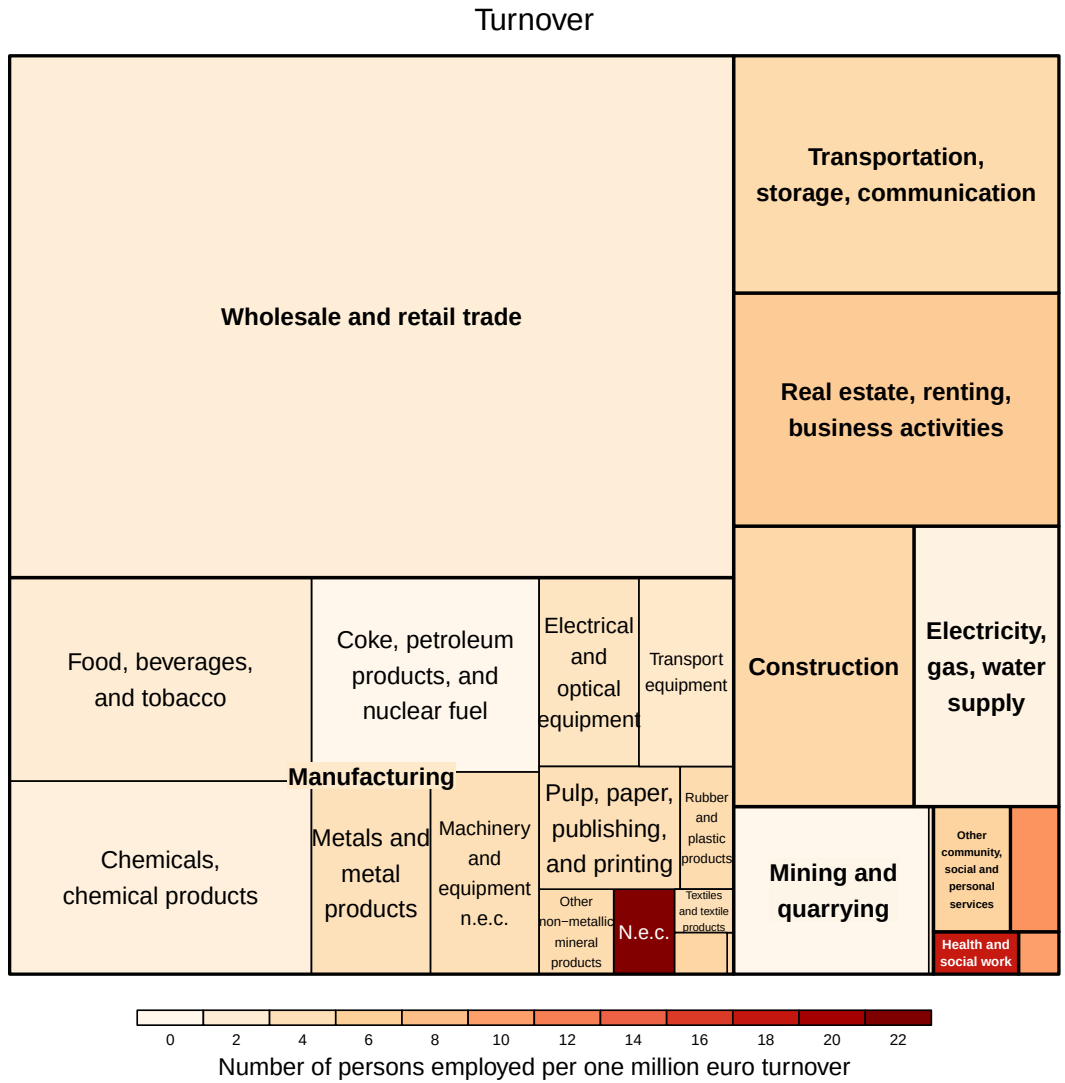
Density treemaps

Goal

Analyze the relationship between two variables

Sizes Aggregated variable y

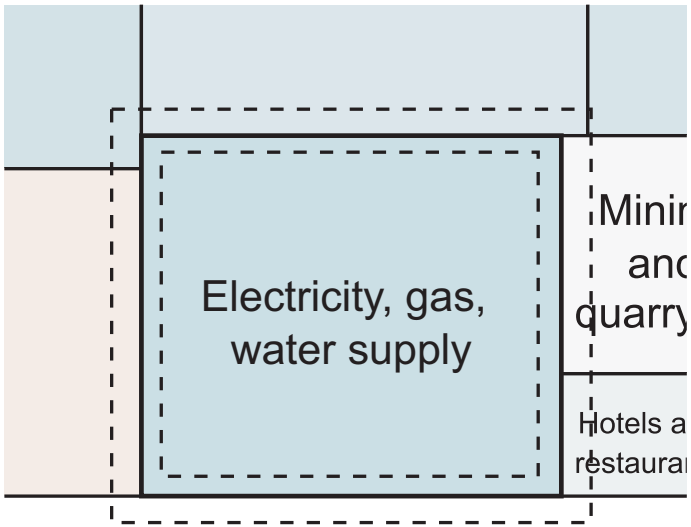
Colors Density parameter x/y



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

  tmPlot
  ```
  tmPlot(myDataFrame,
         index = myIndexVariables,
         vSize = mySizeVariable,
         vColor = myColorVariable,
         ...)
  ```

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

Conclusion and future research

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 and future research

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 and future research

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 and future research

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