

Discovering the unusual with Data Mining

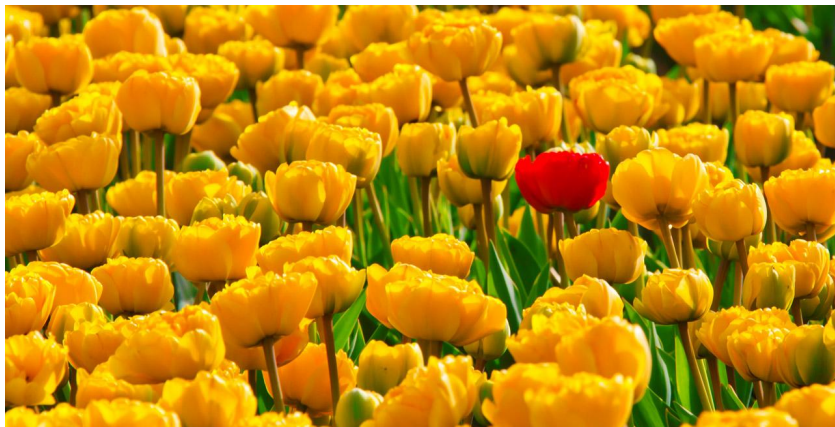
Subgroup Discovery and Preference Learning

Cláudio Rebelo de Sá
24 May 2017

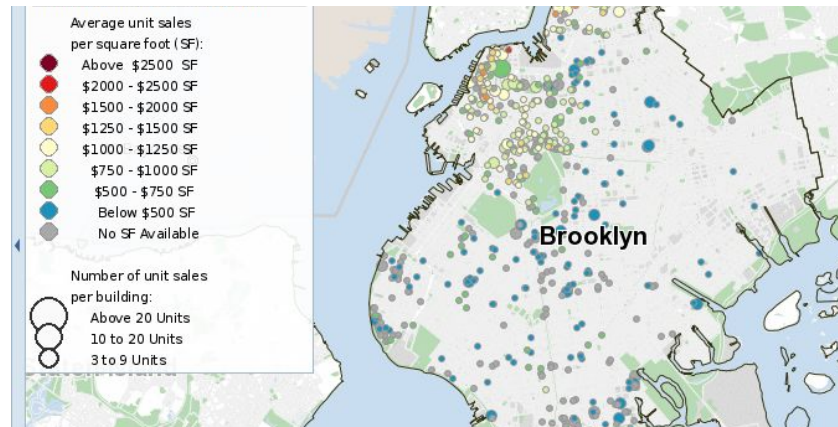


Subgroup Discovery, Outlier Detection and Anomaly Detection

Anomaly Detection / Outlier detection
Finding individual deviating points

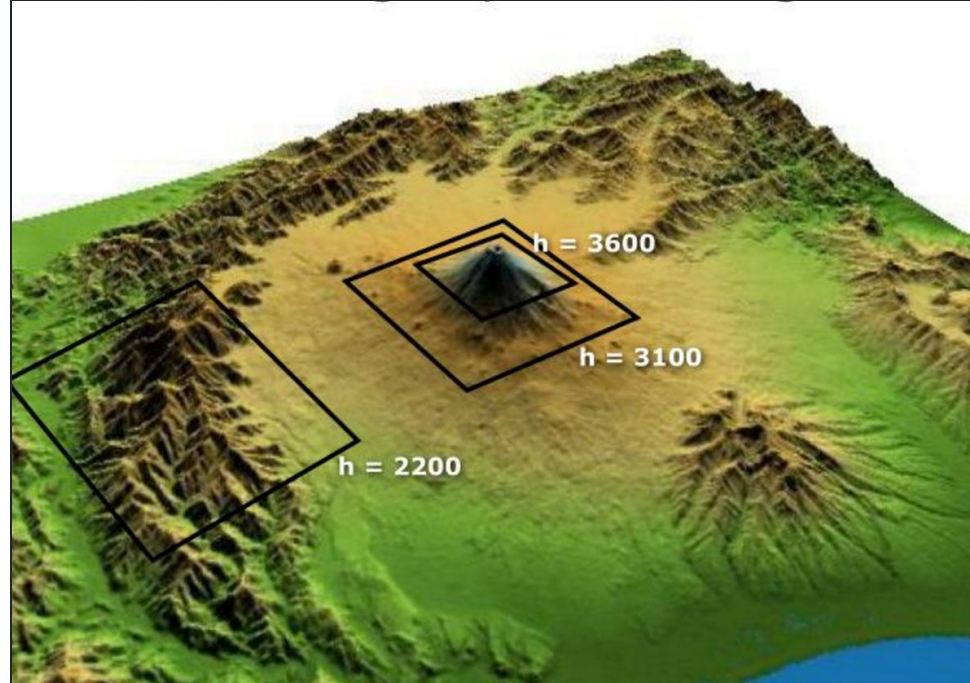


Subgroup Discovery
Finding (descriptions of) deviating groups



Subgroup Discovery

Data mining framework that seeks **subsets** of the dataset where something **exceptional** is going on.



Subgroup Discovery

In typical pattern mining, interestingness is measured by the **frequency** of the pattern.

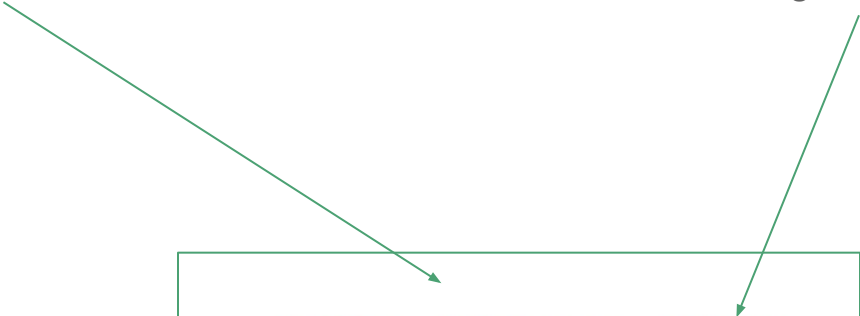
Subgroup Discovery (Klosgen et al. 2002) on the other hand, measures interestingness in a **supervised form**.

$\text{Age} \geq 30 \wedge \text{Likes} = \text{Salmon Roe}$ is unusual

Quality measures

- ★ Specifies what makes a subgroup exceptional.

- ★ Favours larger subgroups (hence avoiding singular cases)


$$QM_S = distance_S \cdot size_S$$

Preference Learning

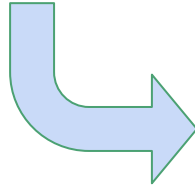


Representation of preferences

A_1	π				alternative π
	λ_1	λ_2	λ_3	λ_4	
0.1	4	3	1	2	$\lambda_3 \succ \lambda_4 \succ \lambda_2 \succ \lambda_1$
0.2	3	2	1	4	$\lambda_3 \succ \lambda_2 \succ \lambda_1 \succ \lambda_4$
0.3	1	4	2	3	$\lambda_1 \succ \lambda_3 \succ \lambda_4 \succ \lambda_2$
0.4	1	3	2	4	$\lambda_1 \succ \lambda_3 \succ \lambda_2 \succ \lambda_4$

Representation of preferences

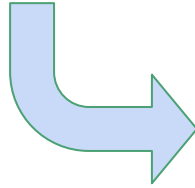
$\lambda_2 > \lambda_3 > \lambda_1$



	λ_1	λ_2	λ_3
λ_1			
λ_2			
λ_3			

Representation of preferences

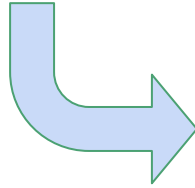
$$\lambda_2 > \lambda_3 > \lambda_1$$



	λ_1	λ_2	λ_3
λ_1			
λ_2			
λ_3			

Representation of preferences

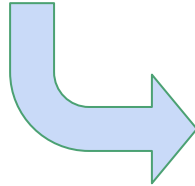
$$\lambda_2 > \lambda_3 > \lambda_1$$



	λ_1	λ_2	λ_3
λ_1			
λ_2			
λ_3			

Representation of preferences

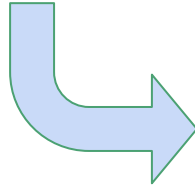
$$\lambda_2 > \lambda_3 > \lambda_1$$



	λ_1	λ_2	λ_3
λ_1			
λ_2			
λ_3			

Representation of preferences

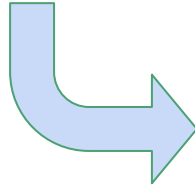
$$\lambda_2 > \lambda_3 > \lambda_1$$



	λ_1	λ_2	λ_3
λ_1			
λ_2			
λ_3			

Representation of preferences

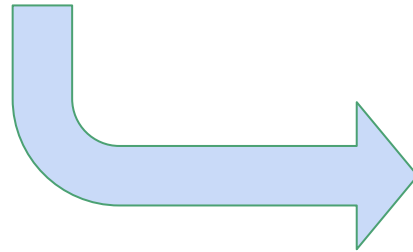
$$\lambda_2 > \lambda_3 > \lambda_1$$



	λ_1	λ_2	λ_3
λ_1	0	-1	-1
λ_2	1	0	1
λ_3	1	-1	0

Aggregation of preferences

A_1	π				alternative π
	λ_1	λ_2	λ_3	λ_4	
0.1	4	3	1	2	$\lambda_3 \succ \lambda_4 \succ \lambda_2 \succ \lambda_1$
0.2	3	2	1	4	$\lambda_3 \succ \lambda_2 \succ \lambda_1 \succ \lambda_4$
0.3	1	4	2	3	$\lambda_1 \succ \lambda_3 \succ \lambda_4 \succ \lambda_2$
0.4	1	3	2	4	$\lambda_1 \succ \lambda_3 \succ \lambda_2 \succ \lambda_4$



$$\begin{bmatrix} 0 & 0 & 0 & 0.5 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ -0.5 & 0 & -1 & 0 \end{bmatrix}$$

Pairwise

If most people prefer:

tamago > **kappa-maki**

a subgroup where most people prefer:

kappa-maki > **tamago**

will be interesting

$$L_{\hat{S}_1} = \begin{bmatrix} 0 & -0.5 & -0.5 & -0.25 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 \end{bmatrix}$$

Labelwise

If a subgroups ranks **tekka-maki** consistently in the **top 3** while the majority in the dataset ranks it in the **last 3**, this measure will find it to be very interesting.

$$L_{\hat{S}_1} = \begin{bmatrix} 0 & -0.5 & -0.5 & -0.25 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 \end{bmatrix}$$

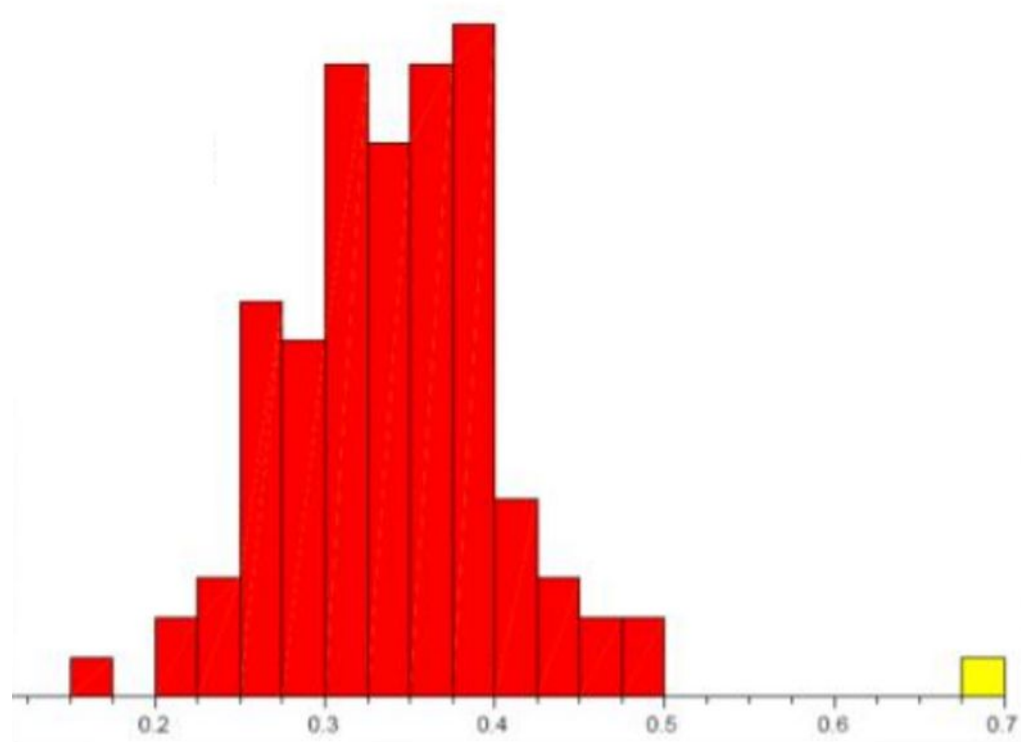
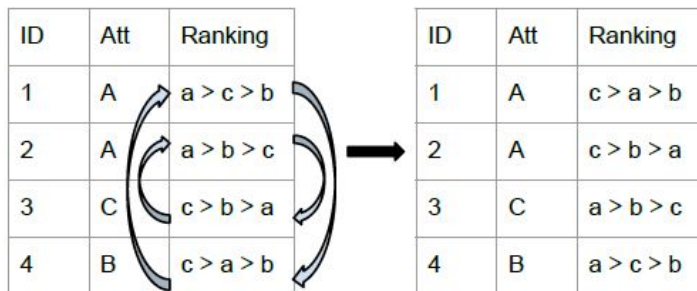
Rankingwise

If a subgroups ranks **a set of labels** in the **opposite order** of the majority in the dataset, this measure will find it to be very interesting.

$$L_{\hat{S}_1} = \begin{bmatrix} 0 & -0.5 & -0.5 & -0.25 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 \end{bmatrix}$$

Distribution of False Discoveries validation (Duivesteijn et al. 2011)

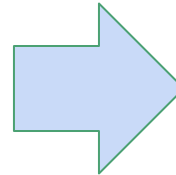
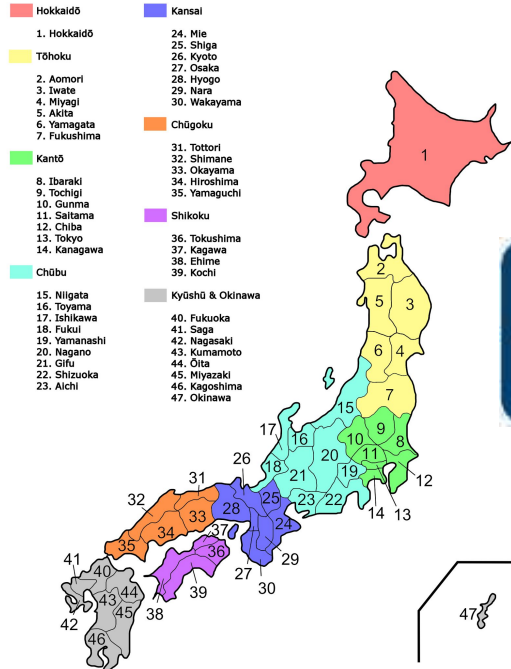
- 100 random datasets
- best score of all the subgroups per dataset



Case studies



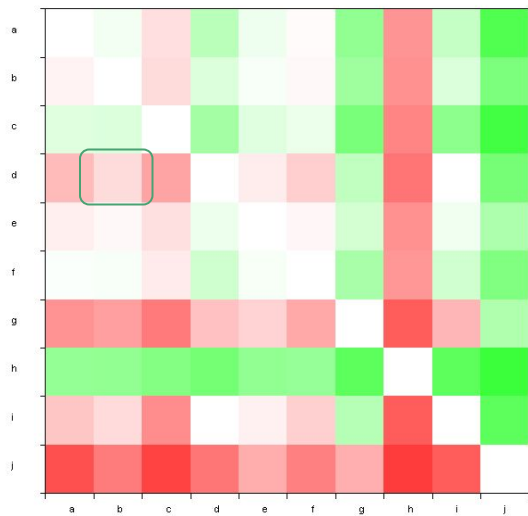
Regions and Prefectures of Japan



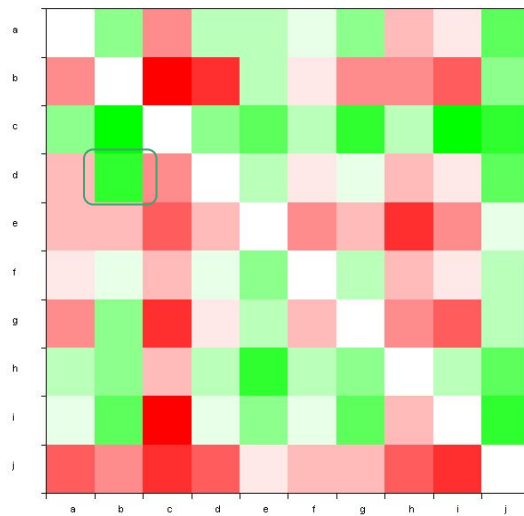
Label Ranking Matrix: (sex = '1' AND lives_in_prefecture = '13:Tokyo' AND lived_in_prefecture = '8:Ibaraki') c>h>i>a>d>f>g>b>e>j



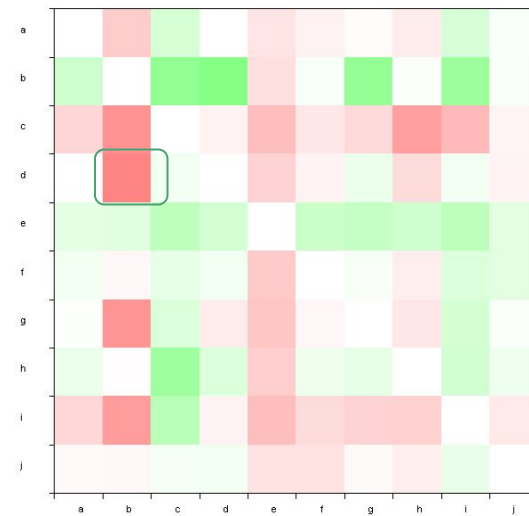
Base Matrix



Subgroup Matrix



Difference

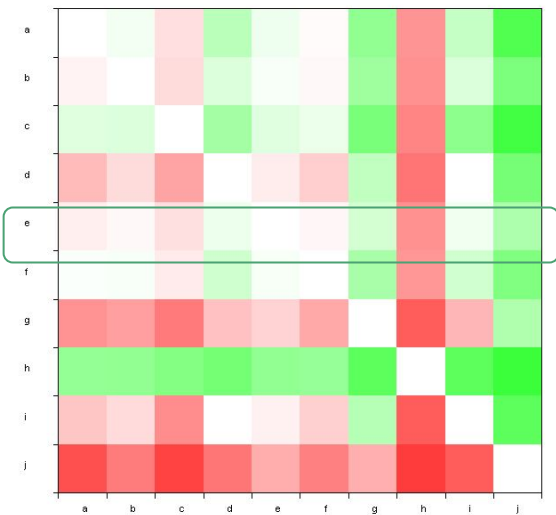


Sushi

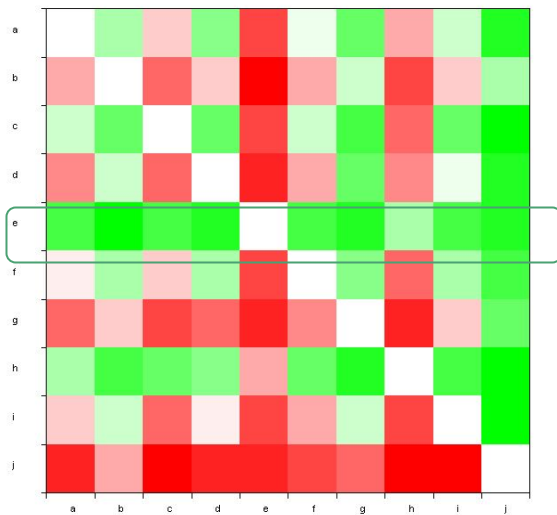
Label Ranking Matrix: (lived_in_prefecture = '0:Hokkaido' AND lives_in_region = '3:Kanto+Shizuoka' AND age >= 3.0) e>h>c>a>f>d>i>b>g>j



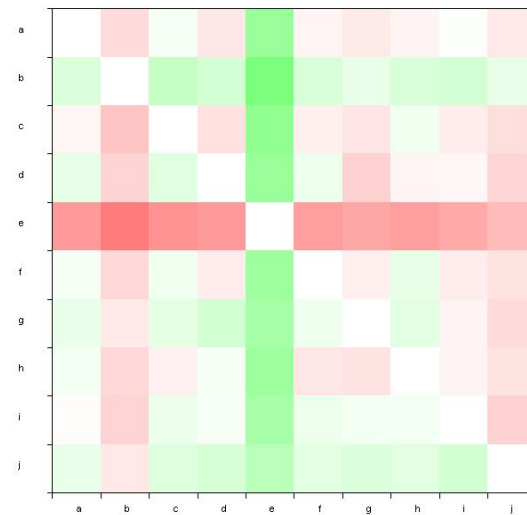
Base Matrix



Subgroup Matrix



Difference

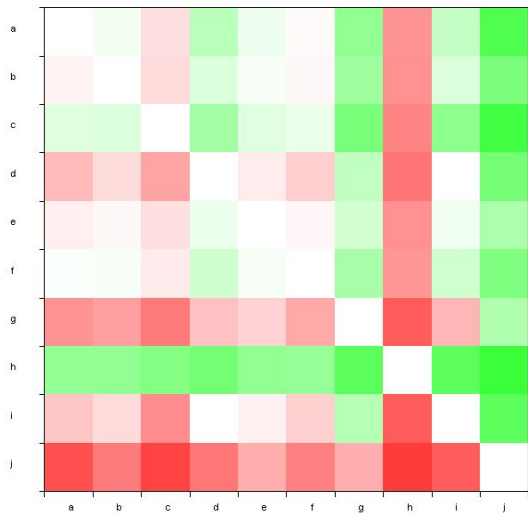


Sushi: Sea urchin

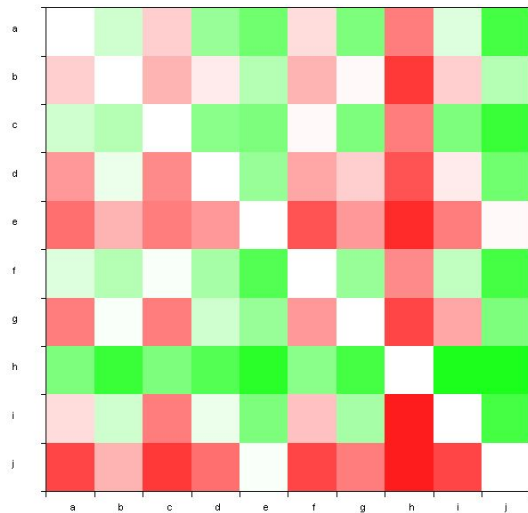
Label Ranking Matrix: (lives_in = '0:Eastern_Japan' AND lived_in_prefecture = '8:Ibaraki' AND age <= 1.0) h>c>f>a>i>d>b>g>e>j



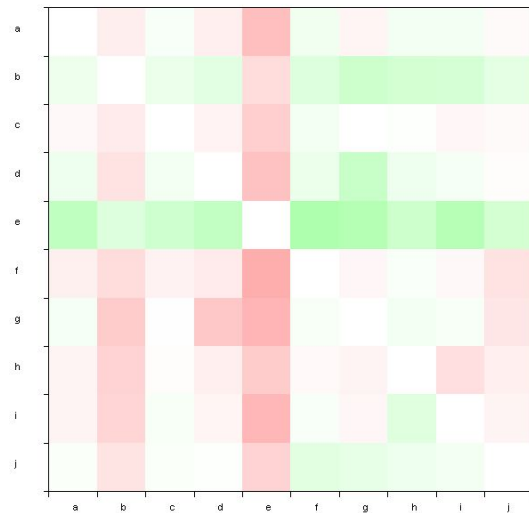
Base Matrix



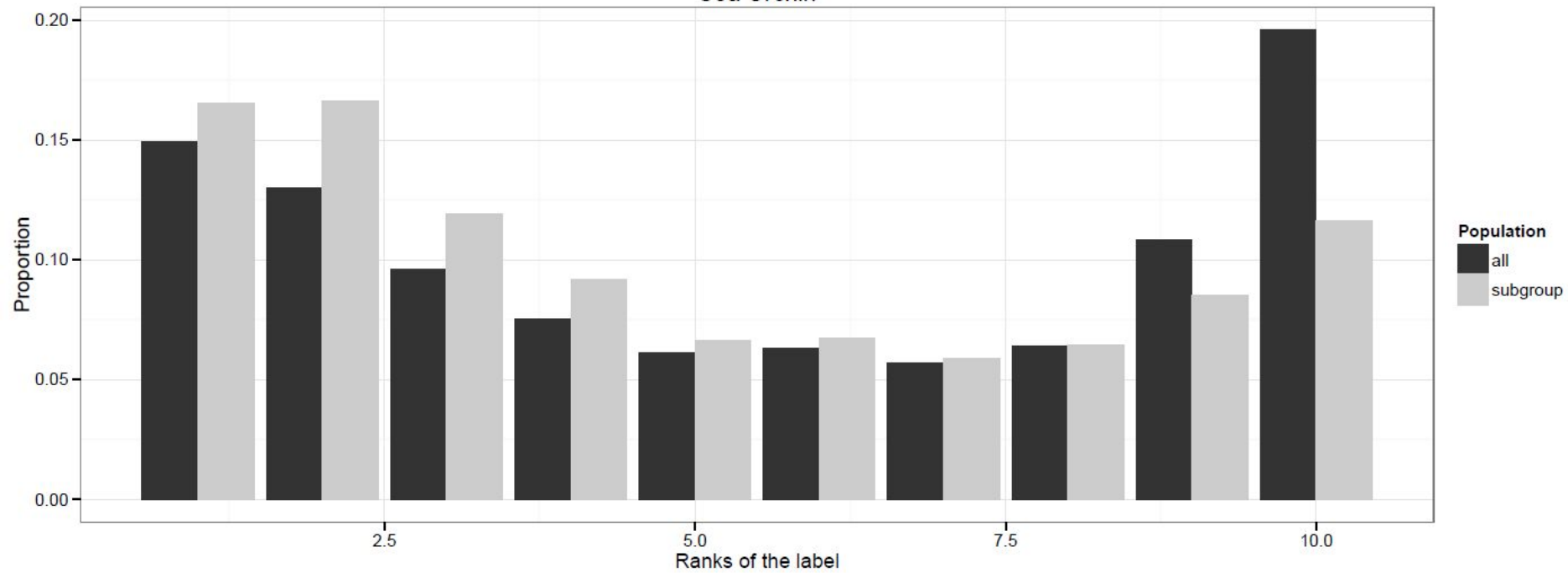
Subgroup Matrix



Difference



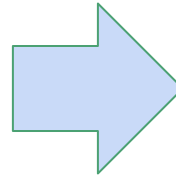
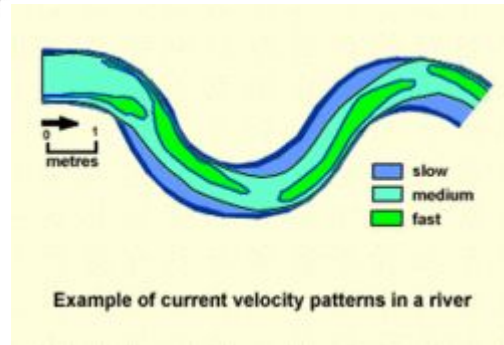
Sea Urchin

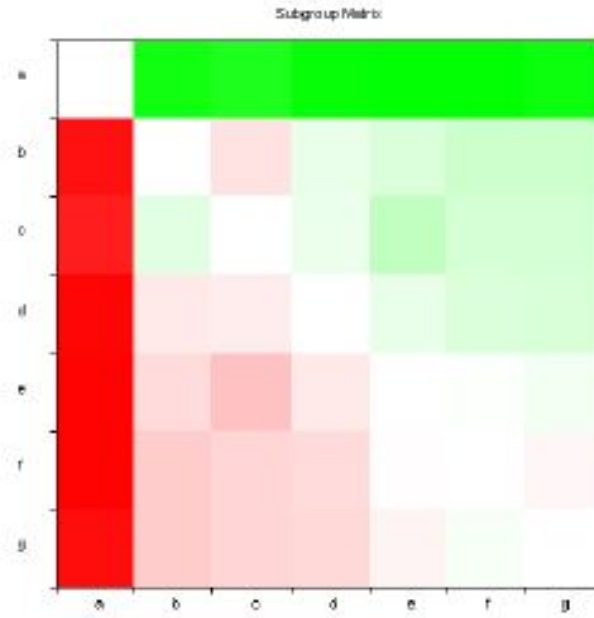
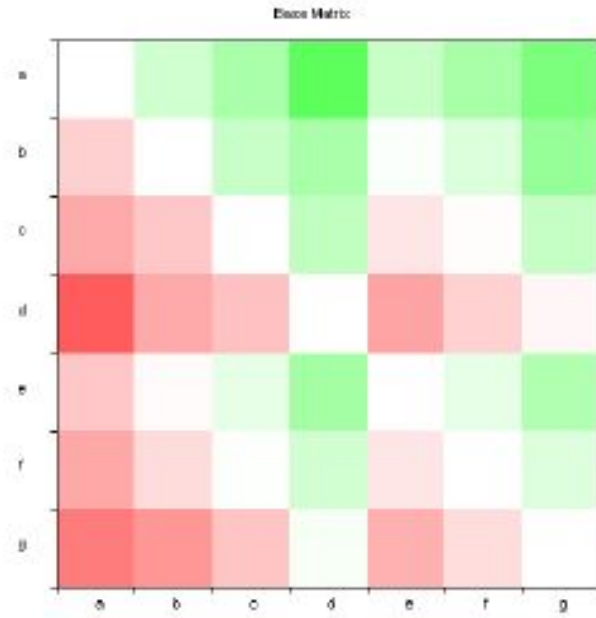




Sea Urchin

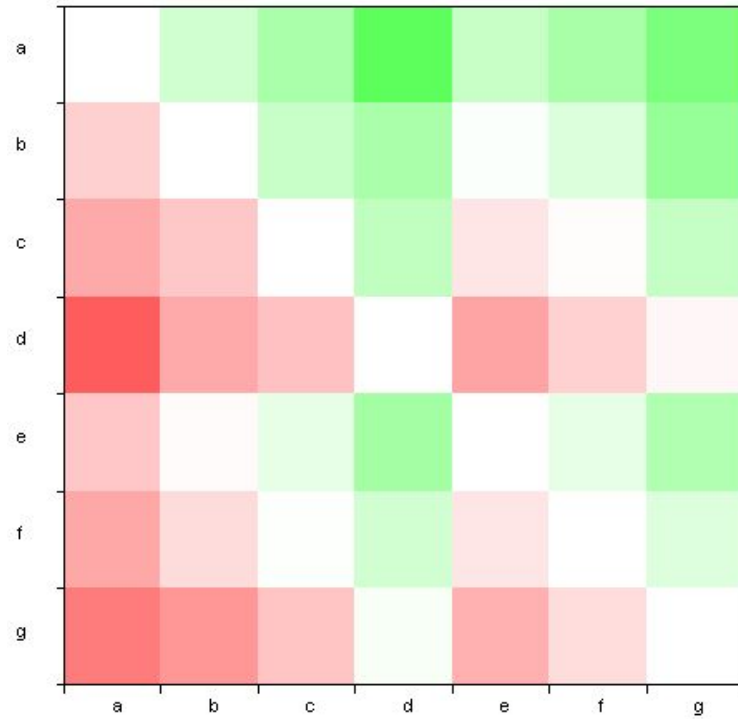




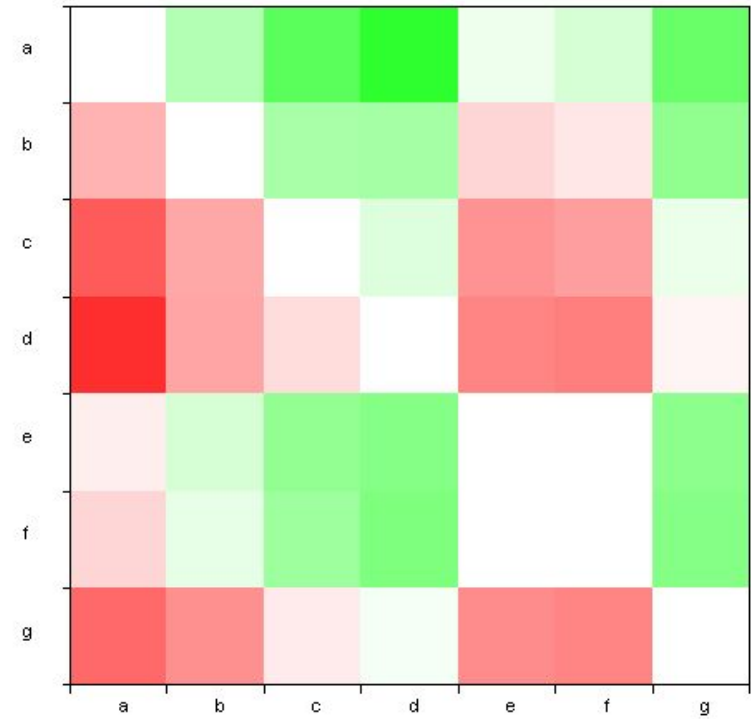


Algae: V10 \square 59 and V6 \square 11.867

Base Matrix

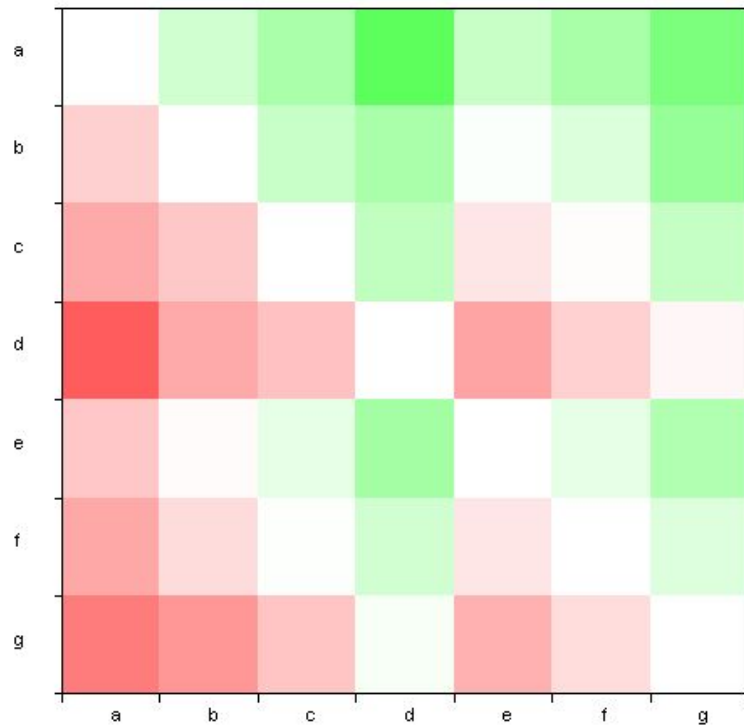


Subgroup Matrix

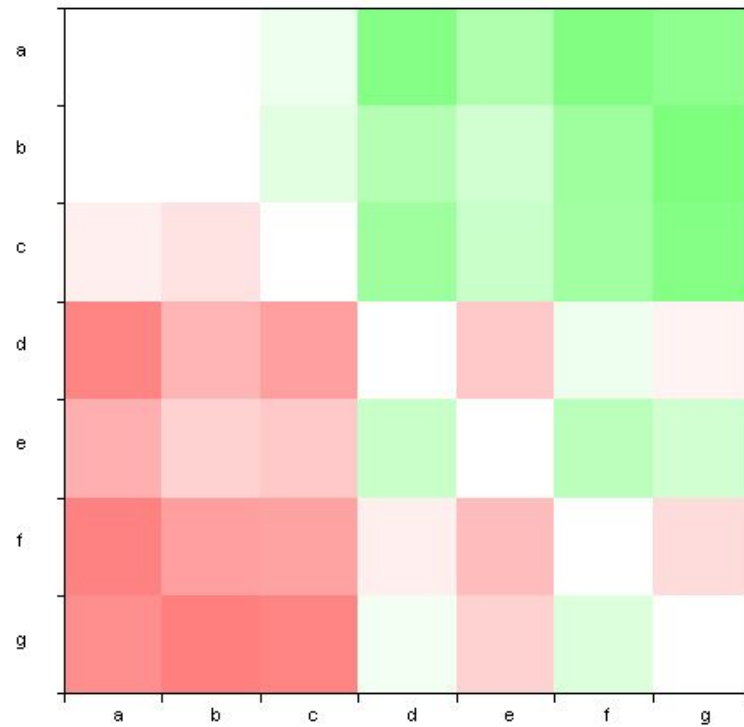


Algae: Season = Autumn

Base Matrix



Subgroup Matrix



Algae: Season = Spring



Age and education of the population, economic indicators (e.g., GDP growth, percentage of unemployment), indicators of the labor workforce in different sectors such as production, public service, etc. In terms of the target, the election results of the five major political parties for the federal elections in 2009 represented as rankings.

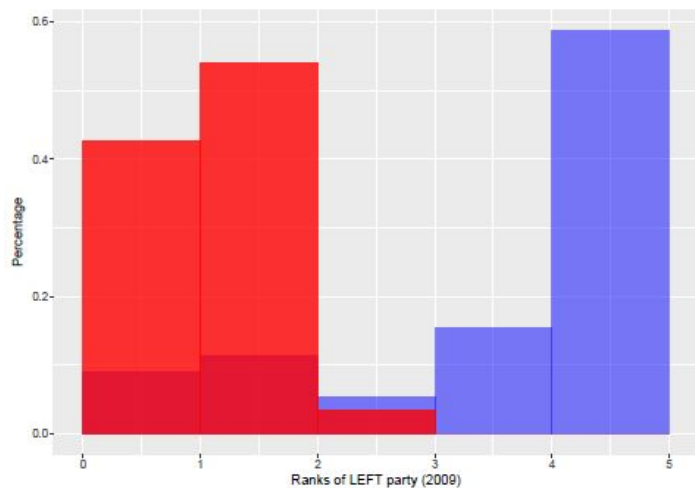


Fig. 2 Histograms representing the relative position of the LEFT party obtained in the 2009 elections of districts in Germany. In red, the subgroup Region = East and in blue the distribution for all districts.

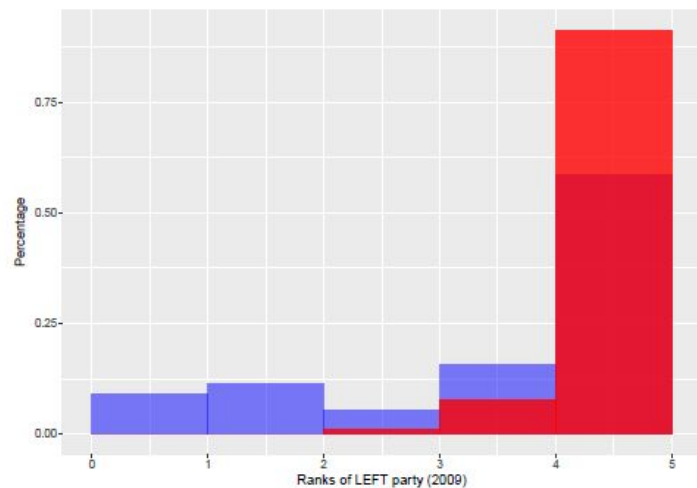
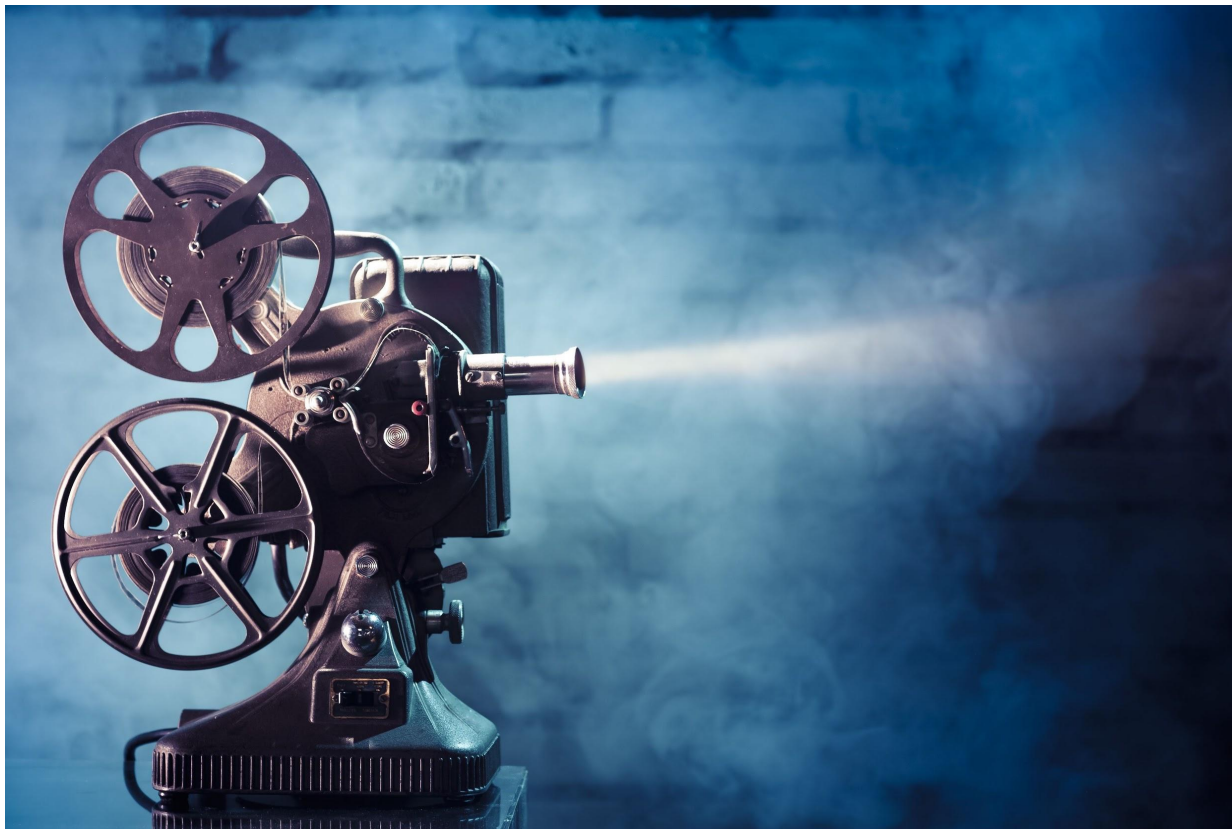


Fig. 3 Histograms representing the relative position of the LEFT party obtained in the 2009 elections of districts in Germany. In red, the subgroup Income ≥ 18442 and in blue the distribution for all districts.



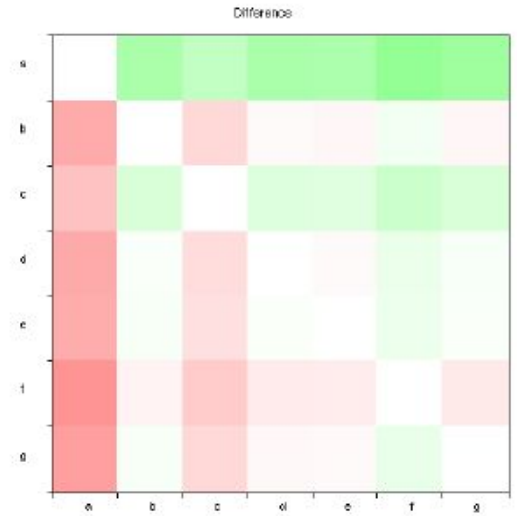
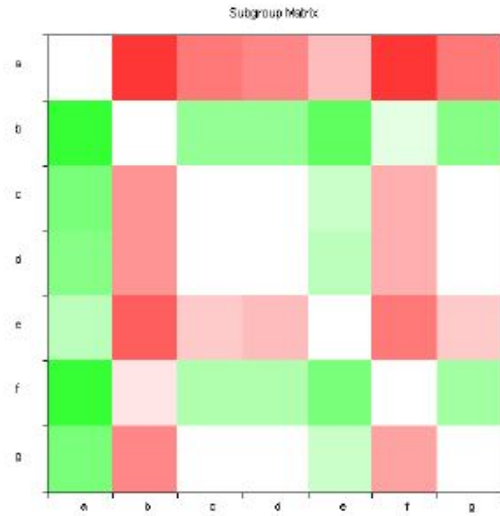
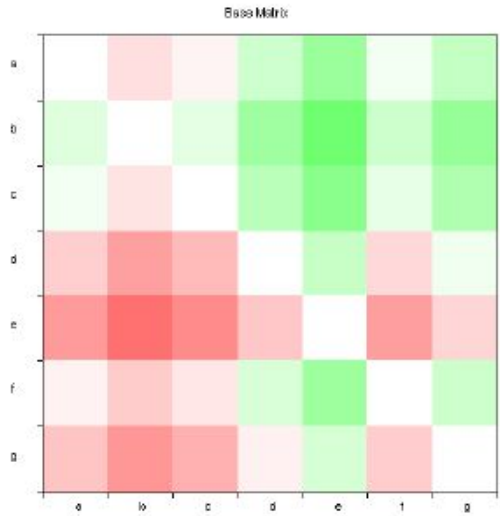


$CDU \succ \mathbf{LEFT} \succ SPD \succ FDP \succ GREEN$ (Thuringia)
 $\mathbf{LEFT} \succ \mathbf{SPD} \succ CDU \succ FDP \succ GREEN$ (Brandenburg)
 $\mathbf{LEFT} \succ CDU \succ SPD \succ FDP \succ GREEN$ (Saxony-Anhalt)
 $CDU \succ \mathbf{LEFT} \succ SPD \succ FDP \succ GREEN$ (Saxony)
 $CDU \succ SPD \succ FDP \succ \mathbf{GREEN} \succ LEFT$ (Bavaria)
 $CDU \succ SPD \succ FDP \succ LEFT \succ GREEN$ (All states)

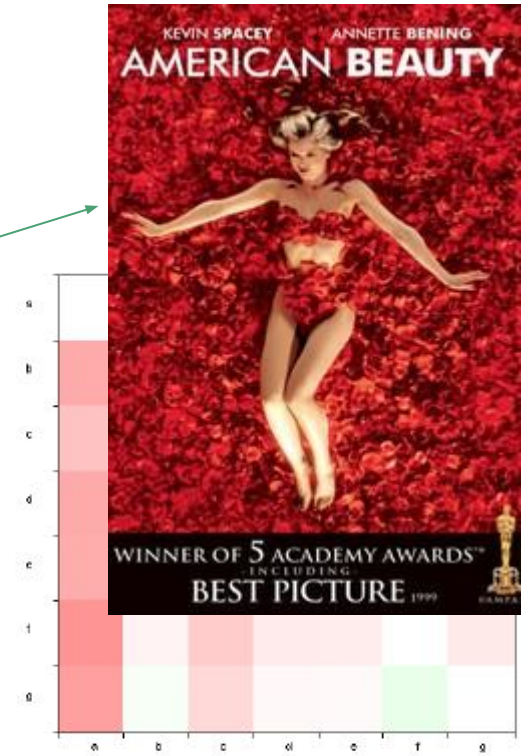
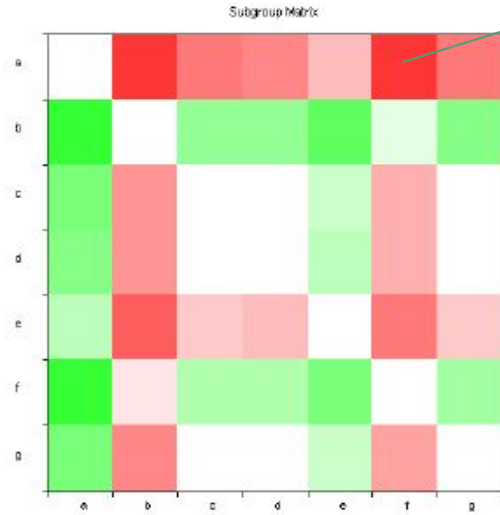
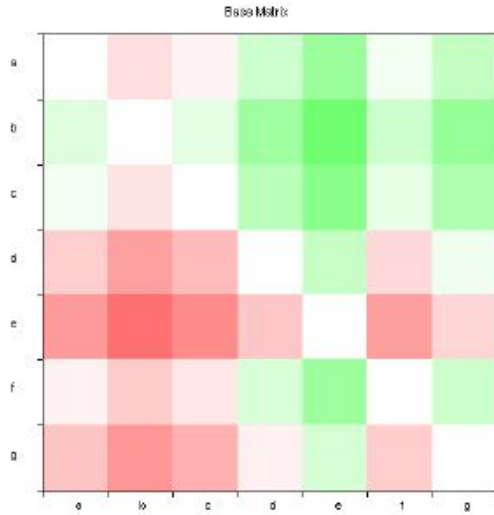


The Top7movies dataset is a subset of the [MovieLens 1M Dataset](#). The original dataset has 1 million ratings from 6000 users on 4000 movies. For each user, we have its demographic data, such as gender, age, occupation and zipcode.

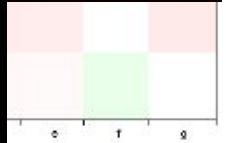
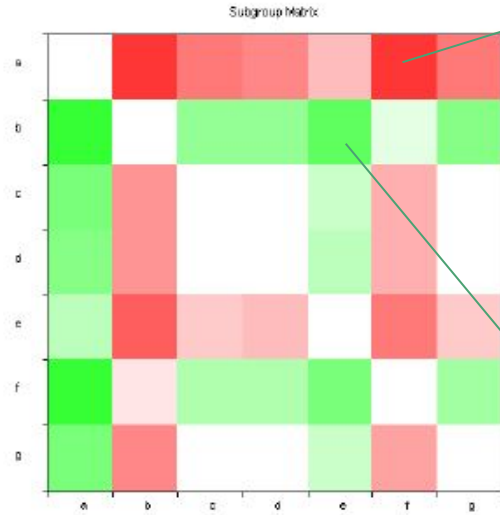
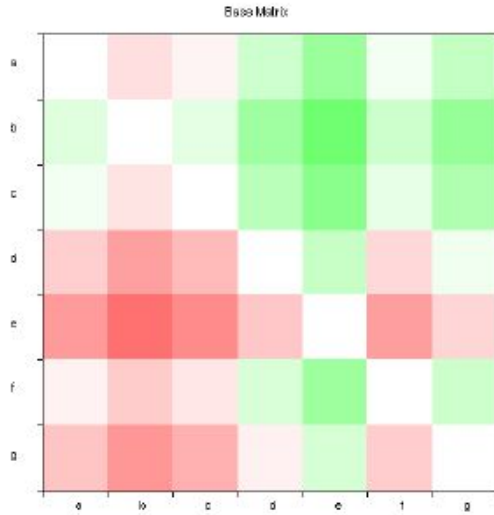
Subgroup: Age bigger than 34



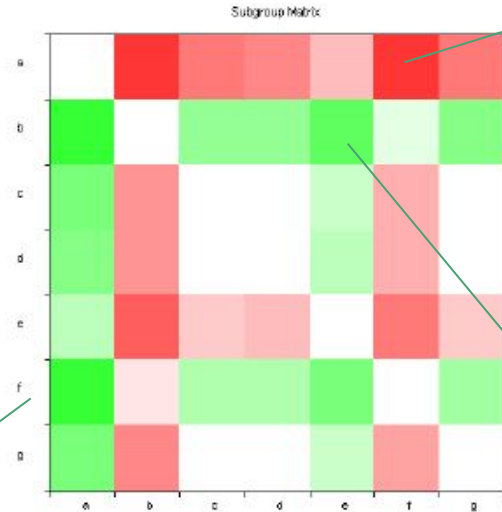
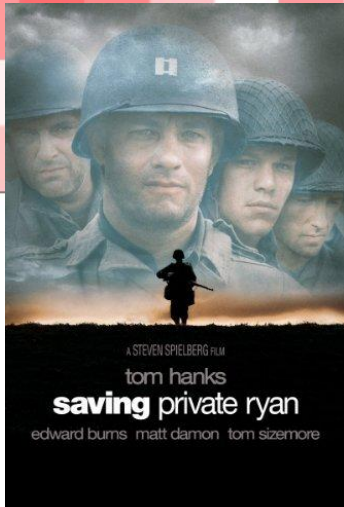
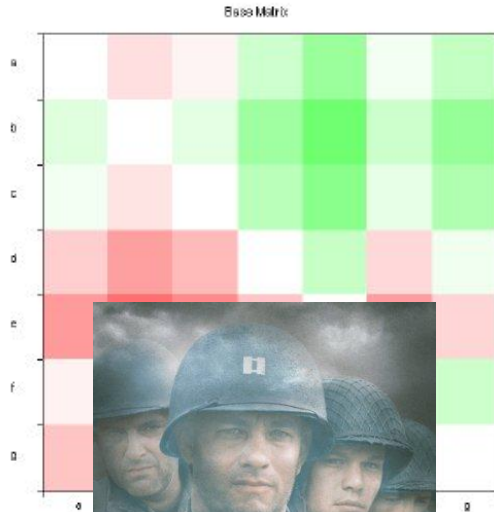
Subgroup: Age bigger than 34



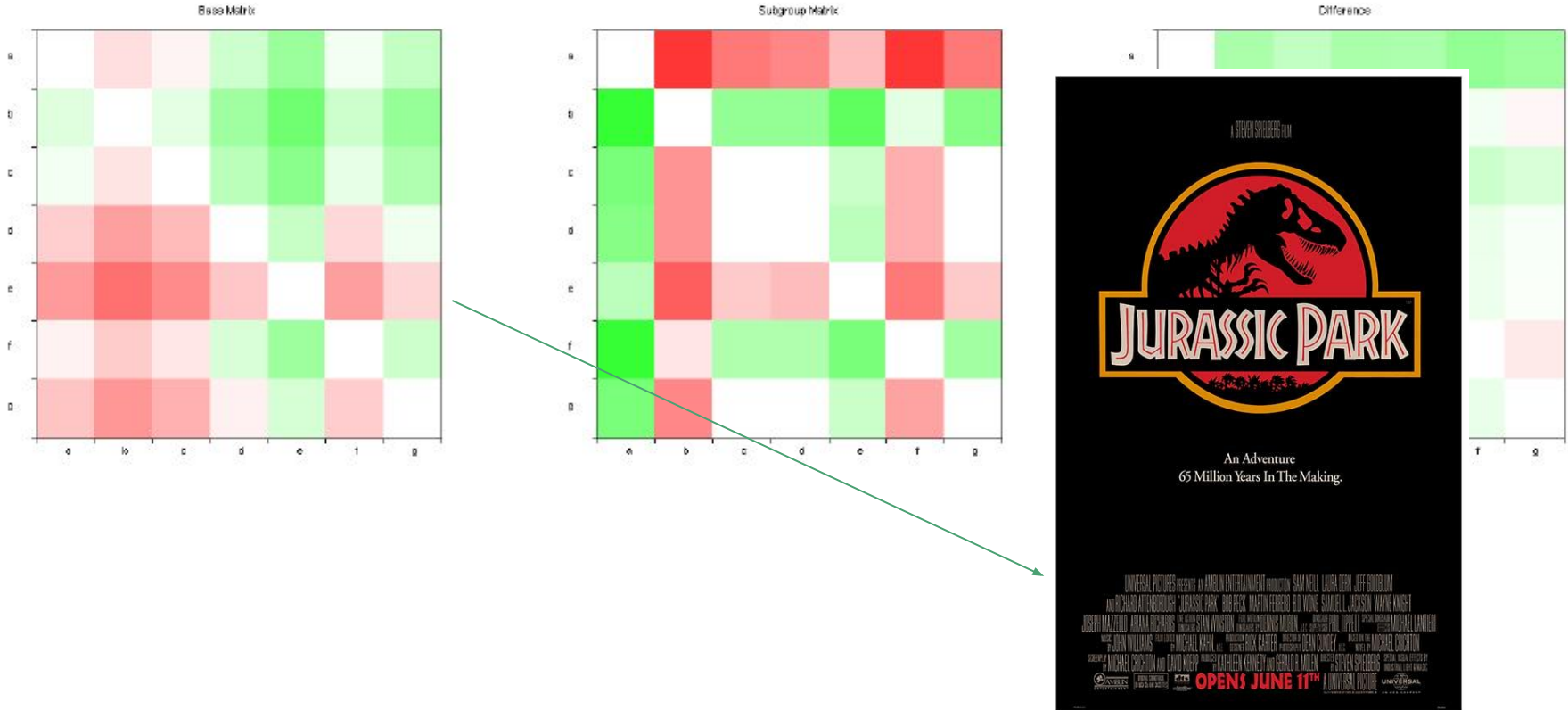
Subgroup: Age bigger than 34



Subgroup: Age bigger than 34



Subgroup: Age bigger than 34



Subgroup Discovery tools



Software

Package 'rsubgroup'

February 20, 2015

Type Package

Title Subgroup Discovery and Analytics

Version 0.6

Date 2014-09-10

Author Martin Atzmueller

Maintainer Martin Atzmueller <martin@atzmueller.net>

Description A collection of efficient and effective tools and algorithms for subgroup discovery and analytics. The package integrates an R interface to the org.vikamine.kernel library of the VIKAMINE system (<http://www.vikamine.org>) implementing subgroup discovery, pattern mining and analytics in Java.

Classification/ACM G.4, H.2.8, I.5.1

License GPL (>= 3)

Depends R (>= 2.10), methods, rJava (>= 0.6.3), foreign (>= 0.8.40)

SystemRequirements Java (>= 6.0)

Collate 'AAAnLoad.R' 'randomSeed.R' 'classes.R' 'subgroup.R'

URL <http://www.rsubgroup.org>

Repository CRAN

Repository/R-Forge/Project subgroup

Repository/R-Forge/Revision 51

Repository/R-Forge/DateTimeStamp 2014-09-10 15:40:17

Date/Publication 2014-09-11 11:00:24


NeedsCompilation no

[rsubgroup](#)

[Cortana](#)

Cortana

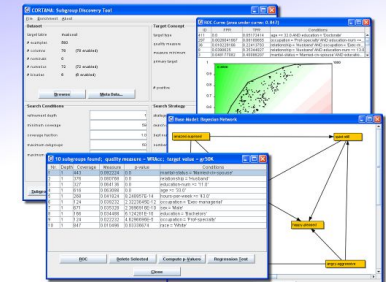
Open Source Subgroup Discovery



Universiteit Leiden

Cortana Subgroup Discovery

Cortana is a Data Mining tool for discovering local patterns in data. Cortana features a generic Subgroup Discovery algorithm that can be configured in many ways, in order to implement various forms of local pattern discovery. The tool can deal with a range of data types, both for the input attributes as well as the target attributes, including nominal, numeric and binary. A unique feature of Cortana is its ability to deal with a range of Subgroup Discovery settings, determined by the type and number of target attributes. Where regular SD algorithms only consider a single target attribute, nominal or sometimes numeric, Cortana is able to deal with targets consisting of multiple attributes, in a setting called *Exceptional Model Mining*.



Cortana's main features

- Generic parameterized Subgroup Discovery algorithm.
- Multiple data types supported.
- Implemented in Java, so works on all major platforms, including Windows, Linux and Mac OS.
- Works on propositional (tabular) data from flat files, .TXT or .ARFF.
- Includes Exceptional Model Mining settings.
- Statistical validation of mining results.
- Graphical presentation of results, such as ROC curves, scatter plots, and exceptional models.
- Additional bioinformatics module for literature-based gene set enrichment (see bioinformatics below).
- Free binary version and open-source access.

Download Cortana

Contacts:

claudio.r.sa@inesctec.pt

<https://www.linkedin.com/in/cláudio-rebelo-de-sá-92895562>