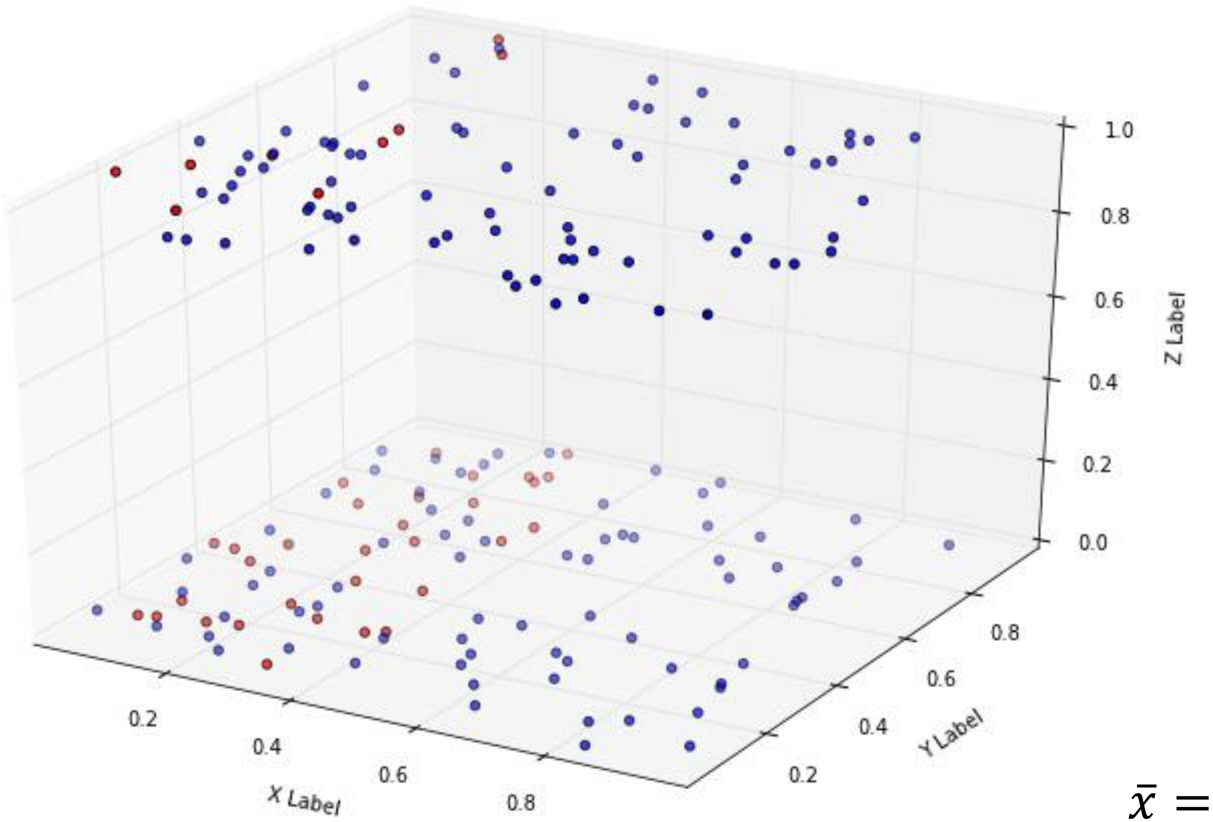
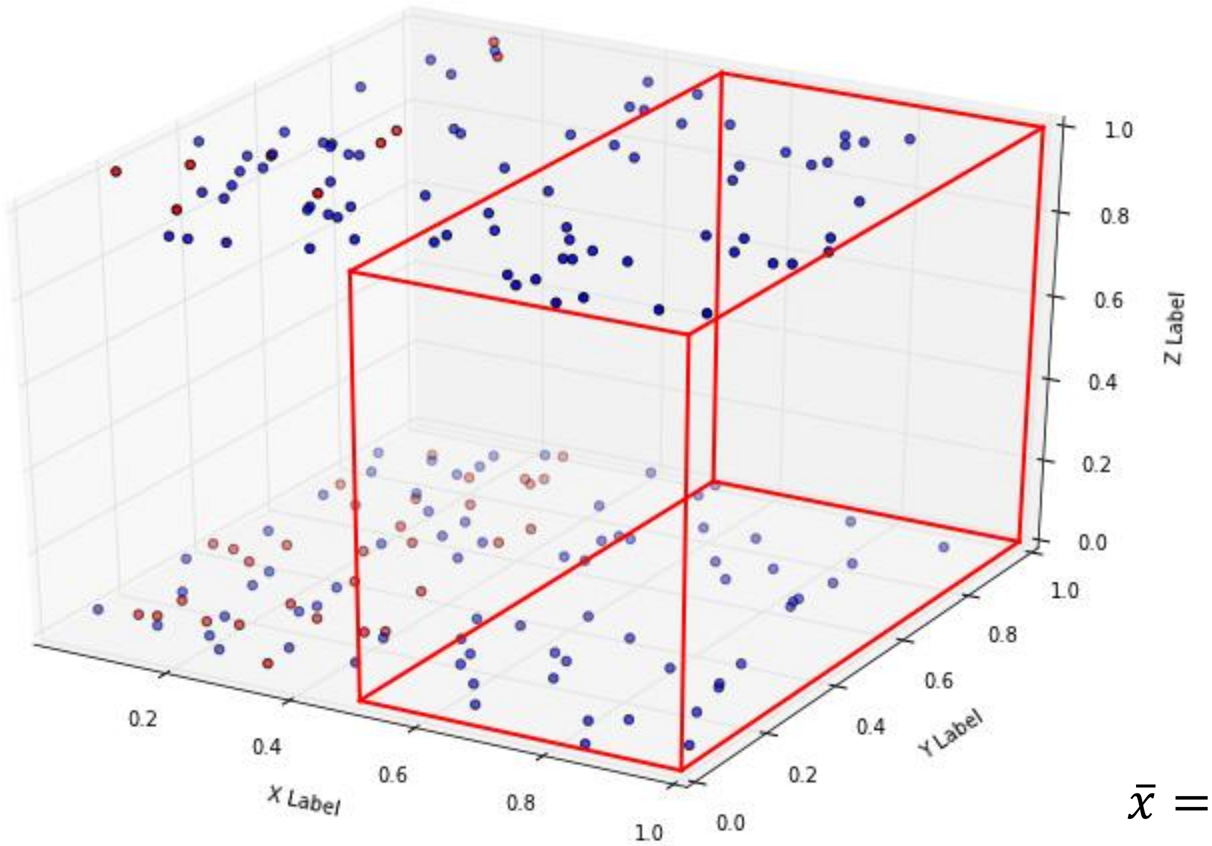
Scenario discovery in heterogeneously typed data

Dr.ir. Jan H. Kwakkel

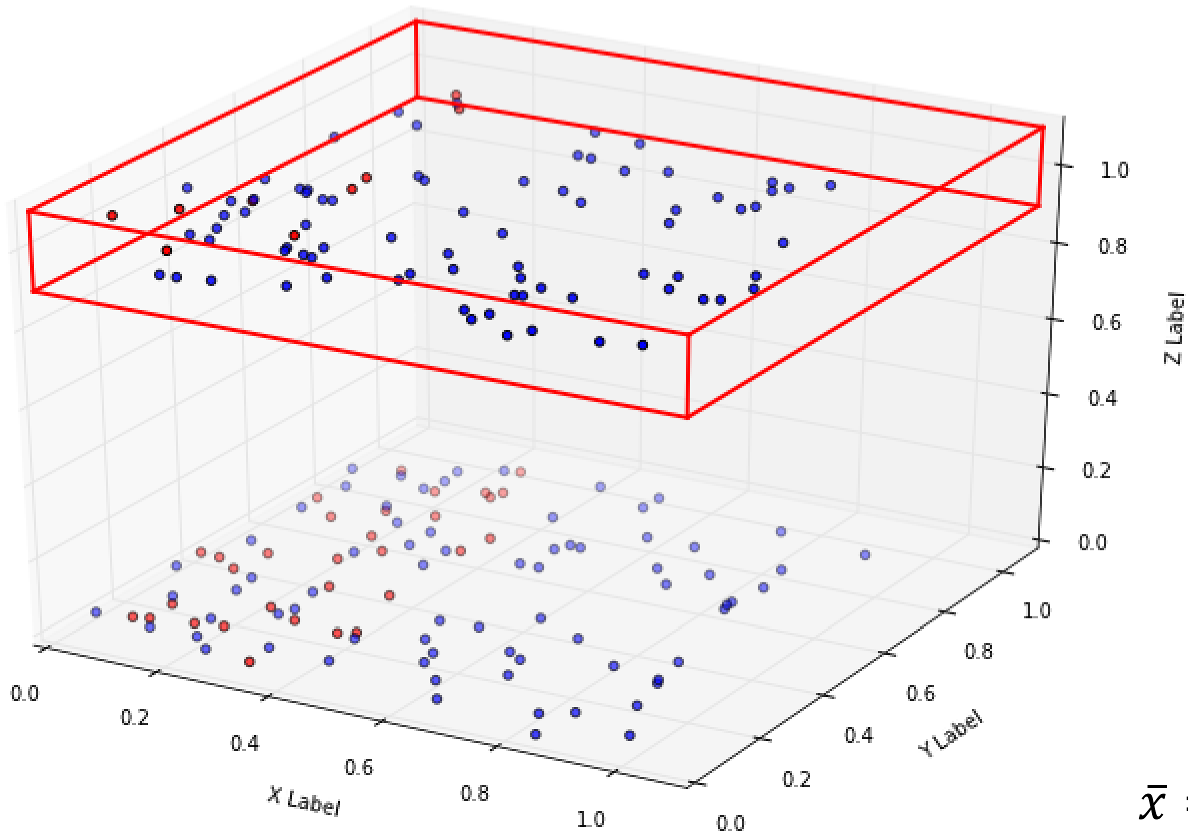
TU Delft



$$\bar{x} = \frac{40}{190} = 0.2105$$



$$\bar{x} = \frac{40}{115} = 0.3478$$



$$\bar{x} = \frac{31}{73} = 0.4247$$

Problem and Solution

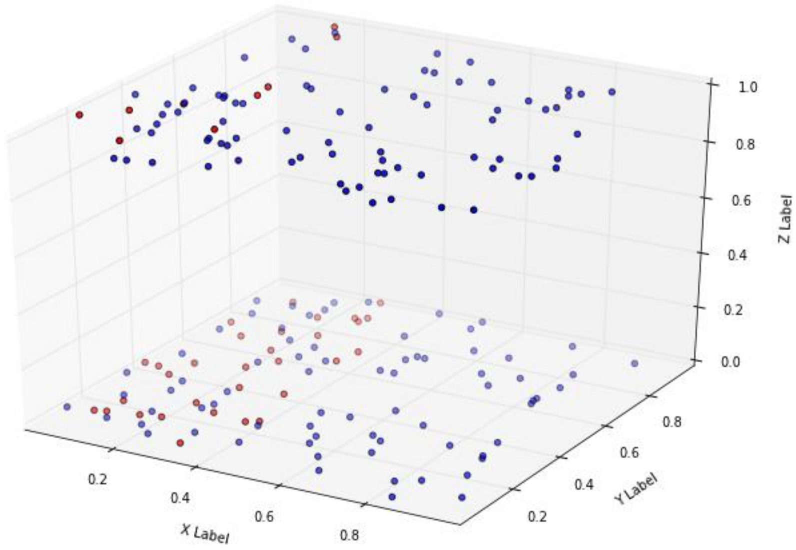
Problem

- PRIM aims at maximizing the mean of the data inside the box
- Lenient peeling only defined for floats

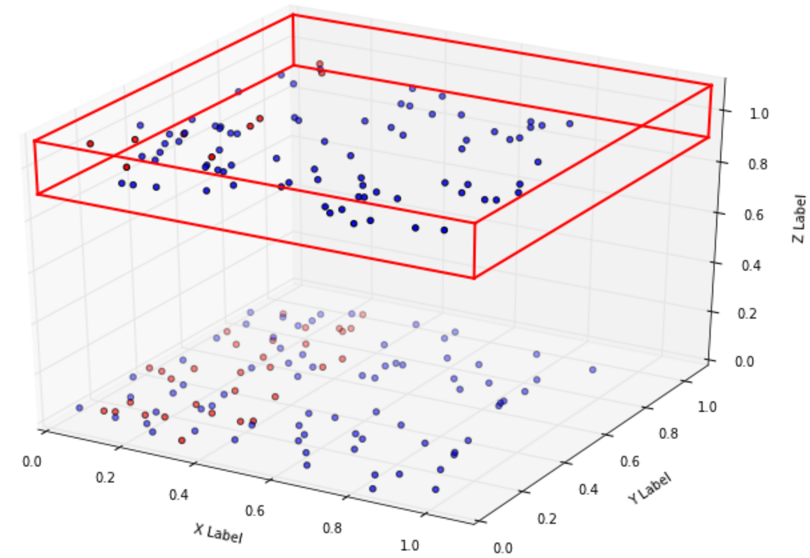
Solution

- In selecting next box, consider the gain in the mean offset by the loss in mass

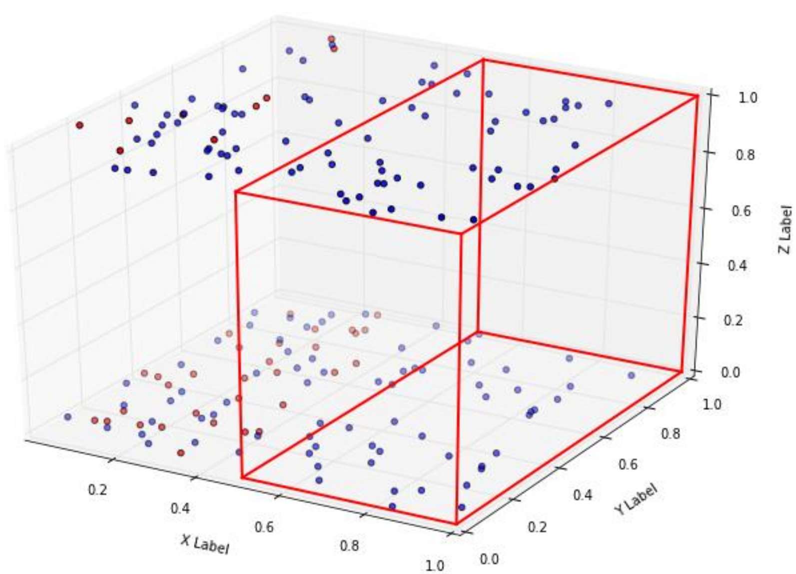
Note: the SD toolkit supports this more lenient criterion since its latest update



$$\bar{x} = \frac{40}{190} = 0.2105$$



$$\bar{f}_B = \frac{\frac{31}{73} - \frac{40}{190}}{190-73} = 0.0018301$$



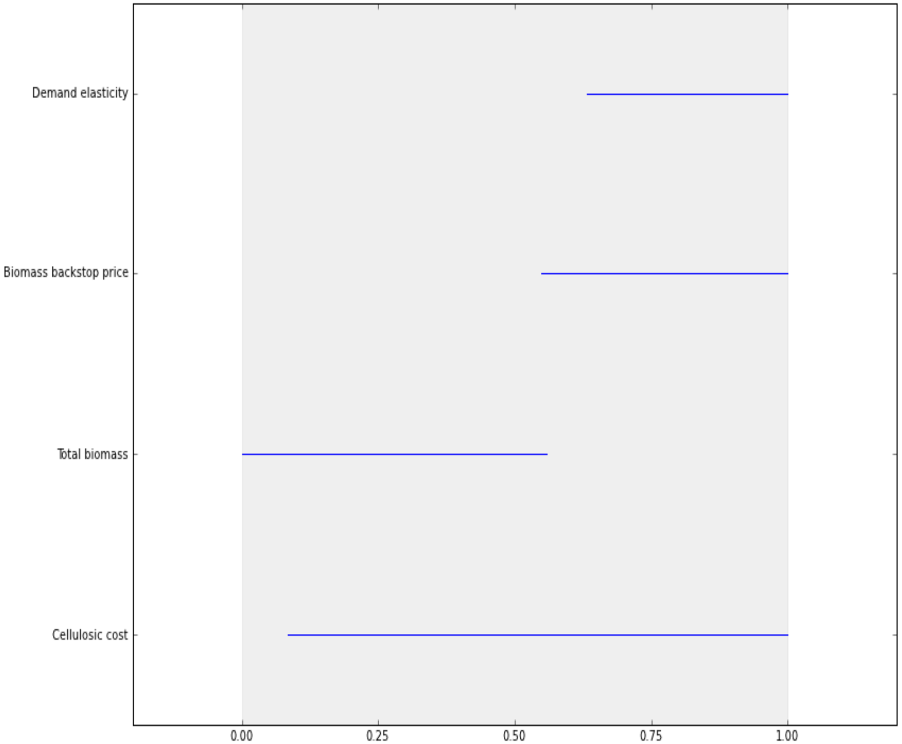
$$\bar{f}_B = \frac{\frac{40}{115} - \frac{40}{190}}{190-115} = 0.0018306$$

Bryant and Lempert (2010)

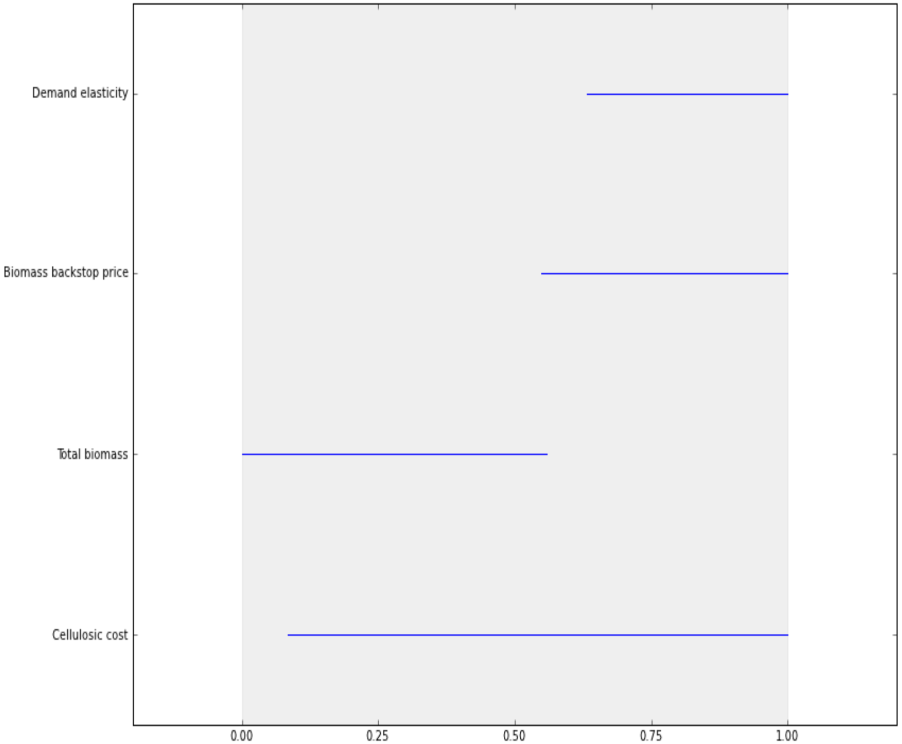
- 832 cases
- 89 cases of interest
- 8 uncertainties
- All uncertainties are floats

Bryant and Lempert (2010)

original criterion



lenient criterion

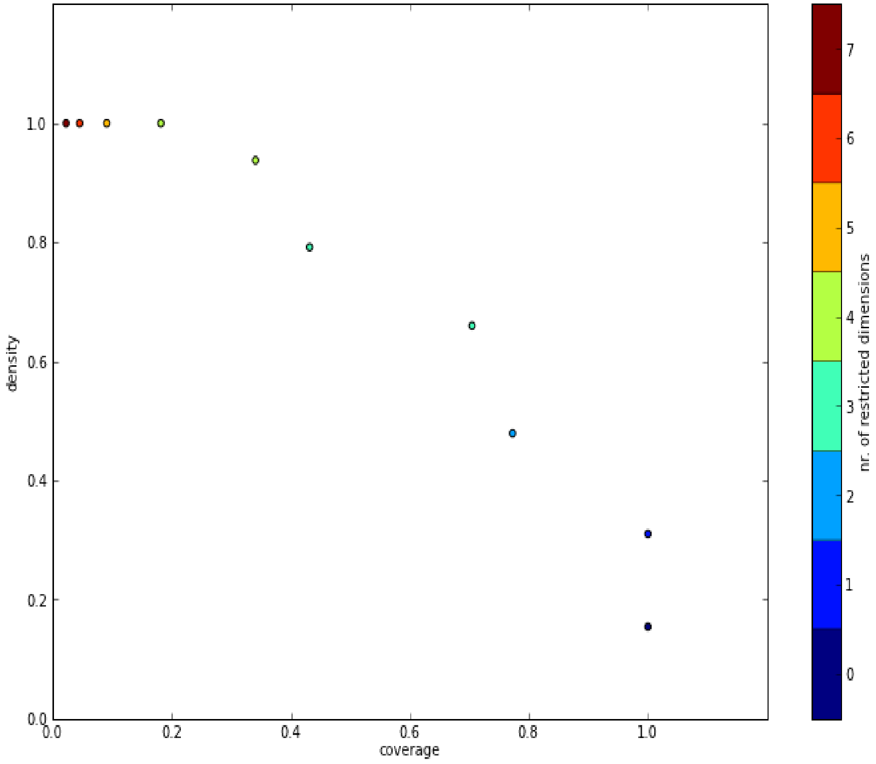


Rozenberg et al (2013)

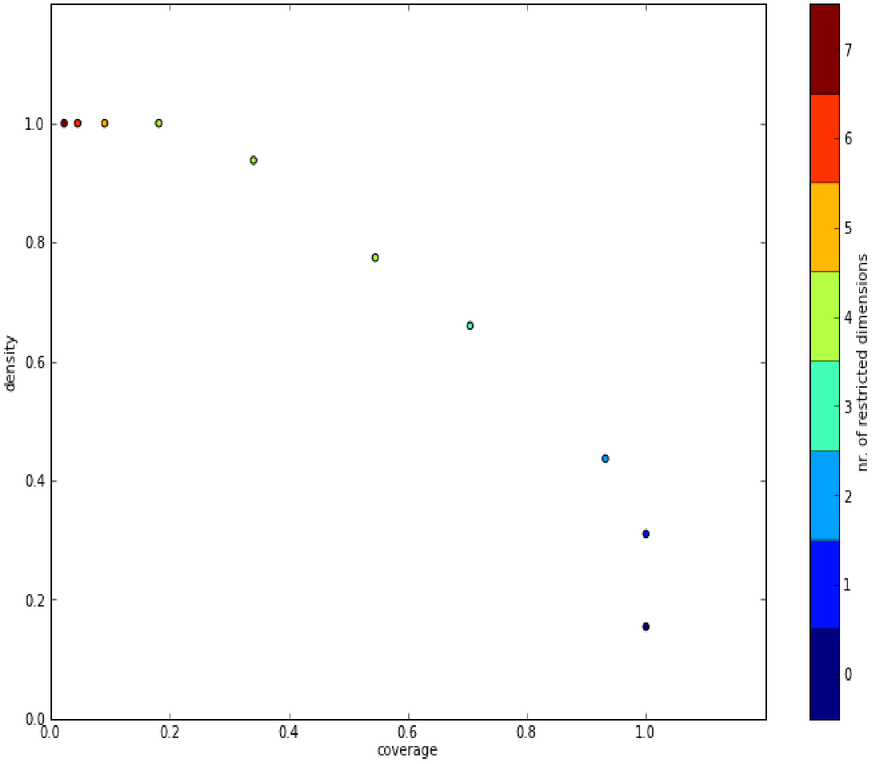
- 286 cases
- 44 cases of interest (SSP1)
- 7 uncertainties
- All uncertainties are integers

Rozenberg et al (2013)

original criterion



lenient criterion



Rozenberg et al (2013)

box	mass	coverage	density	res. dim.
4	0.054	0.43	0.79	3

	limits	qp-values
behaviors	1	9.14E-05
population	0	1.75E-03
inequalities	0	1.03E-02

box	mass	coverage	density	res. dim.
4	0.11	0.55	0.77	4

	limits	qp-values
behaviors	1	3.40E-06
population	0	3.79E-03
inequalities	0-1	8.40E-03
convergence	1-2	7.22E-02

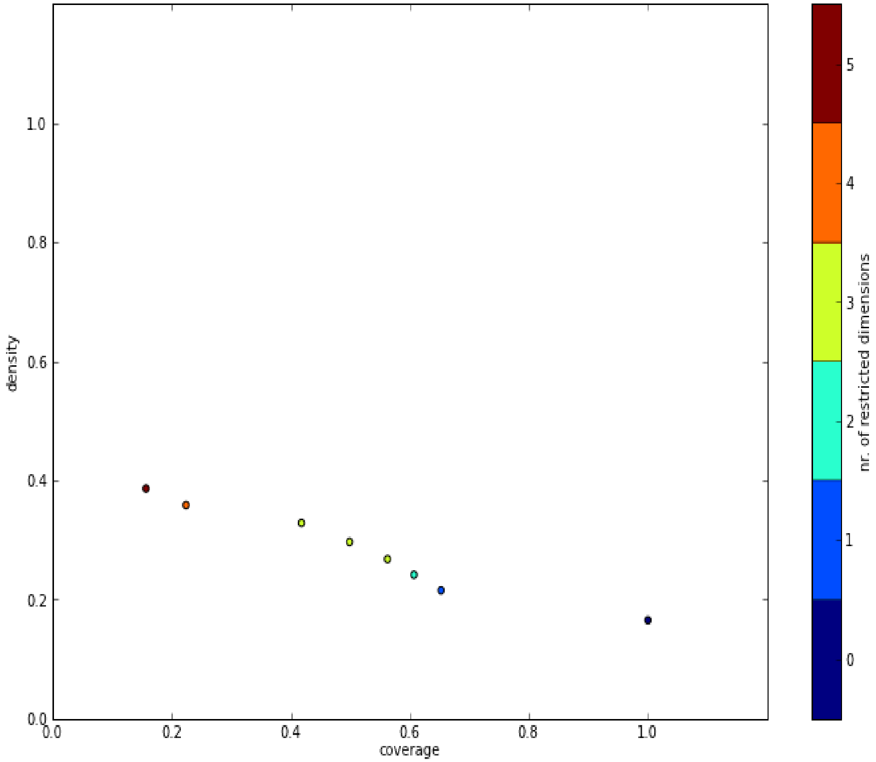
Hamarat et al (2014)

- Data characteristics

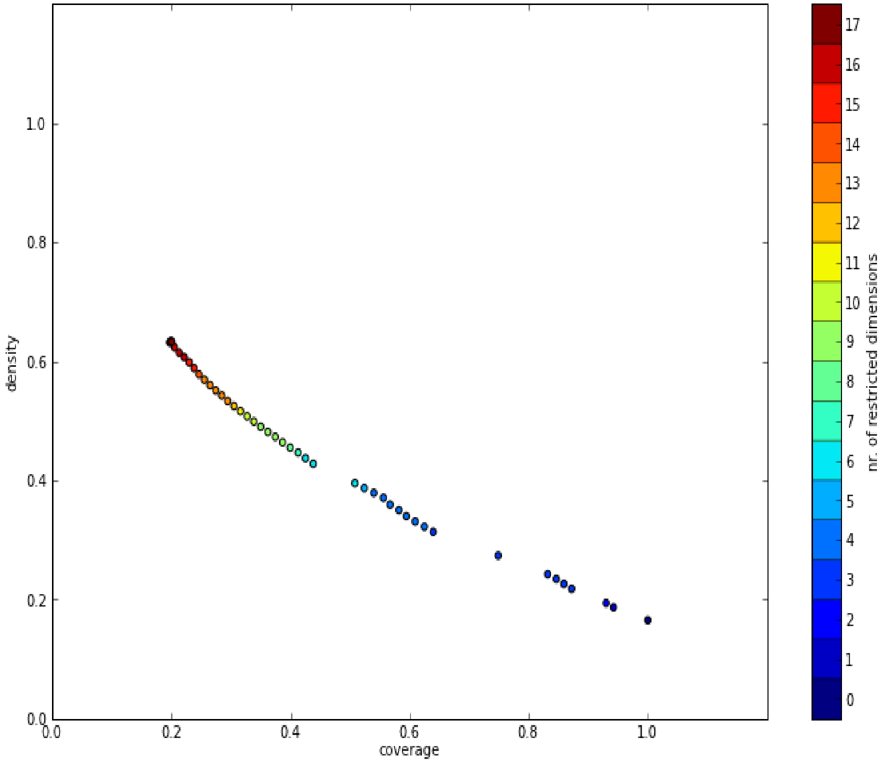
- 20.000 cases
- 3298 cases of interest
- 48 uncertainties
  - 36 uncertainties are floats
  - 12 uncertainties are categorical

Hamarat et al (2014)

original criterion



lenient criterion



Hamarat et al (2014)

box	mass	coverage	density	res. dim.
5	0.21	0.42	0.33	3

	limits	qp-values
SWITCH electrification rate	1, 2, 5	3.98E-37
SWITCH physical limits	1	9.82E-17
SWITCH economic growth	1, 2, 3, 4, 5	2.67E-05

box	mass	coverage	density	res. dim.
15	0.24	0.54	0.38	4

	limits	qp-values
SWITCH electrification rate	1, 2, 5	4.24E-79
progress ratio wind	0.89-1.00	8.87E-22
time of nuclear power plant ban	2029.02-2100	1.78E-14
SWITCH economic growth	1, 2, 3, 4, 5	7.22E-02

Scenario Discovery and multiclass data

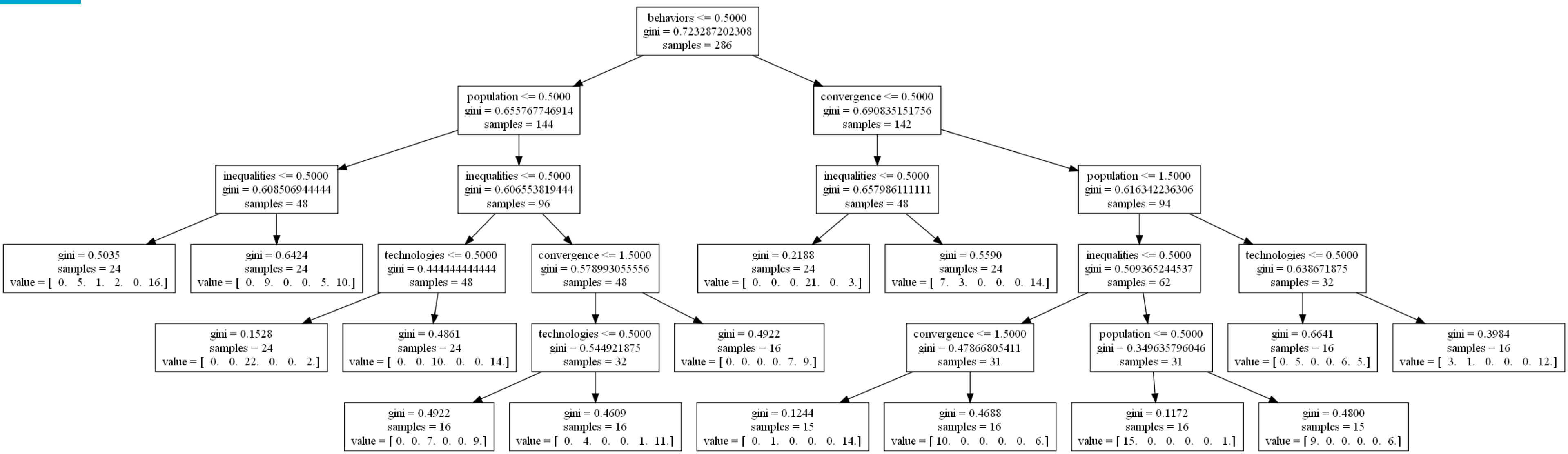
- Problem:
  - Multiclass instead of binary classification for dependent variable

- Possible solutions
  - CART (Gerst et al (2013)
  - Iterated PRIM (Rozenberg et al (2013))

- Idea
  - Adapt PRIM to use GINI impurity as objective function

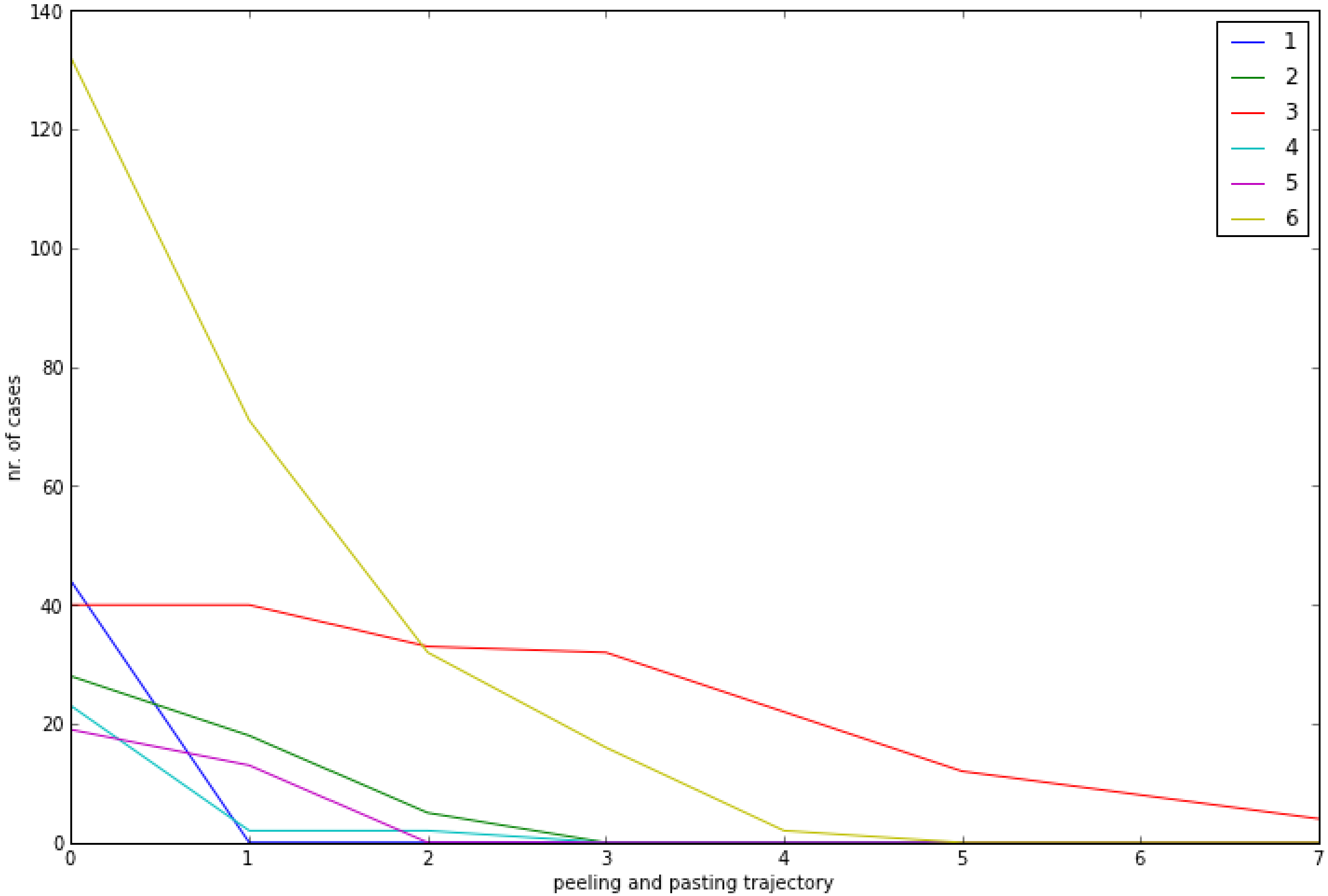
$$\bar{f_B} = 1 - \sum_{i=1}^{m} {f_i}^2$$

CART



- 15 boxes
- No overlap
- Few pure boxes

GINI PRIM



140
120
100
80
60
40
20
0
nr. of cases
0
1
2
3
4
5
6
7
peeling and pasting trajectory
1
2
3
4
5
6

GINI PRIM

box	box composition						mass	res. dim.
	1	**2**	**3**	**4**	**5**	**6**		
1	0	0	32	0	0	16	0.17	3
2	0	5	1	2	0	16	0.084	2
3	0	13	7	0	13	39	0.25	1
4	19	0	0	0	0	5	0.084	2
5	1	1	0	0	1	10	0.045	4
rest	24	9	0	21	5	46	0.4	0

uncertainty	box 1	box 2	box 3	box 4	box 5	rest box
behaviors	0	0	0			0-1
inequalities	0	0		1	1	0-1
population	1-2			0		0-2
capital markets					1	0-1
convergence				1-2		0-2
technologies					0	0-1

Closing remarks

- In case of heterogeneous data types for the independent variables, using the more lenient criterion is worthwhile

- GINI PRIM appears interesting but needs work
  - How to define coverage and density in the multiclass case?
  - What about overlap between classes?
  - How to visualize results and make trade offs?

- What about PCA preprocessing?
  - PCA is only defined for floats