Laboratory for
information systems
Rudjer Boskovic Institute, Croatia

Descriptive modeling in social sciences

Dragan Gamberger
Rudjer Boskovic Institute, Croatia

11.7.2013.

Motivation

FP7 project FOC "Forecasting financial crises"

http://www.focproject.net/

- *IMF Working paper (2008)* Systemic Banking Crisses: A New Database [L.Laeven, F.Valencia] (updated June 2012)

defines147 systemic banking crises in the period 1976-2011.

(e.g. China 1998, USA 1988 and 2007)

- **World bank data about countries:**

current account balance as percentage of GDP,

central government debt as percentage of GDP,

domestic credit to private sector as percentage of GDP,

foreign direct investments as percentage of GDP,

bank capital to assets ratio

....

percentage of rural population,

life expectancy at birth,

percentage of unemployment with tertiary education,

Which properties are characteristic for countries having banking crises ?

<u>descriptive modeling</u>

Banking crises dataset

Examples

country 1
country 2
country 3
..
country 147
country 148
..
country 434

Attributes
(WB data)

2.1	13.2	0.7	1.1	...	crisis
2.5	11.9	1.3	4.0	...	crisis
2.7	9.7	2.7	?	....	crisis
7.7	18.2	?	1.4	...	crisis
2.1	1.0	1.3	2.0	...	non-crisis
4.0	2.7	2.7	1.1	....	non-crisis
					...

945 numerical attributes

147 positive cases
287 negative cases

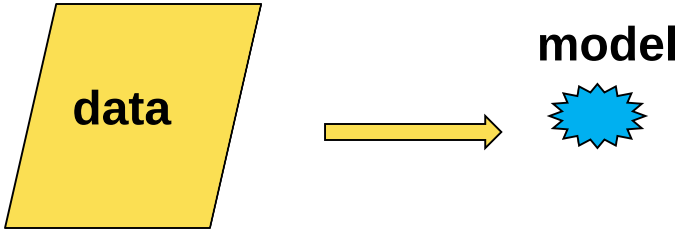
105 indicators

for each indicator a period of 3 years *before* the crisis or non-crisis

_t_3
_t_2
_t_1
_max
_index_max
_min
_index-min
_average
_slope

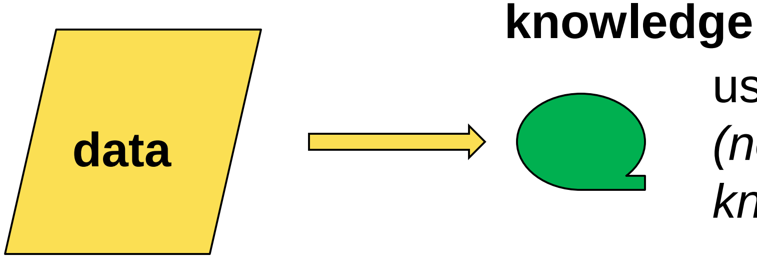
9 attributes for each indicator

Descriptive <-> Predictive modeling



used for:
(automatic) classification of unclassified data

evaluated by:
predictive quality on unseen examples (objective measure)



used as:
(novel) human knowledge

evaluated by:
novelty
actionability
interestness
...
(subjective measures of human expert)

Subgroups

1: *Fast growing credit activity in economies with aging population*
slope of credits in the period of three years before crisis > 5.8 % per year
life expectancy for females in the year before the crisis > 80.2 years.

2: *High credit activity in economies with high social security*
under-five mortality rate in the period of three years before crisis < 6.3 (per 1000)
population ages 65 and more three years before the crisis > 14.2 % of total population.

3: *Increasing credit activity in developing economies*
increasing credit activity in the period of three years before the crisis
population aged 15-65 one year before the crisis < 64.3 % of total population
rural population three years before the crisis < 33.7 % of total population

4: *Socioeconomic problems recognized by decreasing life expectancy*
slope of life expectancy for females in the period of three years before crisis < -0.3 years per year

5: *Socioeconomic problems recognized by non-increasing quality of public health*
non-increasing life expectancy for females in the period of three years before crisis
under-five mortality rate in the period of three years before crisis > -0.5 (per 1000)

Subgroups

1: *Fast growing credit activity in economies with aging population*

slope of credits in the period of three years before crisis > 5.8 % per year
life expectancy for females in the year before the crisis > 80.2 years.

Supporting conditions: low mortality of children, low percentage of young population, high percentage of elderly population, high capitalization of companies.

List of 16 included crises: Sweden in year 1991, USA and UK in year 2007, Belgium, Denmark, France, Greece, Ireland, Island, Italy, Luxemburg, Netherlands, Portugal, Slovenia, Spain and Sweden in year 2008.

5: *Socioeconomic problems recognized by non-increasing quality of public health*

non-increasing life expectancy for females in the period of three years before crisis
under-five mortality slope in the period of three years before crisis > -0.5 (per 1000)

Supporting conditions: high money and quasi money growth before the crisis.

List of 25 included crises: Sierra Leone in year 1990, Finland, Liberia, Nigeria, Norway, and Sweden in year 1991, Kenya and Poland in year 1992, Burundi in year 1994, Belarus, Central African Republic, Latvia, Lithuania, Swaziland, and Zimbabwe in year 1995, Bulgaria in year 1996, Ukraine in year 1998, Uruguay in year 2002, Belgium, Greece, Hungary, Island, Italy, Portugal, and Spain in year 2008.

Comments

A) Subgroup discovery approach does segmentation of the target set of examples and the methodology is useful when the positive class is a result of a few different models. Especially if these models have condratictory conditions.

B) Rules (including subgroup descriptions) are constructed as conjunctions of **features.**

Example:

1: Fast growing credit activity in economies with aging population

slope of credits in the period of three years before crisis > 5.8 % per year

life expectancy for females three years before the crisis > 80.2 years.

A feature-based view as a unifying framework for rule induction is perhaps a most distinguishing chracteristic of the book !

Examples are defined by attributes

attributes

examples

NAME	AGE	SEX	EDU	PROF	WEIGHT	INCOME	SMOKER
peter	30	male	low	worker	27.3	14000	yes
carl	55.5	male	medium	worker	90	20000	no
dora	?	female	high	teacher	65.2	1000	no
tanja	18	female	medium	student	55.1	0	no
tom	70	male	high	?	60	9000	yes
steve	35	male	medium	prof	33	16000	no
mirko	42.2	male	low	driver	27	7500	yes
marc	29	male	?	waiter	31	8300	yes

nominal
(categorical)

numerical

Features

Features:

Income > 1000
Slope of credits < 5.5

For each attribute many different features may be constructed !

The first step of the rule induction process is feature construction.

Features may have only values true and false.

Features are different from binary attributes.

Features may not have unknown values.

Features may be complex in the sense that they may include information from more than one attribute or represent information from a relational database.

Why features are so important ?

- There is a well-defined procedure how to construct features.

- Once the features are constructed, the rule construction process is identical regardless of the type of attributes, how features have been obtained and what is their meaning.

- Feature relevancy is well defined. It enables that irrelevant features may be immediately discarded and that only really relevant features enter the rule induction process.

- Unknown attribute values may be solved in a very systematic way in the feature construction process.

- Imprecise attribute values can be effectively handled.

- Cut-off values in the conditions of features used in rule bodies present a valuable information. They are also the basis for the transformation of subgroups into risk models.

Handling imprecision of numerical attributes as unknown values

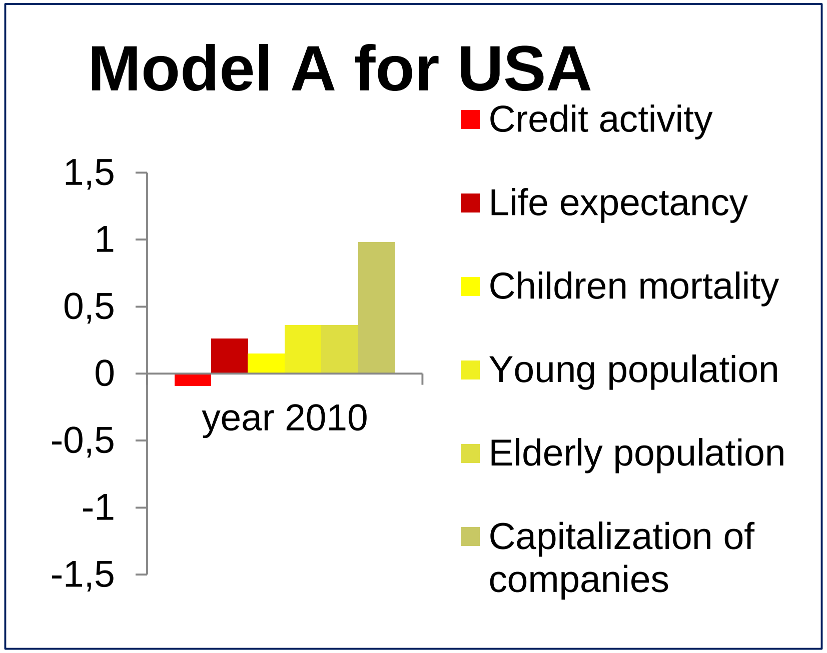
				features with δ=0		features with δ=.17	
	A1	A2	class	A1<1.95	A2<1.95	A1<1.95	A2<1.95
ex1	1.60	1.60	positive	*true*	*true*	*true*	*true*
ex2	1.70	1.65	positive	*true*	*true*	*true*	*true*
ex3	1.80	1.70	positive	*true*	*true*	***false***	*true*
ex4	1.90	1.80	positive	*true*	*true*	***false***	***false***
ex5	2.00	2.10	negative	*false*	*false*	***true***	***true***
ex6	2.10	2.20	negative	*false*	*false*	***true***	*false*
ex7	2.20	2.25	negative	*false*	*false*	*false*	*false*
ex8	2.30	2.30	negative	*false*	*false*	*false*	*false*

In the situation when δ=.17 is assumed the feature based on *A2* is more relevant than the feature based on *A1* and it will be used in the rule construction process

Subgroup -> risk model conversion

- Select a relevant subset of supporting conditions
- For each necessary and supporting condition construct one risk factor so that:

- *positive values always denote the existence of risk*
- *larger values always denote larger risk*
- *size = 0 if equal to the cut-off value*
- *size = 1 if equal to the mean value for the examples that are known to be members of the model.*

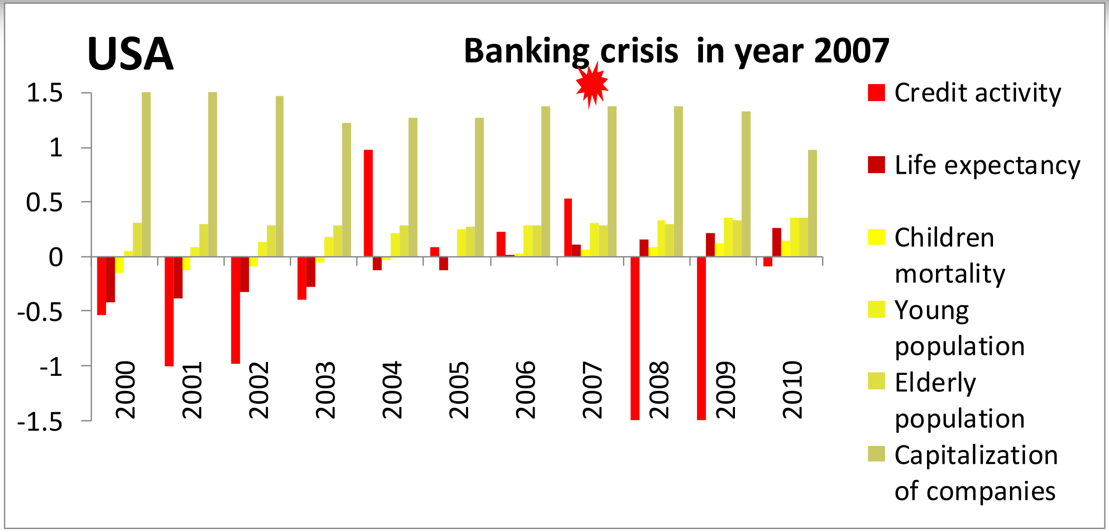


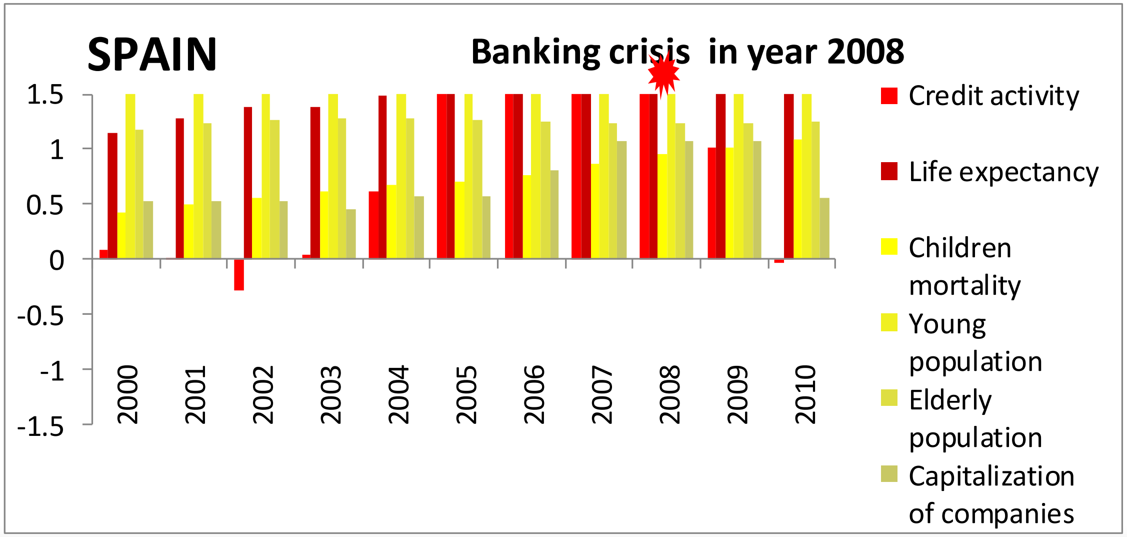
Subgroup -> risk model conversion

Risk factor name	World Bank indicator name	Function	Cut-off	Mean
Credit activity	Domestic credit to private sector (% of GDP)	Slope in three years period	5.8	12.0
Life expectancy	Life expectancy at birth, female (years)	Target year value	80.2	82.2
Children mortality	Mortality rate, under-5 (per 1,000 live births)	Target year value	8.0	4.8
Young population	Population ages 0-14 (% of total)	Target year value	21.6	17.4
Elderly population	Population ages 65 and above (% of total)	Value two years before the target	11.0	15.6
Capitalization of companies	Market capitalization of listed companies (% of GDP)	Maximal value in three years period	51.1	120.0

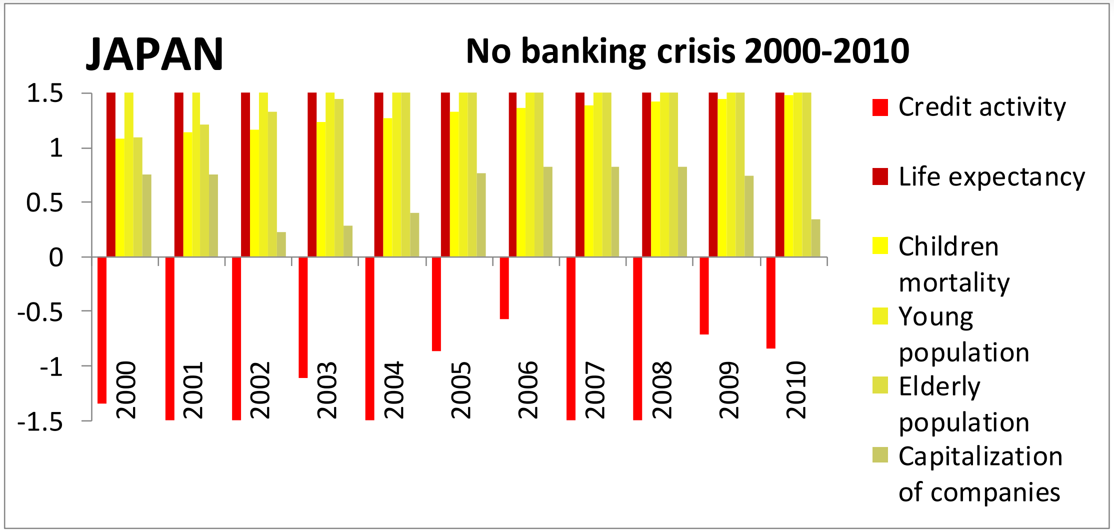
PreseValue = (FuncValue – CutOff) / (Mean – CutOff)

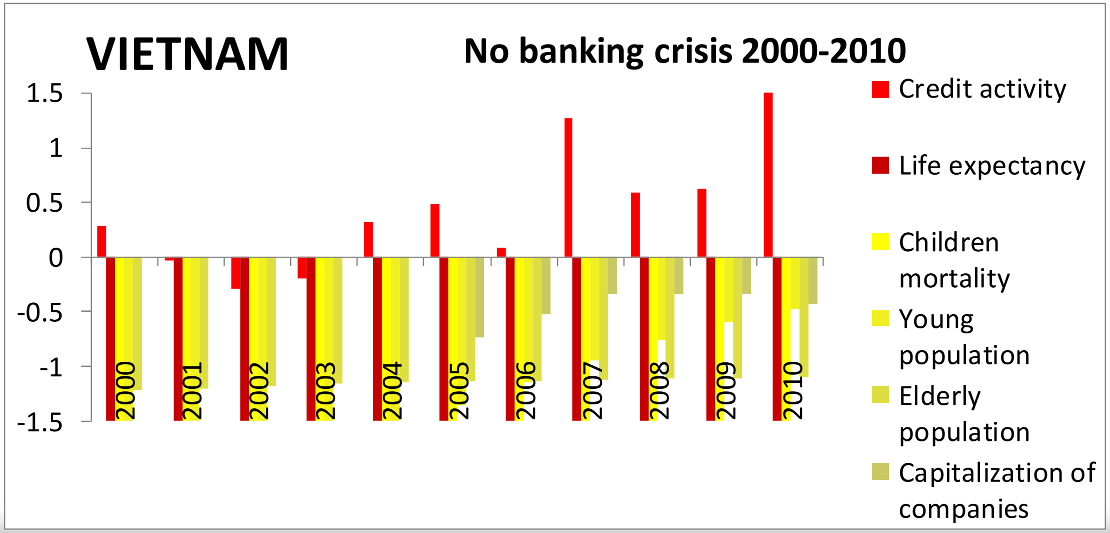
Model A – USA, Spain



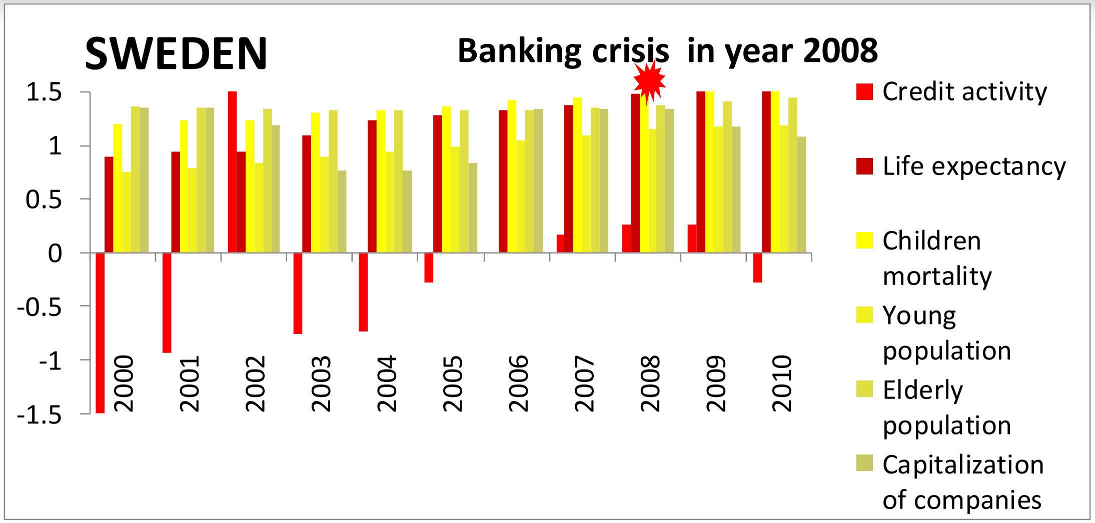


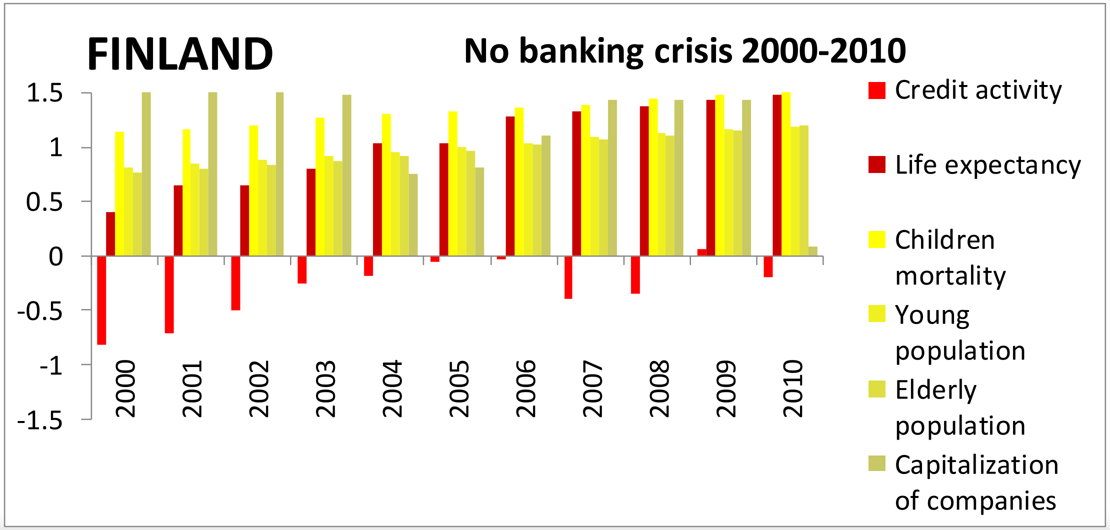
Model A – Japan, Vietnam



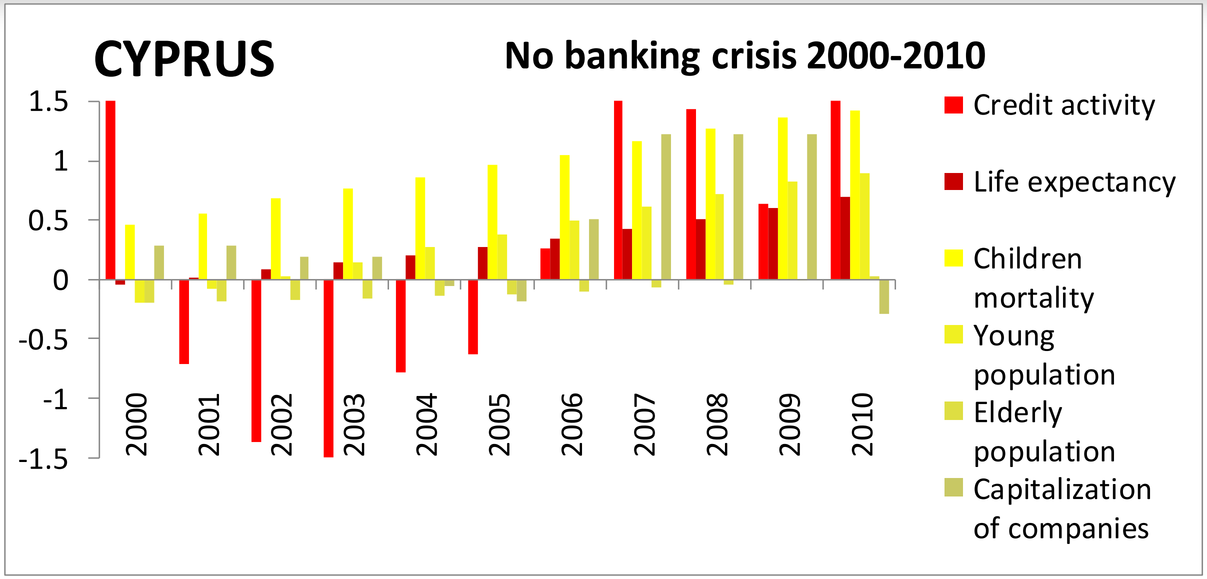


Model A – Sweden, Finland





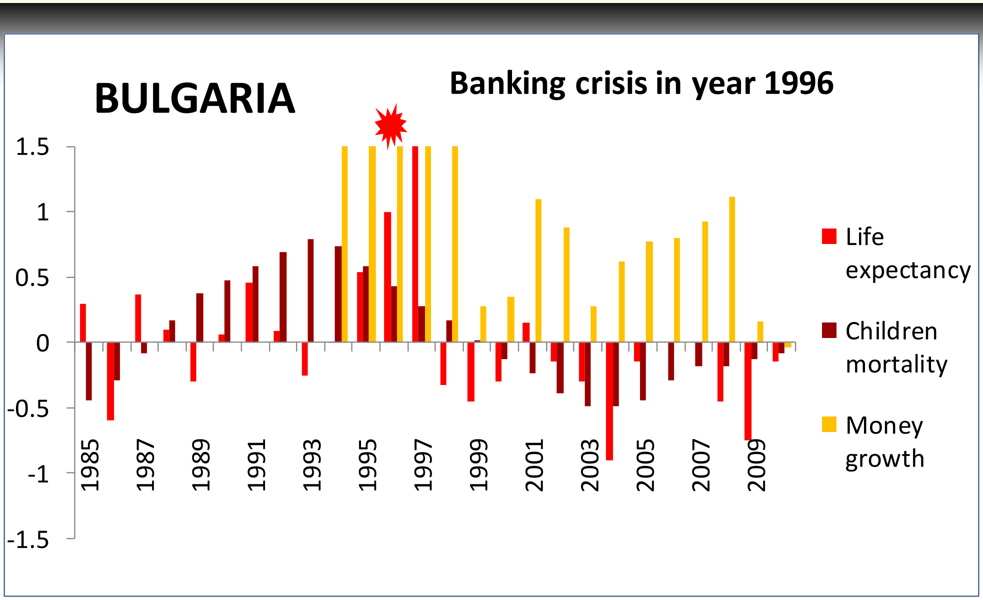
Model A - Cyprus



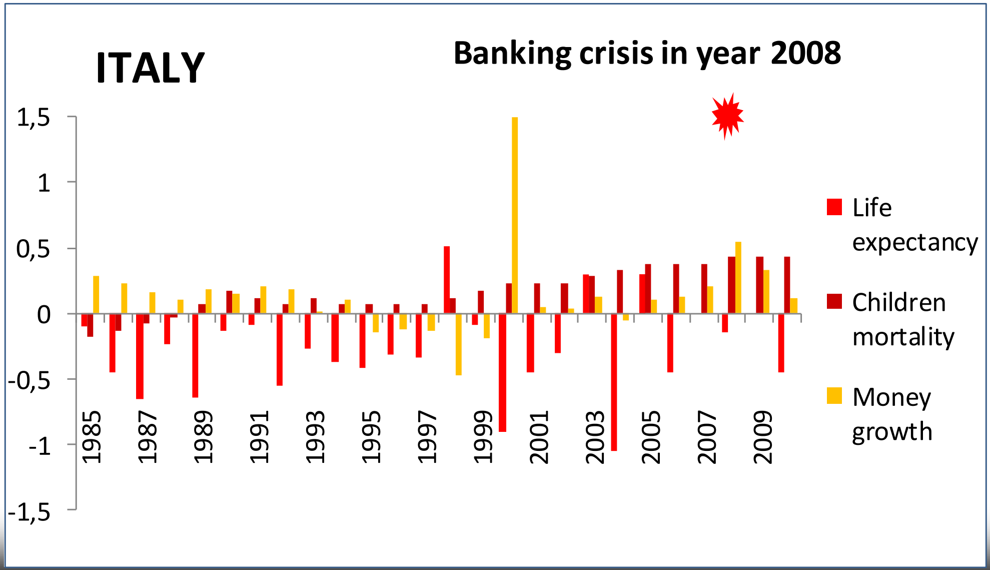
Model B

Risk factor name	World Bank indicator name	Function	Cut-off	Mean
Life expectancy	Life expectancy at birth, female (years)	Difference between maximal value one or two years before the target year and the target year value	0.0	0.7
Children mortality	Mortality rate, under-5 (per 1,000 live births)	Slope in three years period	-0.5	0.5
Money growth	Money and quasi money growth (annual %)	Value in the year before the target year	5.2	28.5

Model B – Bulgaria, Italy

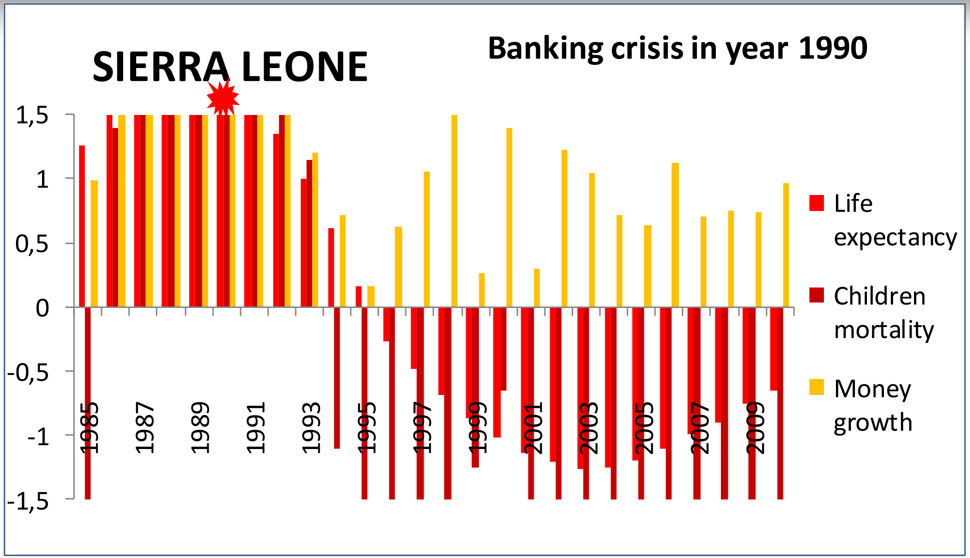

BULGARIA

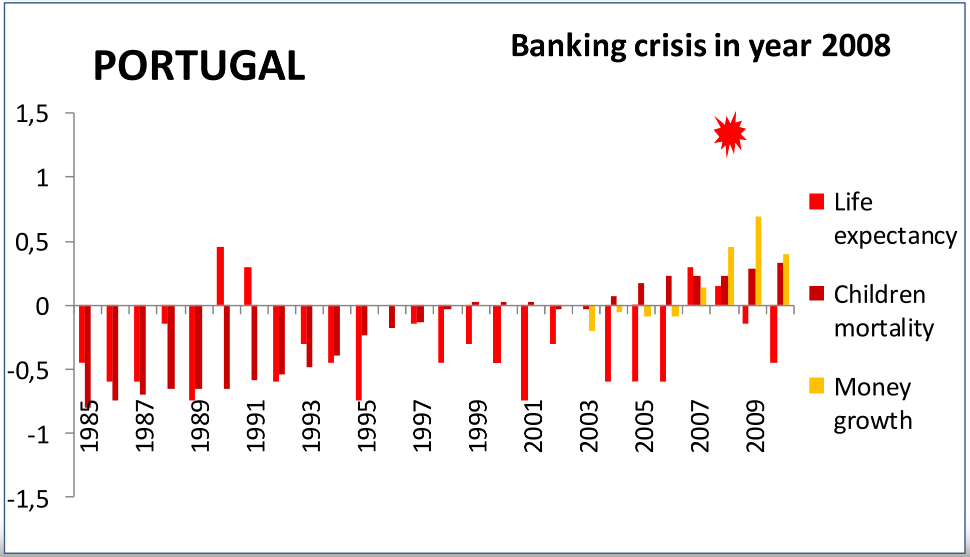
Banking crisis in year 1996


ITALY

Banking crisis in year 2008

Model B – Sierra Leone, Portugal





Comparative analysis
EU countries in Model B

1: *Fast growing credit activity in economies with aging population*

slope of credits in the period of three years before crisis > 5.8 % per year
life expectancy for females in the year before the crisis > 80.2 years.

Supporting conditions: low mortality of children, low percentage of young population, high percentage of elderly population, high capitalization of companies .

List of 16 included crises: Sweden in year 1991, USA and **UK in year 2007, Belgium, Denmark, France, Greece, Ireland,** Island**, Italy, Luxemburg, Netherlands, Portugal, Slovenia, Spain and Sweden in year 2008**.

5: *Socioeconomic problems recognized by non-increasing quality of public health*

non-increasing life expectancy for females in the period of three years before crisis
under-five mortality rate in the period of three years before crisis > -0.5 (per 1000)

Supporting conditions: high money and quasi money growth before the crisis.

List of 25 included crises: Sierra Leone in year 1990, Finland, Liberia, Nigeria, Norway, and Sweden in year 1991, Kenya and Poland in year 1992, Burundi in year 1994, Belarus, Central African Republic, Latvia, Lithuania, Swaziland, and Zimbabwe in year 1995, Bulgaria in year 1996, Ukraine in year 1998, Uruguay in year 2002, **Belgium, Greece, Hungary**, Island**, Italy, Portugal, and Spain in year 2008**.

World Bank governance indicators

Differences in p-ranks for years 2007 and year 2004 for six governance indicators for two groups of EU countries

	Control of corruption	Rule of law	Government effectiveness	Voice and accountability	Political stability	Regulatory quality	Total
Belgium	-4.32	0.00	-1.91	1.92	0.00	3.03	-1.28
Greece	-6.66	-5.74	-4.75	-8.65	0.48	0.22	-25.10
Hungary	-2.80	-1.44	-0.87	-4.33	-2.88	2.10	-10.21
Italy	-6.16	-6.22	-12.02	-1.44	6.25	-2.24	-21.82
Portugal	-4.78	-5.26	-6.23	-1.92	-5.77	-2.28	-26.25
Spain	-7.71	-0.96	-8.20	-4.81	-12.98	-0.85	-35.51
Austria	-0.96	3.35	3.43	2.88	11.06	3.47	23.23
Denmark	0.49	1.44	-0.49	-2.40	3.85	1.97	4.85
France	1.99	-1.91	-3.36	0.00	5.77	1.11	3.59
Germany	-0.94	0.00	3.45	0.00	14.90	3.00	20.41
Netherlands	1.48	0.00	-3.39	1.92	-7.69	-0.47	-8.14
Level of statistical significance	**99.9%**	**97%**	**96%**	Non-sig.	Non-sig.	Non-sig.	**99%**

Results

M. Francis. Governance and financial fragility. Financial System Review pp.73-76, 2003.

> *“Good governance plays a significant role in determining the extent to which a country is likely to have a crisis.”*

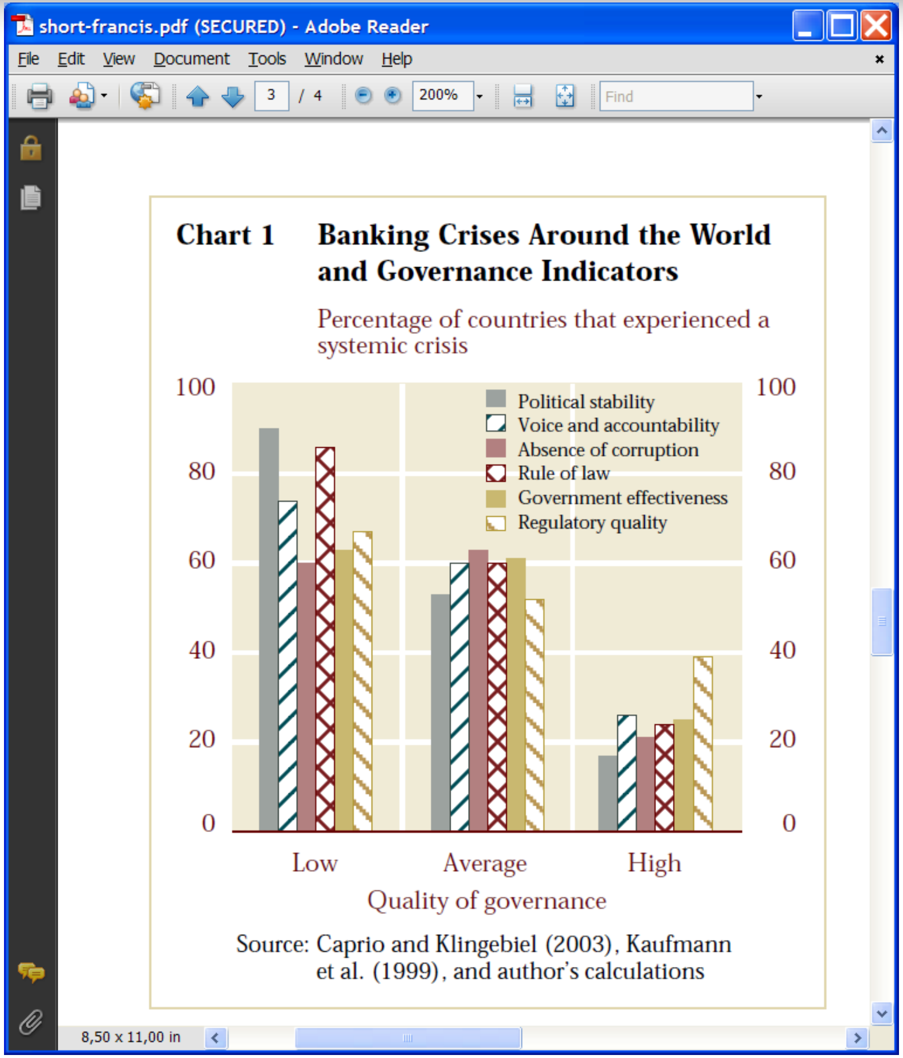


Chart 1 Banking Crises Around the World and Governance Indicators

Percentage of countries that experienced a systemic crisis

Source: Caprio and Klingebiel (2003), Kaufmann et al. (1999), and author's calculations

The result demonstrates that Model B that has been the basis for selecting a subset of 6 countries is reasonable !!

Model B is based on socioeconomic problems recognized ba non-increasing quality of public health.

Now we have:
Good governance problems
Socioeconomic problems
Banking crises

Results

Difference in p-ranks for governance indicators in year 2011 and year 2008

Total for 6 indicators	Total for 3 most relevant indicators	Control of corruption indicator
Greece -39.49	Greece -19.57	Italy -5.76
Malta -29.58	Malta -11.35	Cyprus -5.33
Slovenia -26.84	Austria -8.24	Greece -5.24
Portugal -25.13	Hungary -7.59	Austria -5.09
Ireland -19.08	Cyprus -6.62	Malta -3.84

Conclusions

- Data preparation is important
- Subgroup discovery is useful for different tasks
- Subgroups may be transformed into risks models
- Comparative analysis of examples included into different subgroups may result by interesting novel knowledge

Thank you for your attention!

Questions?