Analysis of world development indicators as predictors of countries' economic status

Blandine Clément
Salmane Kechkar
Antonio Jimenez Gonzalez

A Network Tour of Data Science (NTDS)
EE-558

Introduction

- Data-driven approach to study global development
- Time series data from the World Bank
- Aim : Analyze which indicators or themes are better at predicting economic status of countries

Roadmap

1. Data acquisition
2. Exploration
3. Exploitation
4. Discussion

Data acquisition

- Kaggle dataset → Health, nutrition & population statistics
  - 258 countries, 345 indicators, 56 years (1960-2015)
- World Bank API → queried all the indicators
  - 221 countries, 8266 indicators, 56 years (1960-2015)

Exemples of indicators

- '5.51.01.09.water': 'Access to water'
- 'EN.ATM.CO2E.KT': 'CO2 emissions (kt)'
- 'SE.SCH.LIFE.FE': 'Expected years of schooling, female'
- 'EG.USE.ELEC.KH': 'Electric power consumption (kWh)'
- 'SH.MMR.LEVE': 'Number of weeks of maternity leave'
- 5.51.01.02.malnut': 'Child malnutrition'
- 'AG.PRD.GLVSK.XD': 'Livestock production index
- 'BM.GSR.MRCH.ZS': 'Merchandise imports (BOP): percentage of GDP (%)'
- 'DT.DOD.LTST.CD': 'External Debt, total'
- 'CC.EST': 'Control of Corruption: Estimate'

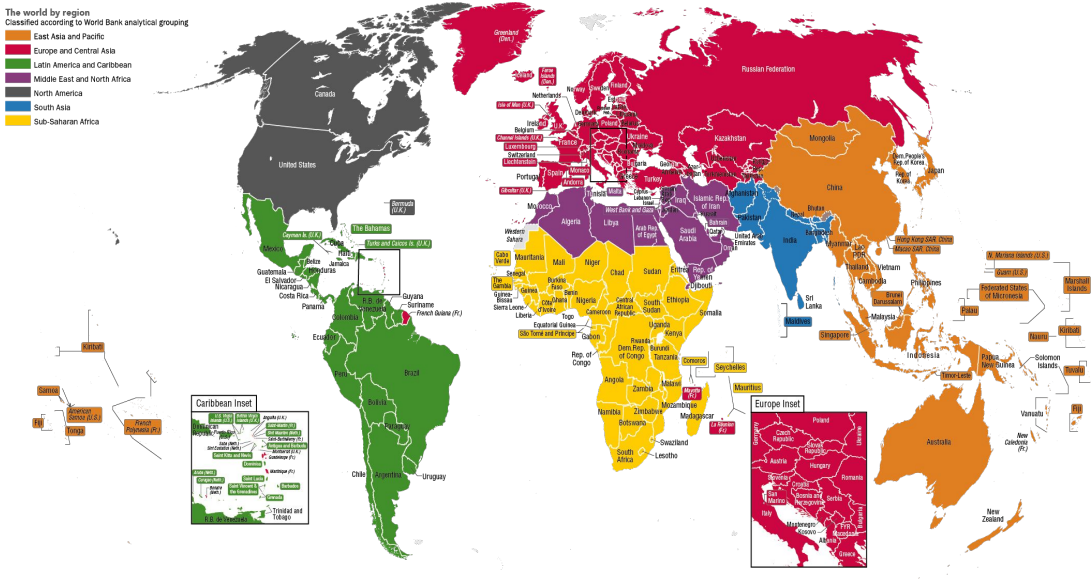
Data acquisition

Labels of the countries from the World Bank:

1) Income level (GNI/capita): **4 classes**
2) Region : **7 classes**

Custom label:

3) Migration rate : **2 classes**



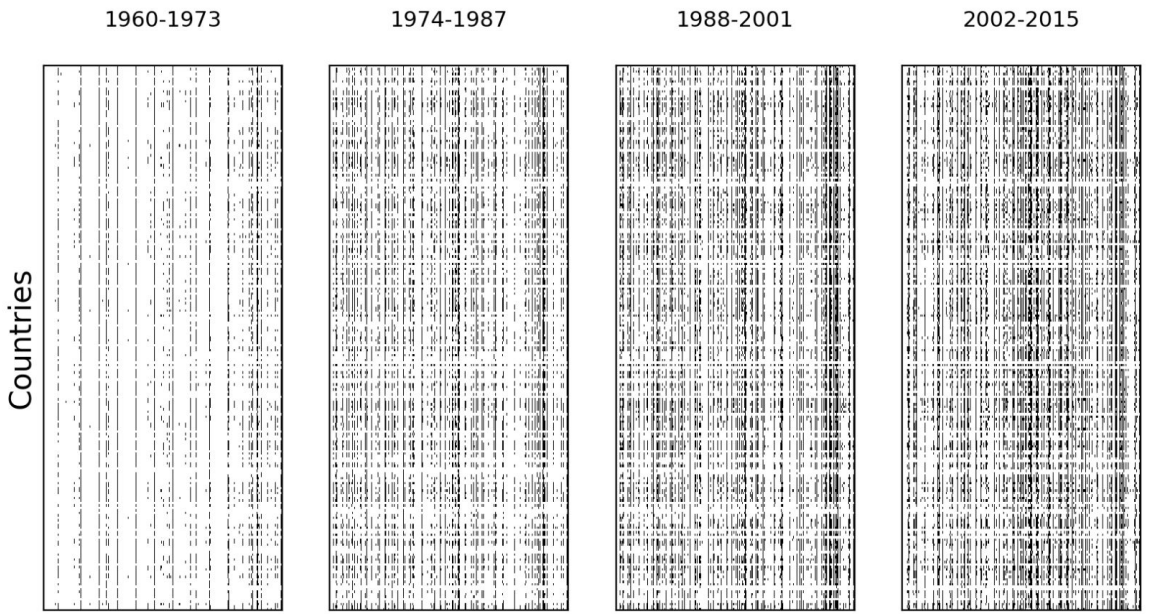
Income level :
- HIC: 70
- UMC: 53
- LMC: 56
- LIC: 35
- NA: 4

Region :
- EAS: 37
- ECS: 58
- LCN: 42
- MEA: 21
- NA: 4
- SAS: 8
- SSF: 48
- NAC: 3

Migration:
- >emigration: 120
- >immigration: 78
- Other: 23

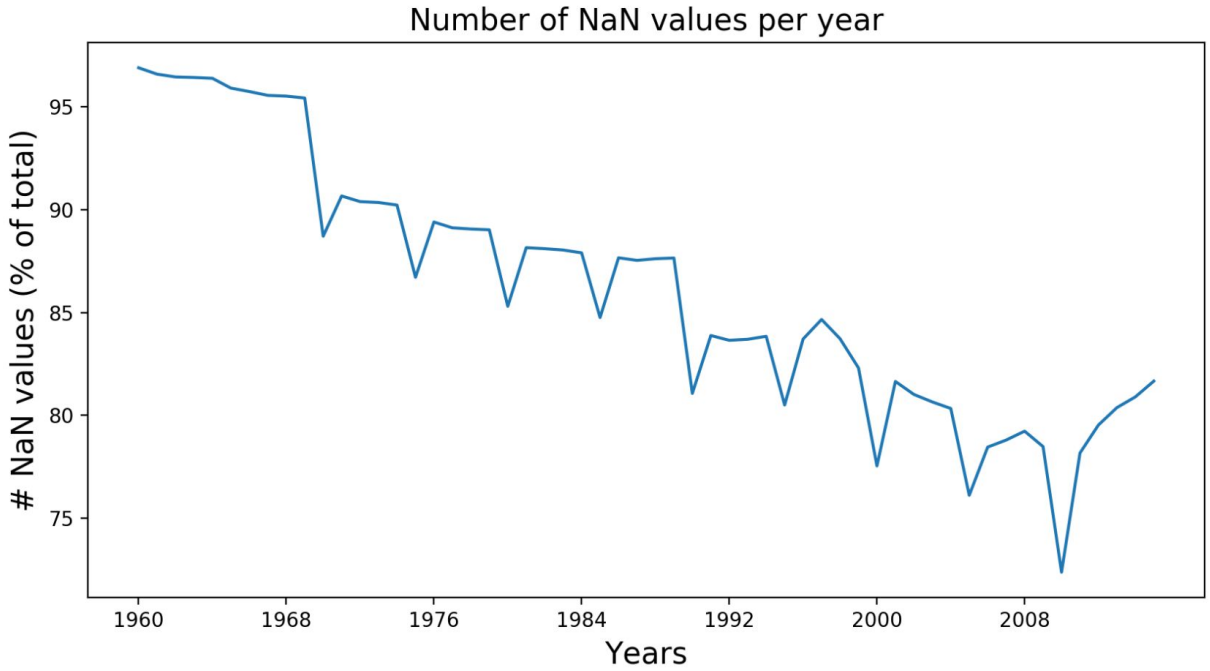
Data visualization

- Raw dataset: **86.13% are NaN**
- Very sparse matrix
- Increasing data by year
- Demographic data always available



Data visualization

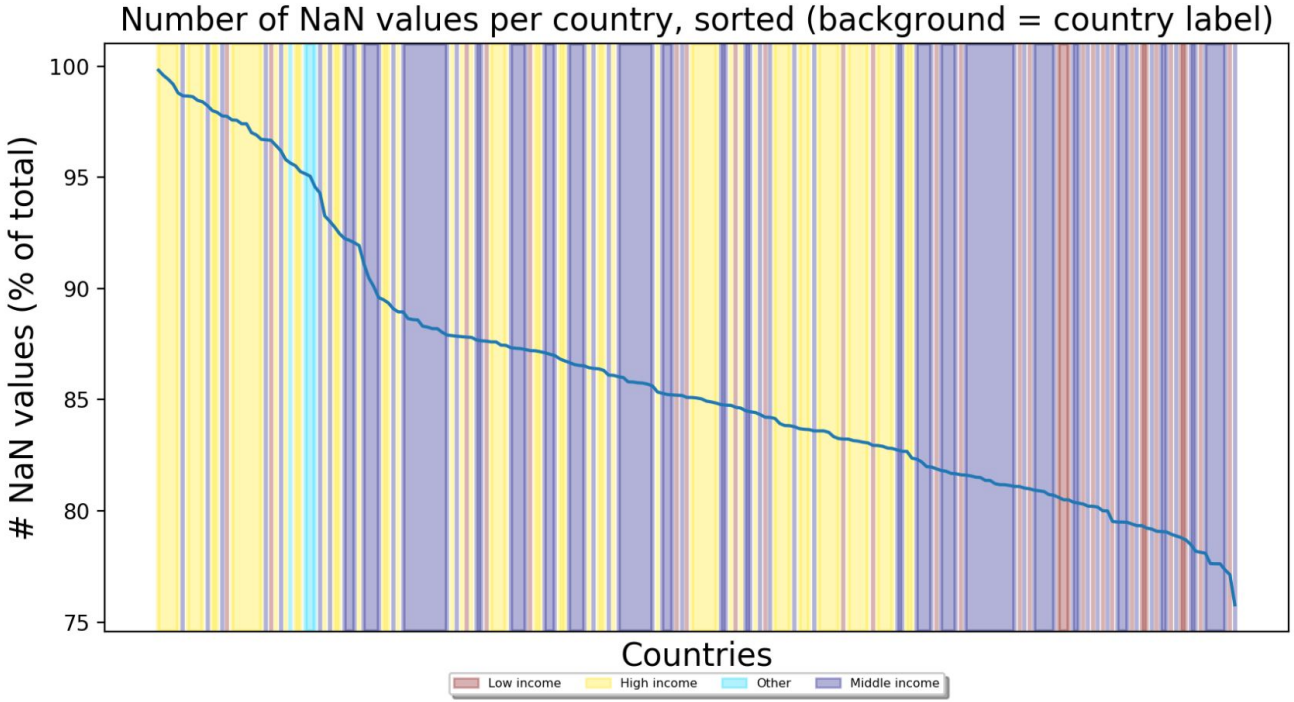
- More data in 5 years interval



Acquisition | Exploration | Exploitation | Discussion

Data visualization

- More data for low income countries



Acquisition | Exploration | Exploitation | Discussion

Data Cleaning – get rid of the NaN!

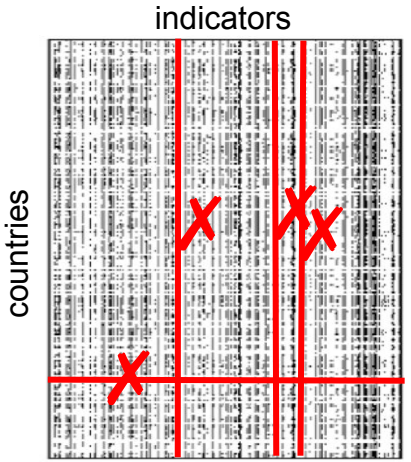
Main issue: too many not existing values (86%)

Reasons?

- some indicators were sampled every x- years
- indicators very specific to a region of the world and/or not accessible

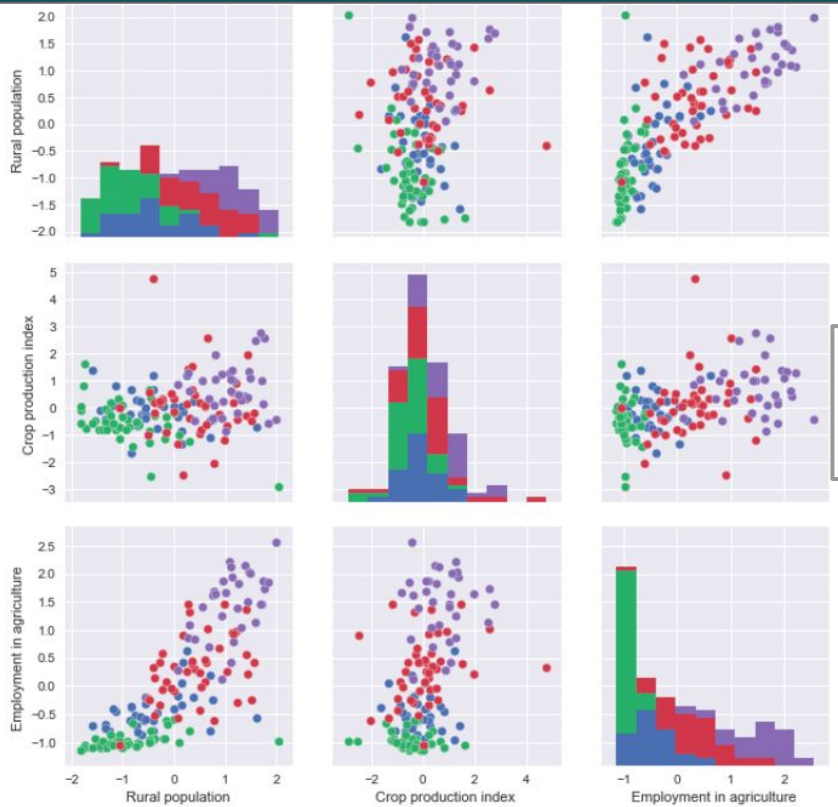
Two strategies:

- Make a three-years average of the data → compensate absent values by existing values.
  - why 3? keep the dynamic of changes over time
- Drop the countries and indicators containing the max number of NaN values
  - keeping the maximum # countries



Resulting in 18 functional datasets:
[1960-1962]: **145** countries and **183** indicators.
.
.
.
[2011-2013]: **157** countries and **645** indicators

Checking some correlations



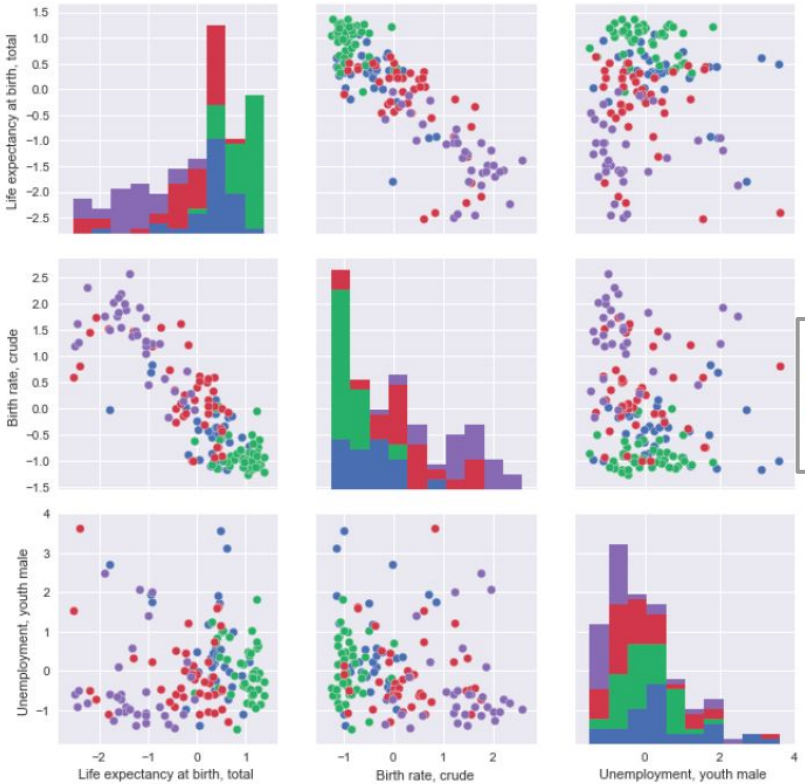
Variables for the period 2008-2010:

1. **Rural population**
2. **Crop production index**
3. **Employment in agriculture**

- High income
- Upper middle income
- Lower middle income
- Low income

Acquisition | Exploration | Exploitation | Discussion

Checking some correlation



Variables for the period 2008-2010:

1. **Life expectancy at birth**
2. **Birth rate, crude**
3. **Unemployment, youth males**

- High income
- Upper middle income
- Lower middle income
- Low income

Acquisition | Exploration | Exploitation | Discussion

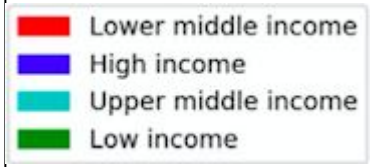
Dealing with the dataset

~500 features per 3y-period

~150 countries left per period
(not many)

3 possible ‘groupings’:

by income level



by regions

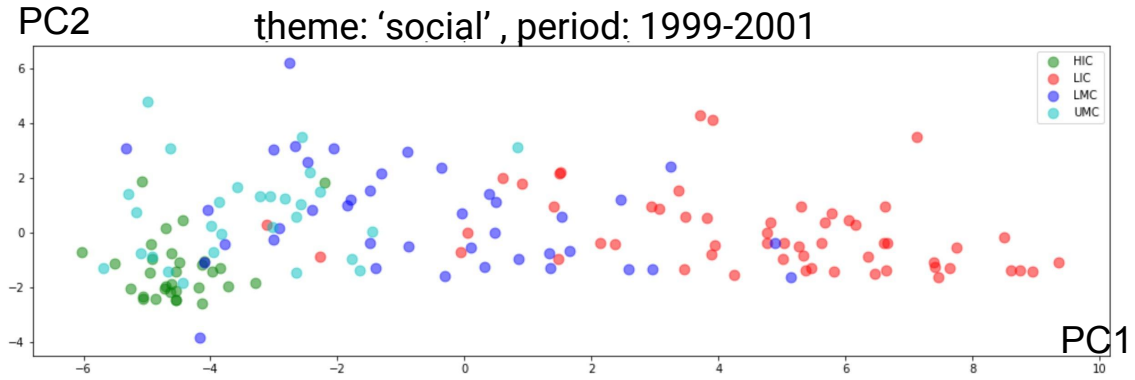
by net migration

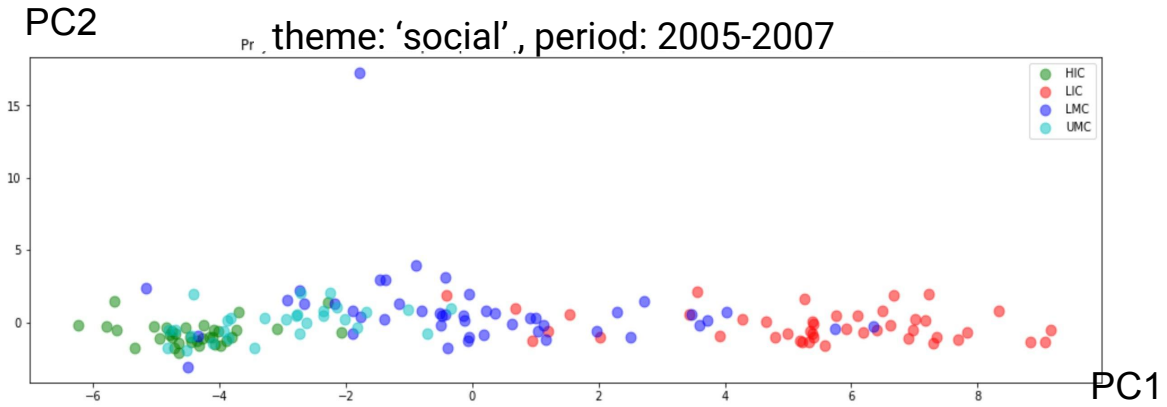
Decomposition into a thematic approach with custom-made features:

- demographic (‘birth’, ’expectancy’,...)
- education (‘school’, ’literacy’,...)
- health
- social
- economic

- political instability
- technology (‘telephone, ’cellular’,’internet’...)
- gas_production

How separable is our data





- data nicely separable
- 32 features related to the theme ‘social’
- In the period 2005-2007, the PC1 seem to be enough to explain most of the variance in the data

Data to Network

Next step: make a similarity graph out of the data.

DATA → WEIGHT MATRIX → SIMILARITY GRAPH

Construction of a weight matrix, weighting the similarity between the countries, based on the different themes-related features.

- Distance metric between features: cosine was chosen
- Kernel: cosine
- Sparsification method: nearest neighbors
- Number of neighbors: ~25

result: Multiple similarity networks: one for each period, each theme

→ Goal: to pre-visualize some clusterization patterns

Network visualization

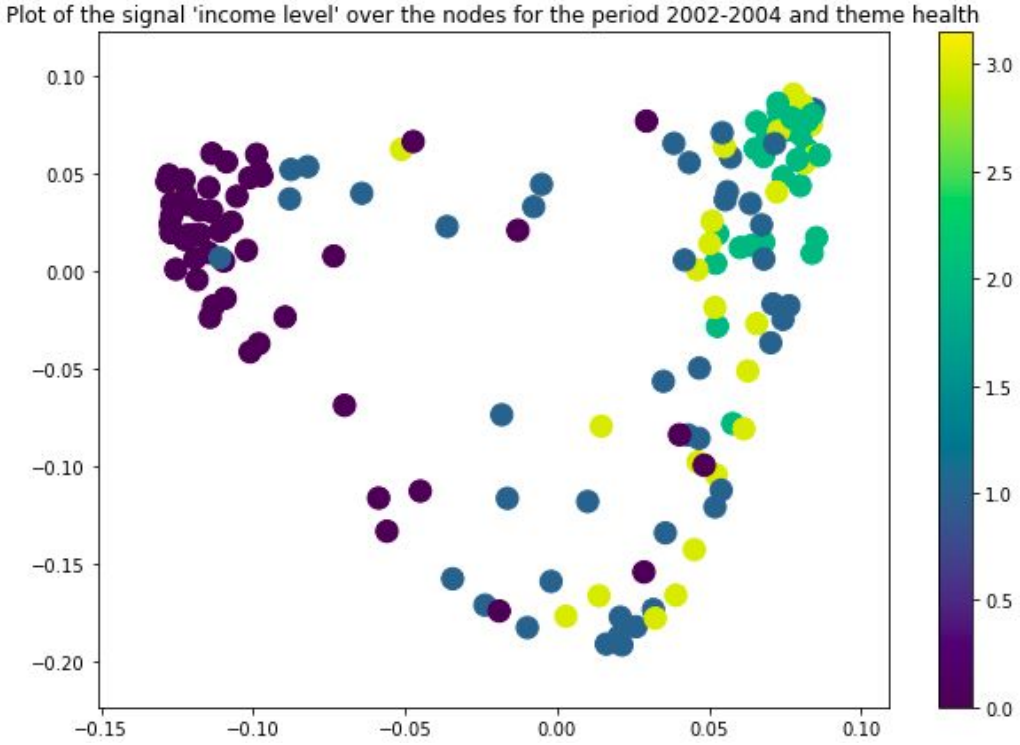
World Data Visualization (Online)

Learning on graph – 'health' , period=2002-2004

- Laplacian was computed
- projection of data on the 2nd and 3rd eigenvectors
- plot of the signal: 'income level' labels

- only used for the 'income level' grouping
  - by regions: too many, for not enough nodes
  - by net migration: the visualization of network did not shown any convincing results

- Masking 75% of the nodes' labels

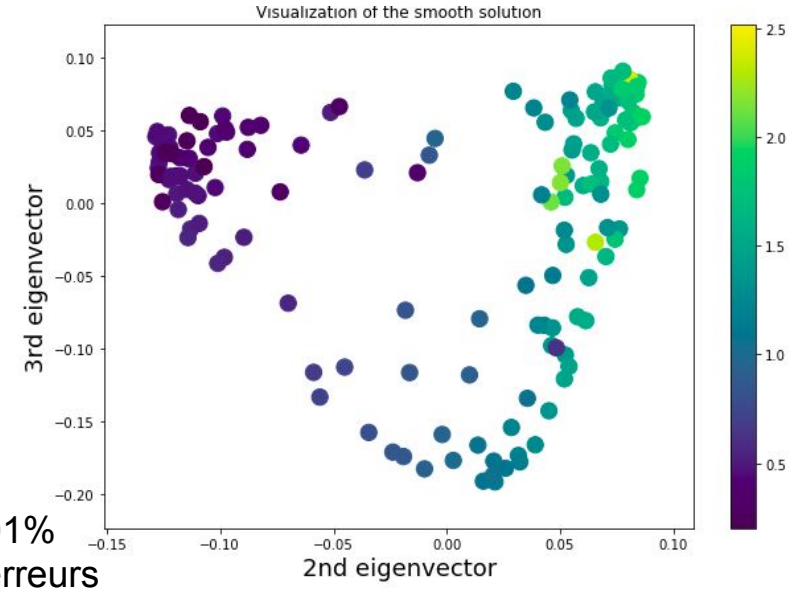


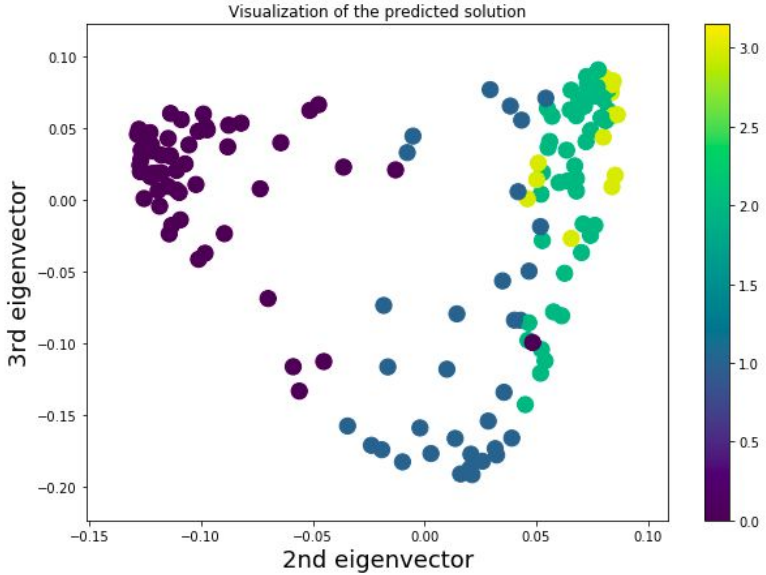
Acquisition | Exploration | Exploitation | Discussion

Learning on graph – 'health' , period=2002-2004

- Transductive learning: by solving the following: $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \|\mathbf{y} - \mathbf{M}\mathbf{x}\|_2^2 + \alpha \mathbf{x}^\intercal \mathbf{L}\mathbf{x},$

  → solution: $\mathbf{x}^* = \mathbf{y}(M + \alpha L)^{-1}$

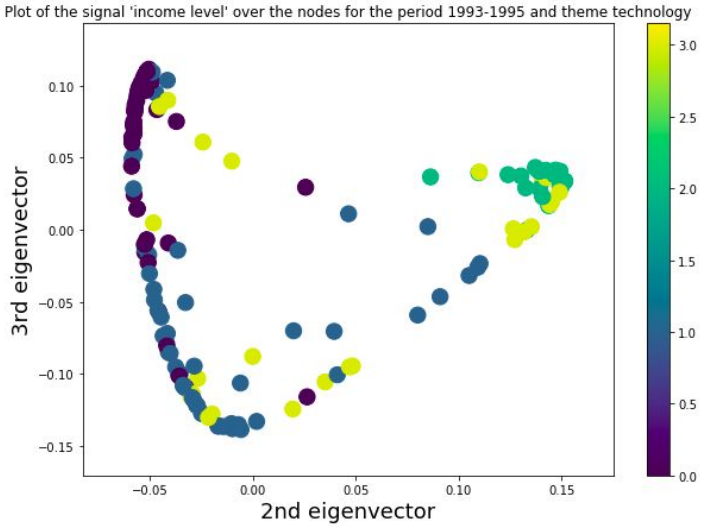




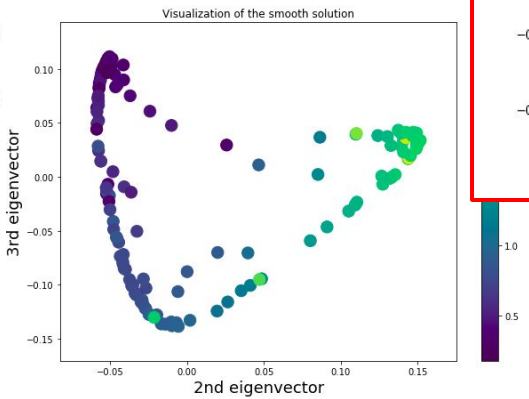
31.01%
49 erreurs

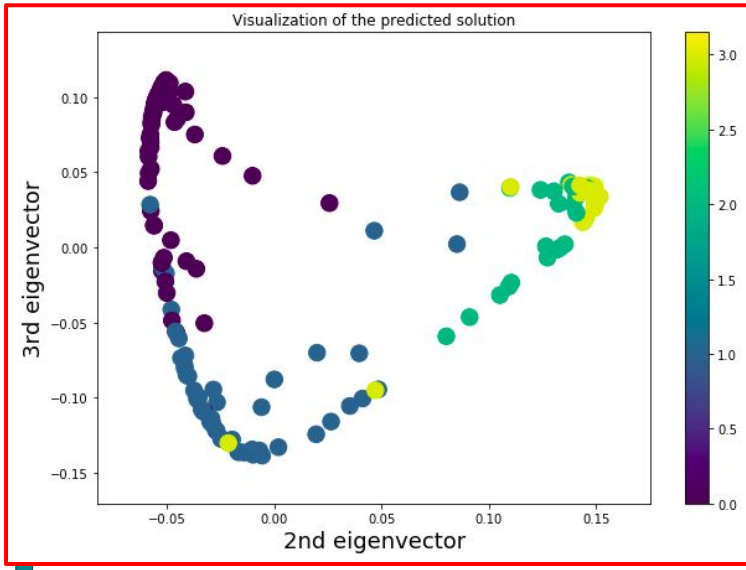
Acquisition | Exploration | Exploitation | Discussion

Learning on graph- ‘technology’ , period=1993-1995



smooth solution



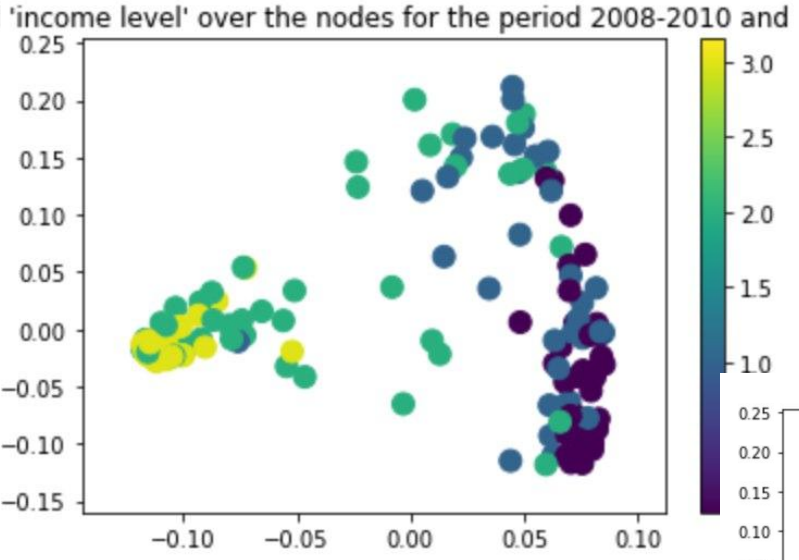


predicted solution

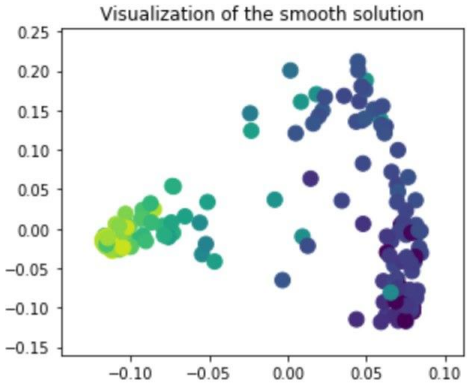
error=34%

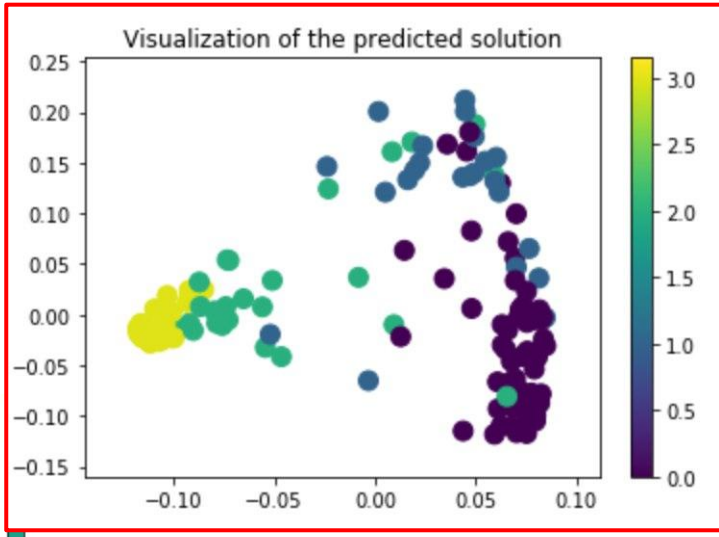
Acquisition | Exploration | Exploitation | Discussion

Learning on graph – 'social' , period:2008-20010



smooth solution





- error=29%

Acquisition | Exploration | Exploitation | Discussion

Discussion

- Different features inside a given theme across the years.
  → lack of comparison throughout the years

- Features were only considered as thematics and were never looked at independently
- Thematics were arbitrary (lack of relevance of some indicators?)

- Initially, very low number of nodes (195 countries in the world) and drop to ~150 countries after cleaning

- Method for cleaning NaN not optimal

- Global consideration of countries without taking into account regional factors

End

Thank you for your attention !