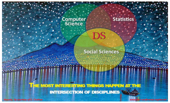
Extracting Social Indicators from Big Data: an Experience in Measuring Wellbeing

R.Campagni[1], L. Gabrielli[2,3], F. Giannotti[2], R. Guidotti[2,3], F. Maggino[1], D. Pedreschi[3]

[1]University of Firenze, [2]ISTI-CNR Pisa, [3]University of Pisa

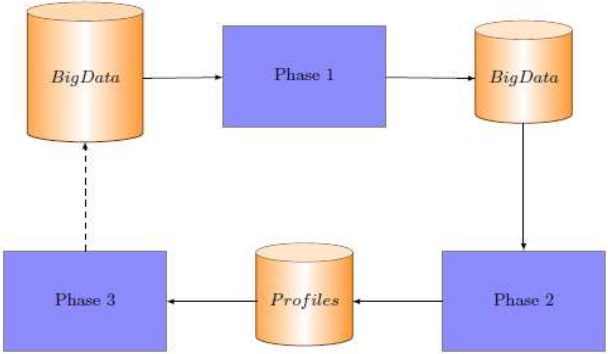
Febrary, 17-19 2016 Naples (Italy)



Aims and phases of project

- *Extracting* from big data concerning the purchases useful information to construct indicators describing social phenomena.
- *Analyzing* the behavior of different families in a crucial period, by paying attention to possible changes in the lifestyle of the people and the role of crisis of the last years.
- *Defining* new social indicators to describe customer purchase behaviors, by changing the classical methodological approach by considering data collected for other purposes.

The analytical process



- Phase 1: exstract from data useful information for analysis
- **Phase 2: perform analysis**
- Phase 3: define new indicators

Period and analysis goals

We observe customer's purchases during 2007-2013 to detect crisis signals from data.

- 2007-2008 period before crisis
- 2009-2012 crisis period
- 2013 eventual economic recovering

We are looking for important factors helping us defining new social indicators related to welfare.

New indicators - the aim

Possiblility to:

- obtain timely information;
- discover important signals related to particular behavior;
- predict changes in the macroeconomic context.

New indicators - How to do it

- Grouping customers by clustering techniques;
- discovering particular characteristics about each cluster, such as *how much*, *what* and *when* people buy.

$$\Downarrow$$

Total amount, total quantity and total number of times in which shopping is made (reference period the year and/or month).

Exploring the changes

Observing changes in shopping cart to understand if:

- *amounts, quantities and number of expenses significantly change* $\Rightarrow$ *typologies of products purchased also change.*

For example, during crisis, a group of customers has reduced purchase of niche products, to the benefit of lower-end products.

Matrix construction for year analysis

customer_id	year0	year1	...	...	year(n-1)	type
10	5	4	...	...	8	time
10	100	120	...	...	250	quantity
10	300	600	...	...	1050	amount
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Year Matrix

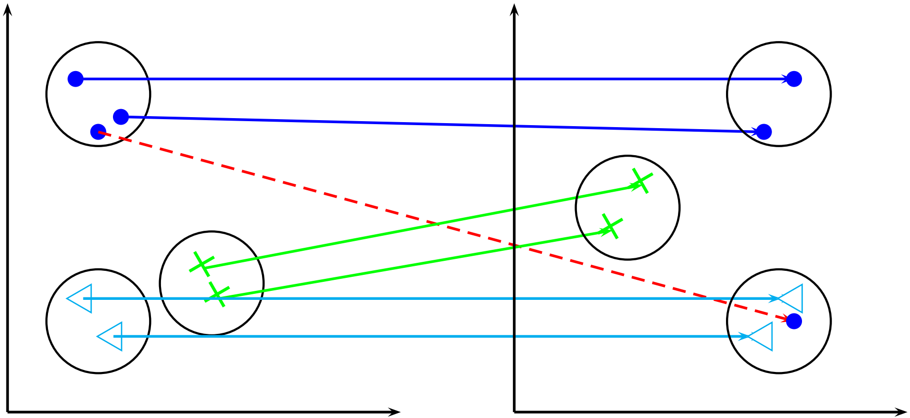
Applying the *K-means* algorithm for each attribute, each customer is assigned to one of the K clusters; each cluster contains customers with similar purchase behavior.

The results

- Checking the quality of the clustering procedure on each attribute (internal consistence, well separated).
- Understanding relationships among results obtained studying different attributes.

Deepening and comparing results

Clustering on attribute **amount** Clustering on attribute **quantity**



Matrix construction for month analysis

customer_id	year	january	febrary	...	...	december	type
10	year1	0	0	...	...	1	time
10	year2	0	0	...	...	0	time
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	year1	0	0	...	...	30	quantity
10	year2	0	0	...	...	0	quantity
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	year1	0	0	...	...	96	amount
10	year2	0	0	...	...	0	amount
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Month Matrix

For each customer and for each year we have a 12-sequence of monthly-values regarding the different attributes that we analyze separately.

Monthly analysis

- For each attribute, we apply the clustering algorithm *K-means* on data regarding the first year.
- The algorithm assigns each customer to one of the K clusters; each cluster contains customers with similar characteristics.
- The obtained output model is used to classify the customer behavior during the next *n-1* years.

At the end a sequence of n clusters is assigned to each customer; sequences describe the different purchase behaviors related to the considered attribute.

The new data matrix

Clusters *passed through* by customers 10 and 20 in the *n* years.

customer_id	**year0**	**year1**	**...**	**...**	**year(n-1)**
10	cluster_0	cluster_0	...	...	cluster_1
20	cluster_1	cluster_0	...	...	cluster_0

- A new clustering step to find groups of customers who, during the observed period, have similar patterns. Customers in the same clusters have similar purchase behaviors by *passing through* similar *clusters paths*.

Interpretating clusters

Labels to describe the clustering results:

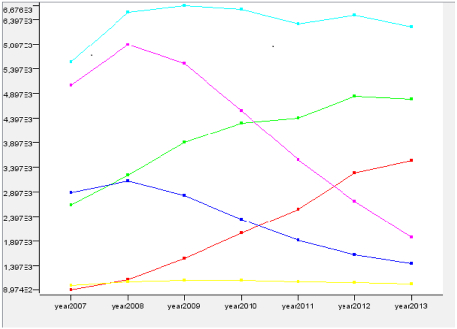
- *L1* indicates the behavior in the lower range for the amount (or for another attribute);
- *LK* indicates the behavior in the higher range for the amount.

$$customer\_id \Leftarrow \{Li_0, ..., Li_{n-1}\},$$

with $i = 1...K$.

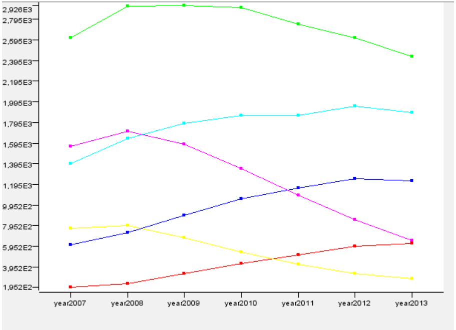
Year analysis on amounts

- Data: 39192 rows for 13064 customers from 2007 to 2013 in a store.
- For each customer information about times, quantities and amounts.



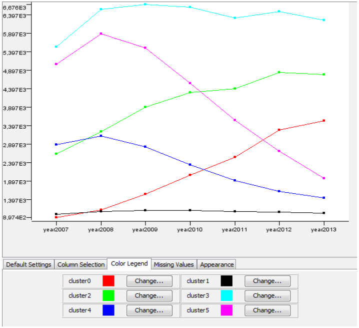
Centroids of clustering on **amounts**, $K = 6$.

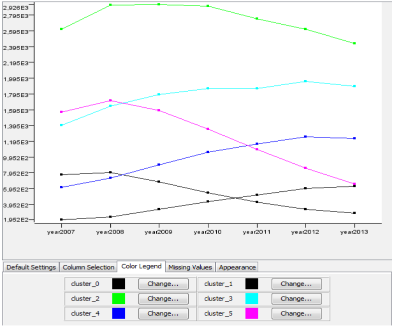
Year analysis on quantities



Centroids of clustering on **quantities**, $K = 6$.

Deepening the results





Centroids of clustering on **amounts** and **quantities**, $K = 6$; in evidence the black lines indicating the cluster with the same customers.

Data for month analysis

Remember Month Matrix:

customer_id	year	january	febrary	...	...	december	type
10	year1	0	0	...	...	1	time
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	year1	0	0	...	...	30	quantity
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	year1	0	0	...	...	96	amount
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Clustering Analysis starts on data (amounts) regarding the first year, 2007.

Clustering about the first year

	cluster_0 (2746) 30%	cluster_1 (2392) 25%	cluster_2 (969) 11%	cluster_3 (1194) 13%	cluster_4 (1408) 15%	cluster_5 (519) 6%
1	31.573	124.728	391.629	343.157	140.522	551.620
2	31.573	124.728	391.629	343.157	140.522	551.620
3	37.737	151.831	339.329	413.850	195.6836	593.312
4	39.446	150.257	300.673	415.407	223.354	601.450
5	40.202	150.161	284.458	424.067	222.698	583.583
6	45.119	152.636	265.659	424.890	252.411	593.513
7	43.453	151.959	217.341	404.070	252.825	565.820
8	45.687	153.132	205.695	405.583	278.411	576.477
9	49.377	160.690	226.807	407.059	290.825	582.036
10	53.454	148.885	234.879	419.744	307.081	590.105
11	53.534	137.962	220.120	398.948	289.469	568.281
12	68.762	168.097	230.822	467.640	375.366	630.116

Coordinates of the centroids clustering on the dataset of the **amounts** for the year 2007.

Final clustering: the result

```
Number of iterations: 4
Within cluster sum of squared errors: 16426.0
Missing values globally replaced with mean/mode

Cluster centroids:
```

Attribute	Full Data (5792)	Cluster# 0 (2115)	1 (656)	2 (873)	3 (849)	4 (634)	5 (665)
Cluster2007	L1	L2	L5	L1	L2	L1	L5
Cluster2008	L2	L2	L5	L1	L2	L1	L5
Cluster2009	L2	L2	L5	L1	L1	L1	L5
Cluster2010	L2	L2	L3	L2	L1	L1	L5
Cluster2011	L2	L2	L4	L2	L1	L1	L5
Cluster2012	L2	L2	L4	L2	L1	L1	L5
Cluster2013	L1	L1	L2	L2	L1	L1	L5

The clustering result, for amounts, performed on the sequences describing, for each customer, the clusters crossed over years.

Conclusions and future works

Obtained profiles suggest some insights about product categories bought by customers:

- some customers bought more products to the quality downside;
- other customers bought fewer products privileging the quality or prices increased.

Defining new indicators to describe changes that impact on people's customs.

This work is supported by the European Community's H2020 Program under the scheme $'INFRAIA - 1 - 2014 - 2015$: Research Infrastructures', grant agreement #654024 'SoBigData: Social Mining & Big Data Ecosystem'. ($http://www.sobigdata.eu$).

Thanks for attention!