POLITECNICO DI MILANO

Summer School: la matematica incontra il mondo

San Pellegrino Terme, 5 - 6 - 7 Settembre 2016

Data Science, il futuro della statistica

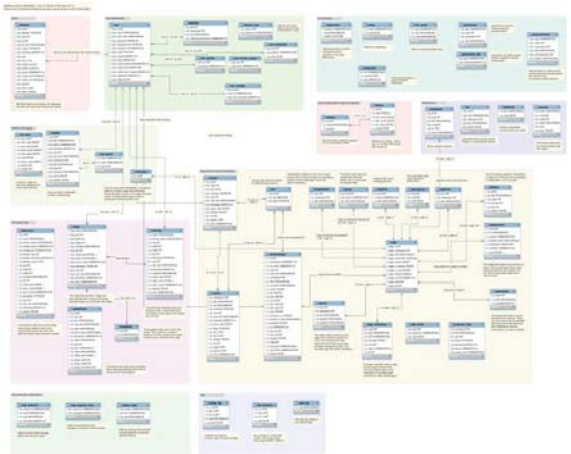
Piercesare Secchi

Dipartimento di Matematica – Politecnico di Milano

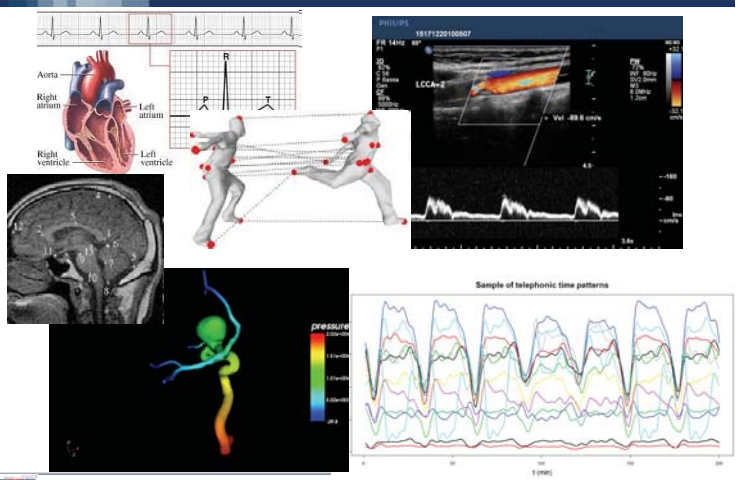
piercesare.secchi@polimi.it

The data deluge era

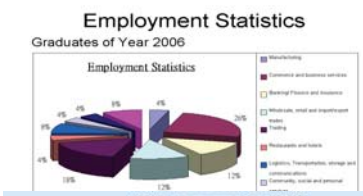
Big data: large datasets

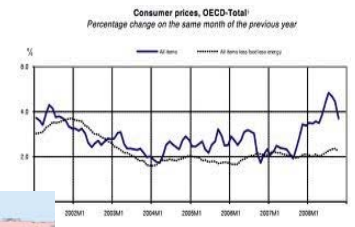


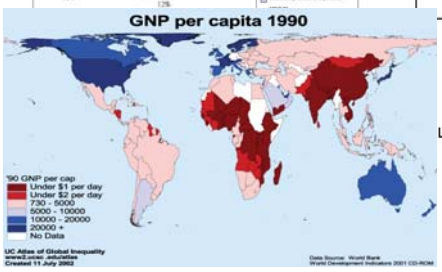
Big data: complex and high dimensional data

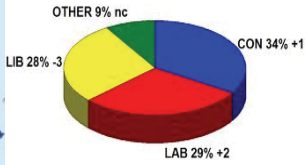


Statistics?









2010 UK General Elections

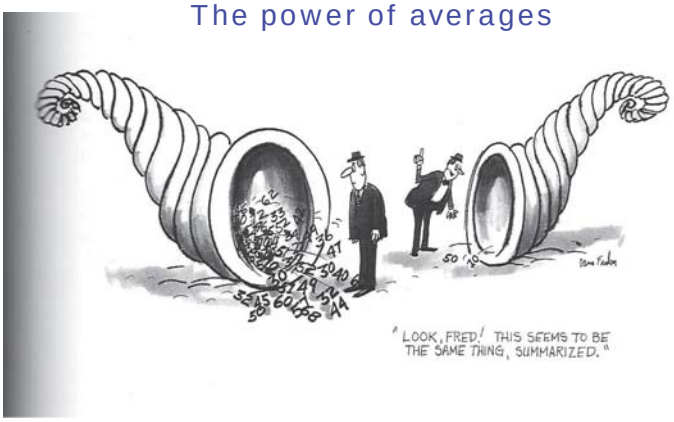
Statistics: science of averages?

Statistics: science of averages?

« *Sai ched'è la statistica? È na' cosa*
che serve pe fà un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che spósa.
Ma pè me la statistica curiosa
è dove c'entra la percentuale,
pè via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perch'è c'è un antro che ne magna due. **»**

(Trilussa, La Statistica)

The power of averages



Energy Statistics (source: United Nations)

Italy

Economy		
GDP growth rate from previous year (%)	2	2006
GDP per capita ($US)	35,585	2007
% Value added agriculture, hunting, forestry, fishing	2	2007
% Value added mining, manufacturing, utilities	21	2007
% Value added other	77	2007
Energy		
Energy consumption (1000t oil eq.)	181,278 [1]	2006
Energy consumption per capita (kg oil eq.)	3,076 [1]	2006
Energy intensity (kg oil eq.) per $1,000 (PPP) GDP	110	2006
Renewable electricity production (%)	17.0 [1]	2006

USA

Economy		
GDP growth rate from previous year (%)	3	2006
GDP per capita ($US)	45,047	2007
% Value added agriculture, hunting, forestry, fishing	1	2007
% Value added mining, manufacturing, utilities	17	2007
% Value added other	82	2007
Energy		
Energy consumption (1000t oil eq.)	1,994,876	2006
Energy consumption per capita (kg oil eq.)	6,684	2006
Energy intensity (kg oil eq.) per $1,000 (PPP) GDP	182	2006
Renewable electricity production (%)	8.0	2006

Energy Statistics (source: United Nations)

Switzerland

Economy		
GDP growth rate from previous year (%)	3	2006
GDP per capita ($US)	56,579	2007
% Value added agriculture, hunting, forestry, fishing	1	2007
% Value added mining, manufacturing, utilities	22	2007
% Value added other	77	2007
Energy		
Energy consumption (1000t oil eq.)	19,785 [1]	2006
Energy consumption per capita (kg oil eq.)	2,644 [1]	2006
Energy intensity (kg oil eq.) per $1,000 (PPP) GDP	103	2006
Renewable electricity production (%)	51.0 [1]	2006

Bangladesh

Economy		
GDP growth rate from previous year (%)	7	2006
GDP per capita ($US)	428	2007
% Value added agriculture, hunting, forestry, fishing	19	2007
% Value added mining, manufacturing, utilities	20	2007
% Value added other	61	2007
Energy		
Energy consumption (1000t oil eq.)	16,966	2006
Energy consumption per capita (kg oil eq.)	120	2006
Energy intensity (kg oil eq.) per $1,000 (PPP) GDP	143	2006
Renewable electricity production (%)	6.0	2006

Statistics: the science of variability

Vive la différence!

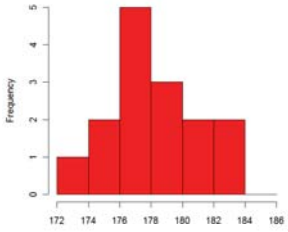
Height distributions in two highschool classes (males) (Milano: 1970')

Class IV A

Class IV B

Analysis 1: Mean Heights are significantly different?

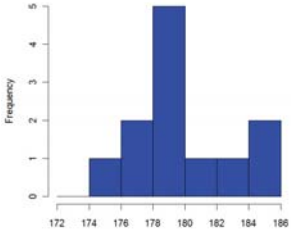
Height: Class IV A



n= 15

Average Height= 178.31cm

Height: Class IV B



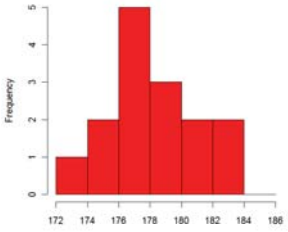
n= 12

Average Height= 180.02cm

Q: The fact that the two averages are different is sufficient evidence to conclude that the distributions generating the two samples have DIFFERENT MEANS?

Analysis 1: Mean Heights are significantly different?

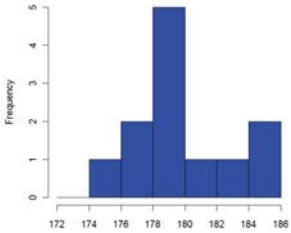
Height: Class IV A



n= 15

Average Height= 178.31cm

Height: Class IV B



n= 12

Average Height= 180.02cm

T-test: p-value = 0.1425

A: NO. There is NOT ENOUGH evidence to conclude that the distributions generating the two samples have different means

The message

Compare the variance

BETWEEN

the two groups with the variance

WITHIN

the two groups

What if VARIABILITY around the averages had been SMALLER?

Height distributions in two classes of comic strips

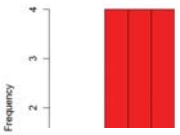
Disney Class

Simpson Class

What if VARIABILITY around the averages had been SMALLER?

Simulated data

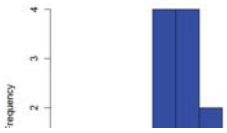
Height: Disney Class

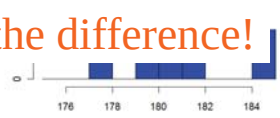


n= 15

Average Height= 178.32cm

Height: Simpson Class





n= 12

Average Height= 180.33cm

Variability makes the difference!

T-test: p-value = 0.001

There is ENOUGH evidence to conclude that the distributions generating the two samples have different means !

Michael Jordan (UC Berkeley) on Big Data: an interwiew on IEEE Spectrum (october 2014)

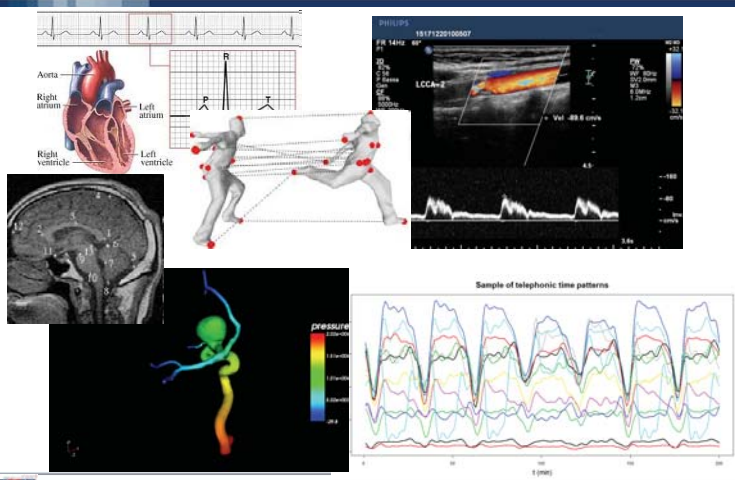
“ And for any particular database, I will find some combination of columns that will predict perfectly any outcome, just by chance alone.

So it's like having billions of monkeys typing. One of them will write Shakespeare.

We have to have error bars around all our predictions.

... if you list all the hypotheses that come out of some analysis of data, some fraction of them will be useful. You just won't know which fraction. ... **unless you're actually doing the full-scale engineering statistical analysis to provide some error bars and quantify the errors, it's gambling.** “

Complex and high dimensional data



A few stories of big data analysis at

MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

Caveat

The analysis of high dimensional, complex data poses new and challenging problems to the statistician:

- Smoothing
- Dimension reduction
- Data alignment and registration
- Landmarks identification
- Classification, Regression, Prediction
- Dependence: spatial or temporal
- Statistical models and methods for non-euclidean data
- Advanced numerical methods for statistical computing
- …

Caveat

The analysis of high dimensional, complex data poses new and challenging problems to the statistician:

- Smoothing
- Dimension reduction
- Data alignment and registration
- Landmarks identification
- Classification, Regression, Prediction
- Dependence: spatial or temporal
- Statistical models and methods for non-euclidean data
- Advanced numerical methods for statistical computing
- …

NO UNIFIED THEORY

I am a data artisan…

…working in a creative workshop

Special Thanks – Statistics@MOX!

https://statistics.mox.polimi.it/people/

Sediments Data from Lake Kassjön

Case study: Sediments Data from Lake Kassjön

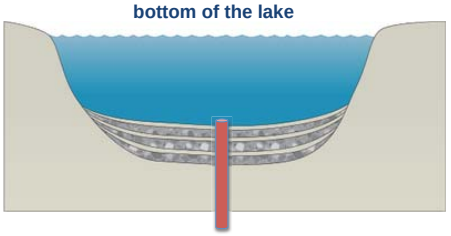
Goal:
Unsupervised classification of misaligned functional data observed on a (1D) lattice.

- Abramowicz, K., Arnqvist, P., Secchi, P., Sjostedt de Luna, S., Vantini, S, Vitelli, V. (2016), Clustering misaligned dependent curves - applied to varved lake sediment for climate reconstruction", forthcoming in *Stochastic Environmental Research and Risk Assessment.*

Case study: Sediments Data from Lake Kassjön

Data: sequence of varves from a sedimentary core drilled at the bottom of the lake

Goal:
Identification of past seasonal climates through the analysis of varves



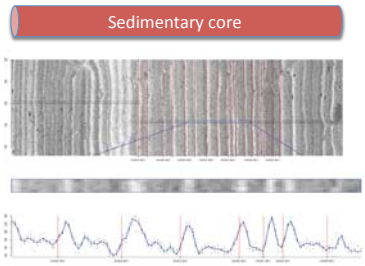
Sedimentary core

K. Abramowicz, P. Arnqvist, P. Secchi, S. Sjostedt de Luna, S. Vantini, V. Vitelli (2016)

Case study: Sediments Data from Lake Kassjön

Annually laminated lake sediments
6385 varves
Years 4486 B.C. – 1901 A.D

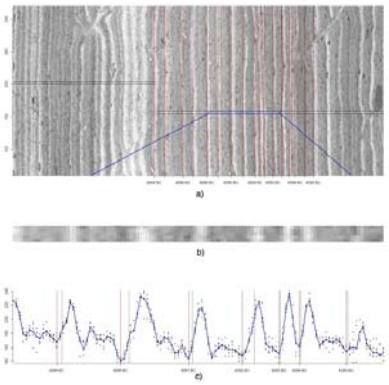
Goal:
Identification of past seasonal climates through the analysis of varves



K. Abramowicz, P. Arnqvist, P. Secchi, S. Sjostedt de Luna, S. Vantini, V. Vitelli (2016)

The varved data

- Each varve reflects wheather conditions and internal biological processes for the year the varve was deposited,
- **In spring**: snow melting and spring runoff → minerogenic deposit (bright layer, high greyscale value)
- **In summer**: lake organism → organic deposit (dark layer, low grayscale value)



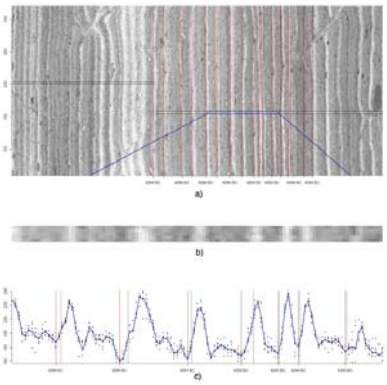
The varved data

- **In spring**: snow melting and spring runoff → minerogenic deposit (bright layer, high greyscale value)
- **In summer**: lake organism → organic deposit (dark layer, low grayscale value)

High spring peak: a winter with a lot of snow

Flat part after spring peak: thick organic sediment, warm summer

Peaks after the spring peak: fall storms with heavy rain

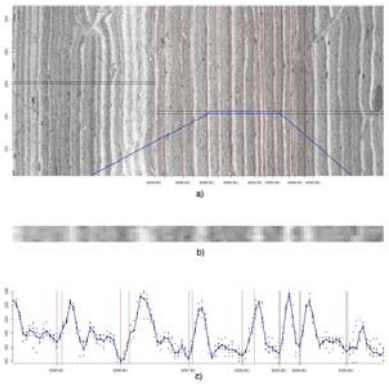


The varved data

- Sedimentation process is continuous: yearly data have been smoothed with cubic B-splines
- Yearly data are functions and they are 1D – spatially (temporally) dependent
- Yearly data are misaligned
- Yearly data are clustered according to different climates

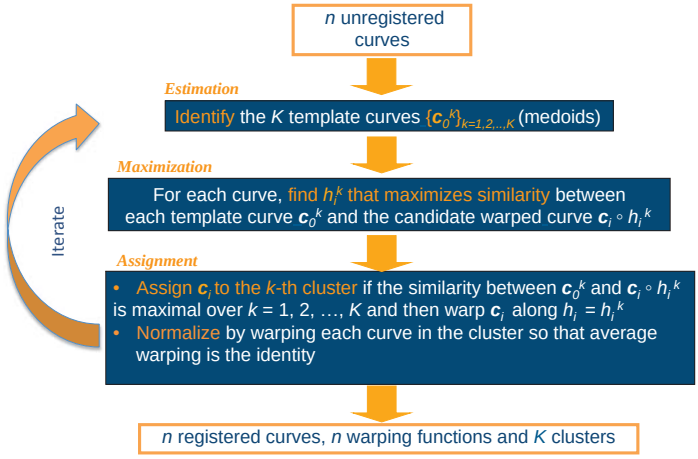
Goal: clusterize, spatially dependent and misaligned functional data

Algorithm: Bagging Voronoi K-Medoid Allignemnt



K-Medoid Alignment at a glance

Sangalli, Secchi, Vantini, Vitelli (2010)



Choosing the size of the Voronoi tesselation

When the final aim is **unsupervised classification**, one can choose ***the number of Voronoi tiles (or equivalently their average size)*** by looking at the **stability** of the final classification as measured by entropy.

Secchi, Vantini, Vitelli (2013)

An **index** for judging the goodness of the final classification, which also exploits the spatial dependence:

A-posteriori minimization of the average spatial **entropy** $\eta_{\mathbf{x}}^K$:

$$\eta_{\mathbf{x}}^K = -\sum_{k=1}^{K} f_{\mathbf{x}}^k * \log(f_{\mathbf{x}}^k)$$

$$\eta^K = \frac{\sum_{\mathbf{x}\in S} \eta_{\mathbf{x}}^K}{\log(K)}$$

Entropy associated to the final distribution of assignment to each of the clusters along iterations.

High entropy => uncertainty about the final classification

Low entropy => very neat final classification

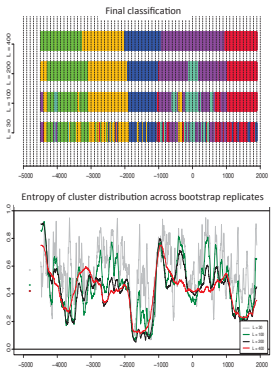
Parameter Tuning

Bagging Voronoi Analysis Parameters

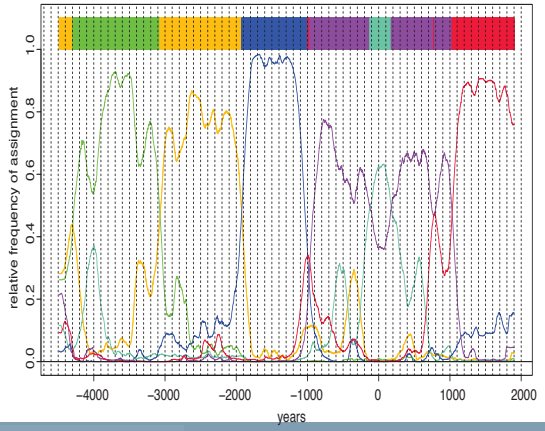
- Seed sampling strategy: uniform
- Average length of Voronoi cells L: L=30,100,200,400
- L=200 chosen by examining entropy

K-mean Alignment Parameters:

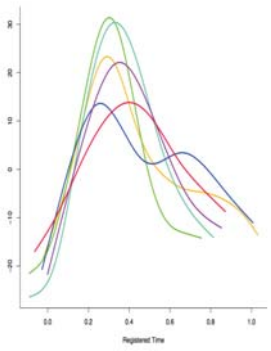
- Number of clusters K: K=1,..,7
- K=7 chosen via Goldilocks principle
- Group of warping functions: positive slope affinities
- Dissimilarity measure: normalized L^2 norm



Climate Clusters



Climate Centroids



BVKMA cluster centroids

The 6 clusters can be associated to 8 major time periods:

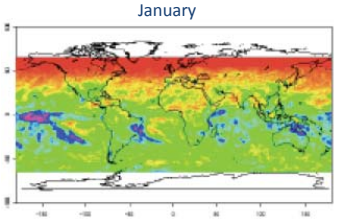
- [4500BC, 4300BC] Lake creation phase (when the lake was separated from the sea due to land uplift) without clear association to climatic factors.
- [4300BC, 3100BC] A sharp high spring peak reflecting large amounts of mineral input linked to rich amounts of snow during winters.
- [3100BC, 1950BC] A less pronounced spring peak and a flatter part thereafter, reflecting a mixture of high, moderate and low spring peaks, and more biological activity during summers.
- [1950BC, 1000BC] Low spring peak indicating warmer winters, and a second lower peak in the flatter part thereafter indicating fall storms and more biological activity during summers.
- [1000BC, 150BC] and [AD150, AD1000] A less pronounced spring peak, perhaps reflecting a mixture of high, moderate and low spring peaks.
- [150BC, AD150] A high spring peak reflecting rich amounts of snow during winters.
- [AD1000, AD1900] A low and wide spring peak, influenced by agricultural activities around the lake the last few hundred years and less from climate.

NASA Irradiance Data Analysis

Case study: NASA irradiance data

Goal:
Unsupervised classification of irradiance temporal profiles (functional data, observed on a spatial lattice)

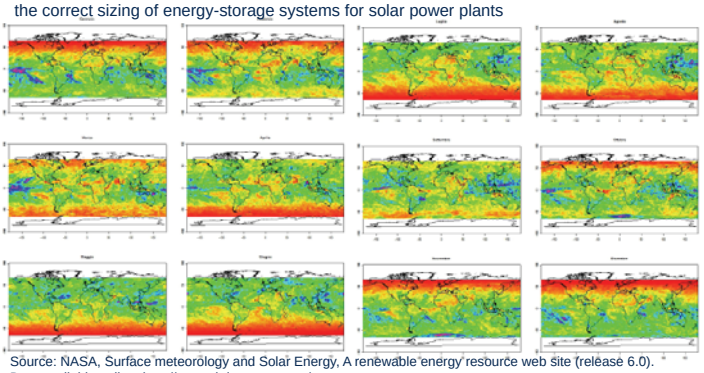
January



Secchi, Vantini, Vitelli (2013)

Case study: NASA irradiance data

Data: maximum deficit below average value of solar radiation incident on a horizontal surface over a consecutive day period (Kwh/m^2) along the year (monthly data computed over the time span from July 1983 to June 2005)
Aim: identify areas of the planet homogeneous with respect to the energy deficit pattern for the correct sizing of energy-storage systems for solar power plants

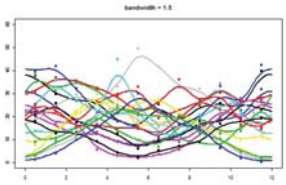


Source: NASA, Surface meteorology and Solar Energy, A renewable energy resource web site (release 6.0).
Data available online: http://eosweb.larc.nasa.gov/.

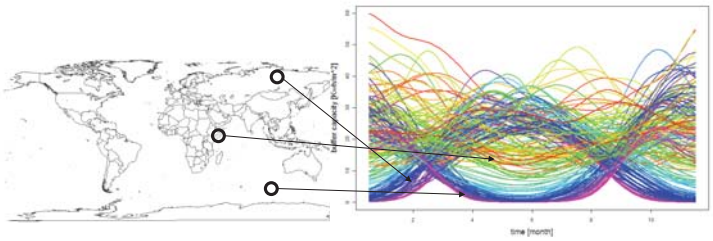
NASA irradiance data

Data: multivariate high-dimensional data indexed by the sites of a spatial lattice (47520 = 132x360 sites)

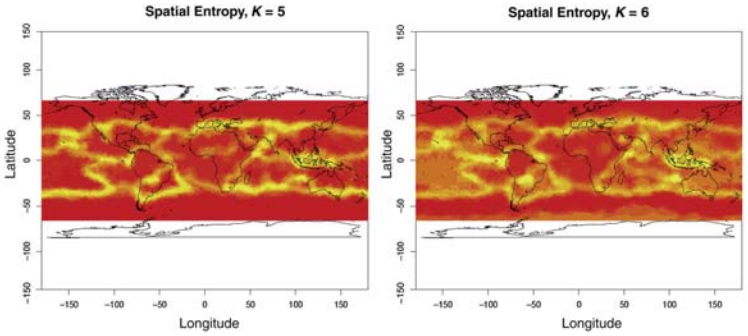
We obtain the **functional form** of the data via a **kernel smoothing** with **Gaussian kernel** with bandwidth ***h = 1.5***



Smoothed data along the year for 200 random sites; rainbow from red (= lat appx 0) to violet (= lat appx 66°)

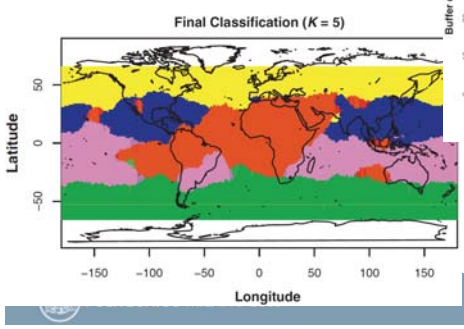


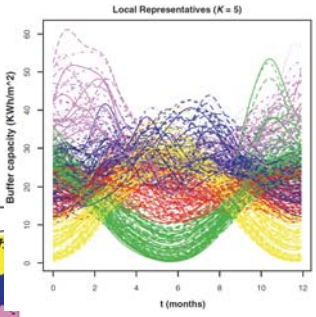
Results: *K=5 or K=6*



Results: *K* = 5

Cluster interpretation
Yellow: Seasonal Low North Cluster
Blue: Seasonal High North Cluster
Red: Non Seasonal Cluster
Violet: Seasonal High South Cluster
Green: Seasonal Low South Cluster

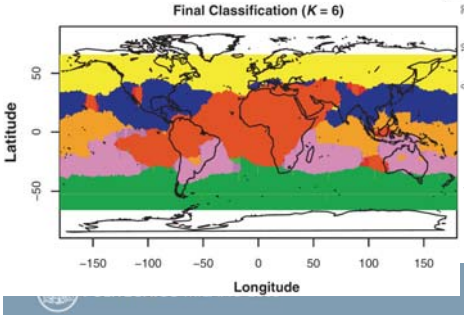


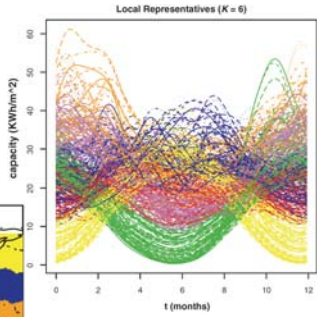


Optimal in an **engineering** perspective:
- the maximal energy demand along the year is the smallest among the clusters;
- constant reliability in time.

Results: *K* = 6

Cluster interpretation
A new cluster (orange) spurts from the former South High (violet) cluster, along the equator.
Characterized by high seasonality, the associated buffer capacity profile makes it strongly unsuited for electricity production by solar power.

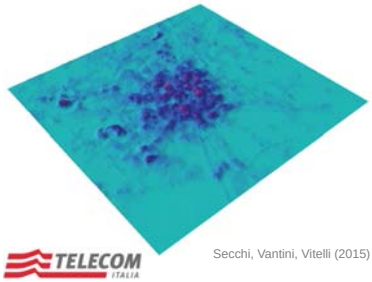




Robustness of the final classification with respect to the choice of *K*: all the other clusters are unchanged, in particular the red one.

Telecom Data Analysis

Goal:
dimensional reduction and representation of functional data observed on a spatial lattice

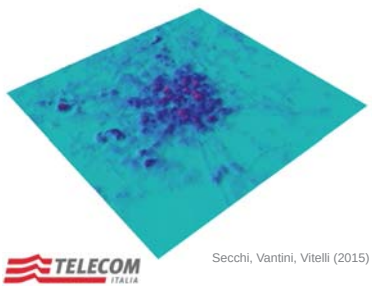


Secchi, Vantini, Vitelli (2015)

Data are courtesy of Convenzione di ricerca
DiAP – Telecom Italia, Politecnico di Milano (Italy)

The Telecom Italia Mobile Network Data Base

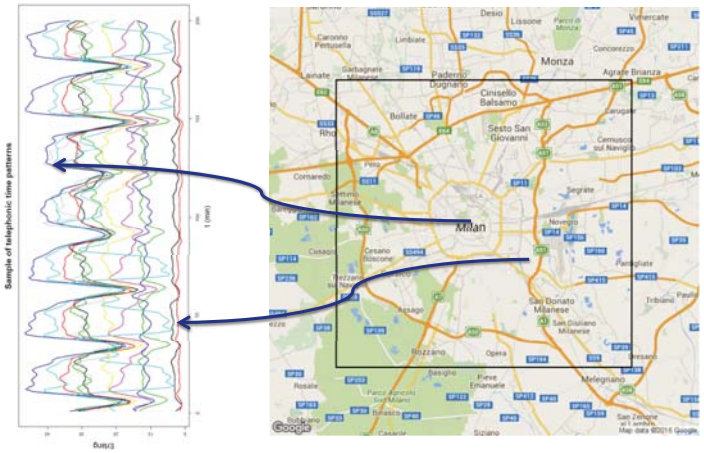
- **Data Description:** Erlang measures along time (every 15 minutes from 18/03/2009, 00:00 a.m., till 24/03/2009, 11:45 p.m.) of the use of the Telecom mobile phone network across a lattice covering the area of Milan (Italy).
- **N = 10573, p = 1308, 13.8M records, 0.1M missing values, plus acquisition errors**
- **Data Features:**
  - Spatial dependence
  - High dimensionality
  - High sample size
  - Presence of macro/micro behaviors
  - Cheap
  - Safe, since data are Eulerian
- **Aim of the Analysis:**
  Identification of subregions of the metropolitan area of Milano which share the same activity pattern along time in terms of population dynamics (“when” and “where”)

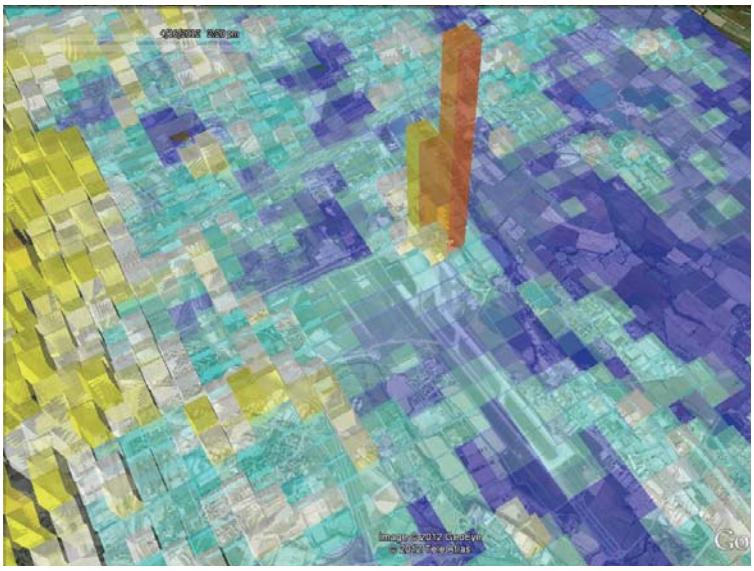


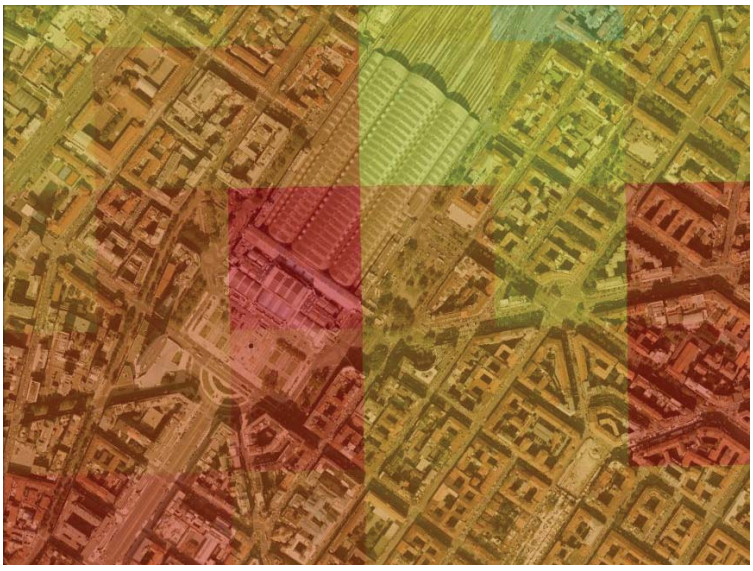
Secchi, Vantini, Vitelli (2015)

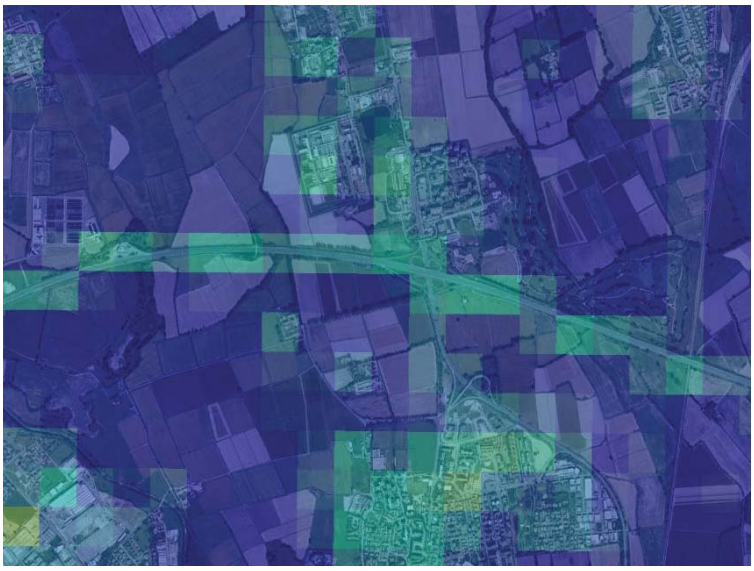
Data are courtesy of Convenzione di ricerca
DiAP – Telecom Italia, Politecnico di Milano (Italy)

The Telecom Italia Mobile Network Data Base

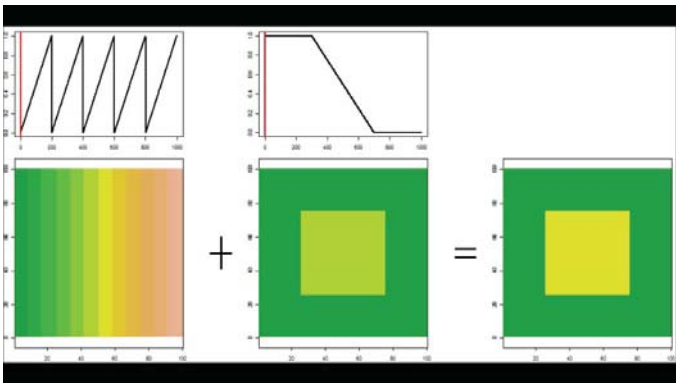






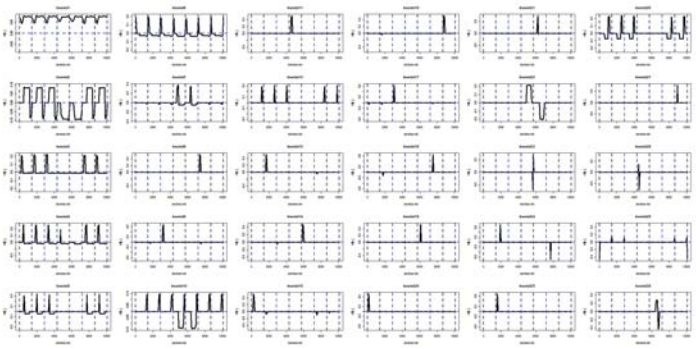


Data representation and dimensional reduction

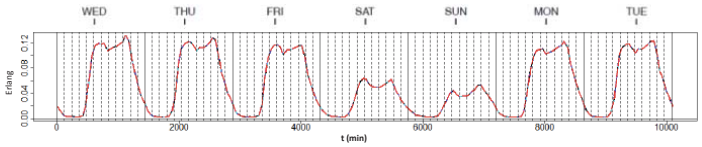


Dimension reduction

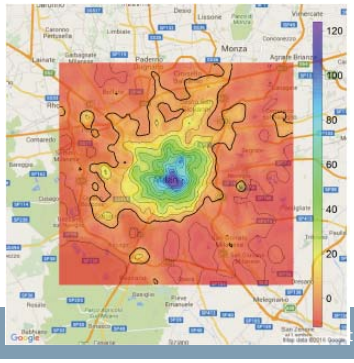
Decomposing within variability and between variability by means of a treelet basis



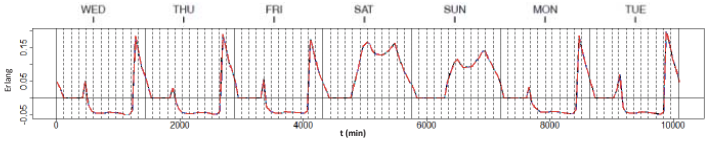
1st Component: Average mobile phone activity



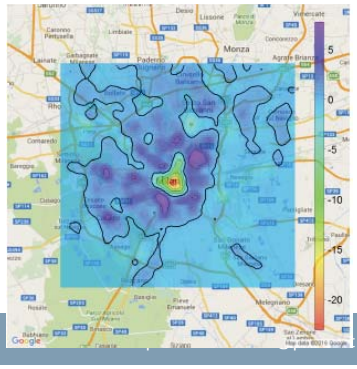
Daytime higly populated Areas
versus
Daytime low density areas



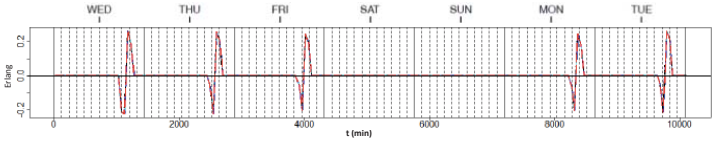
2nd Component: Working/Non Working time



Working Areas
(day-time attractors in working-days)
versus
Non-working Areas
(working-day early morning and evening,
and weekend attractors)

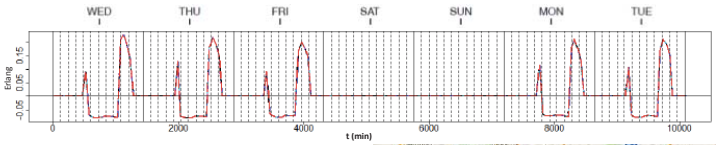


3rd Component: after work attractors

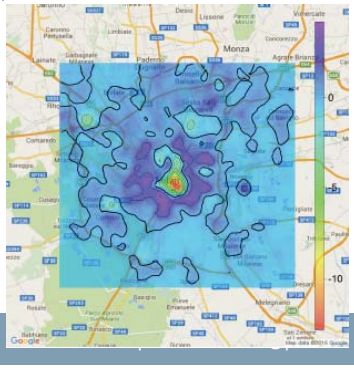


5pm-7pm attractors in working-days
versus
7pm-9pm attractors in working-days

4th Component: working days rush hours



Congested areas at rush hours
versus
Emptier areas during working hours



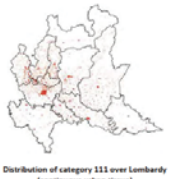
Interpreting the treelets maps through soil use

Passamonti, Secchi, Vantini (2016)

- Data from the **DUSAF** geographical data bank (ERSAF, Regione Lombardia).
- Soil use information obtained through aerial photos, provided as *shapefiles* over Lombardy.
- A specific **soil use category** is assigned to each polygon, following a complex 5-levels structure.

→ Selection of **18** categories as variables, including :

- Continuous urban tissue;
- Agricultural areas;
- Public and private services;
- Industrial and commercial settlements;
- Road networks.



Distribution of category 111 over Lombardy (continuous urban tissue)

piercesare.secchi@polimi.it

DUSAF data – representation examples

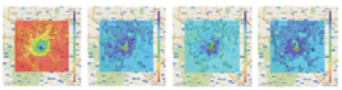
- Agricultural areas (*agri*)
- Road networks (*road*)
- Continuous urban tissue (*urban_cont*)

Relating the two different sources of information

- Objective: **explain** and **support the interpretation** of mobile phone activity spatial patterns (**treelet maps**), using soil use information (**DUSAF data**).

- Global relationship between the two kinds of information
  → **Canonical Correlation Analysis**
- One-by-one explanation of relevant treelet maps
  → **Regression techniques**

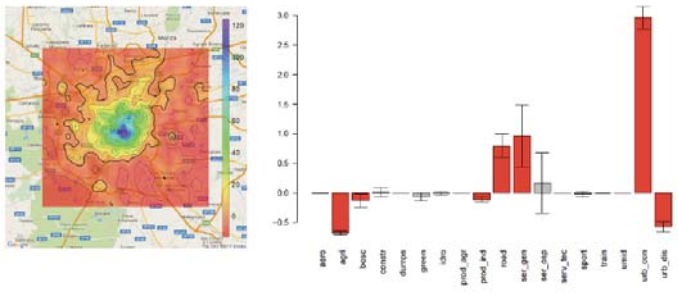
Treelet maps, K=20



18 soil use categories

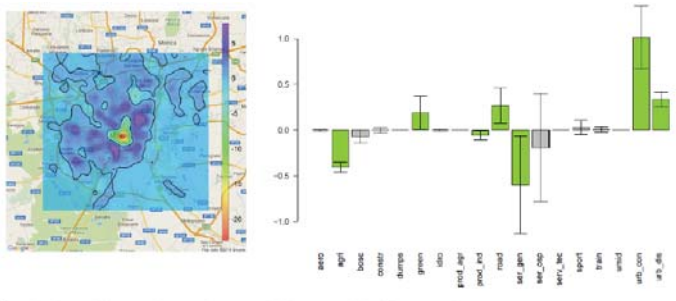
- Spatial dependence must be taken into account in the analyses.

1° component: Average mobile phone activity



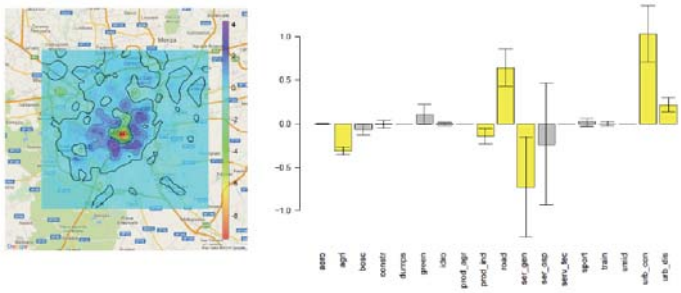
*intervals with semi-width corresponding to one bootstrap standard deviation around the coefficients mean values.

2° component: Working/non Working hours



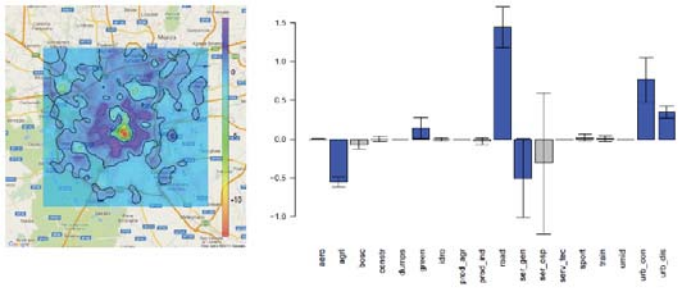
*intervals with semi-width corresponding to one bootstrap standard deviation around the coefficients mean values.

3° component: after work attractors



*intervals with semi-width corresponding to one bootstrap standard deviation around the coefficients mean values.

4° component: working days rush hours



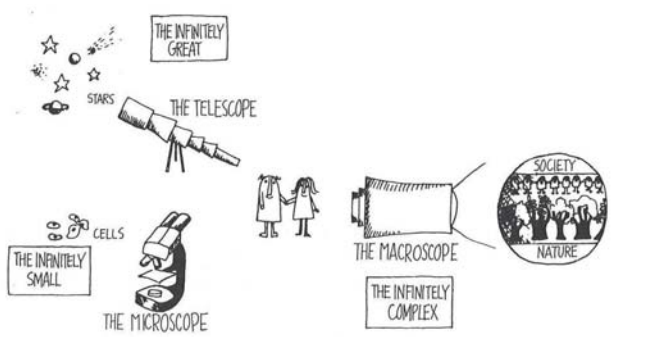
*intervals with semi-width corresponding to one bootstrap standard deviation around the coefficients mean values.

The Urbanscope project

Urban + Macroscope = **Urbanscope**

http://www.urbanscope.polimi.it/

Telescopes, microscopes, macroscopes…



Joël de Rosnay, The Macroscope, 1979

Introduzione: Il senso di Urbanscope

- Urbanscope è un macroscopio per «leggere» lo strato digitale che insiste sulla città (dai dati open, ai social media, ai dati telefonici…).
- Produce viste complementari a quelle generate dagli strumenti esistenti (a partire dai dati amministrativi); emergono segnali più «deboli» ma più tempestivi.
- Urbanscope è progettato per crescere, attraverso l'auspicabile aumento delle fonti di alimentazione e delle "lenti" per la lettura dei dati.

2. Cities into cities: Una città fatta di tante città

Cities into the Cities visualizza la città esplorando le migliaia di messaggi che vengono scambiati a Milano attraverso Twitter. Questa vista, che **segue la lingua dei messaggi**, rivela le tre città digitali attive nella città di Milano.

Le tre città di Twitter:

- una Milano che parla in italiano con se stessa e l'Italia;
- una Milano internazionale che parla in inglese con il resto del mondo;
- una Milano multi-etnica, proiettata verso le nuove comunità cittadine e quelle di origine.

2. Cities into cities (1)

La seconda lente:

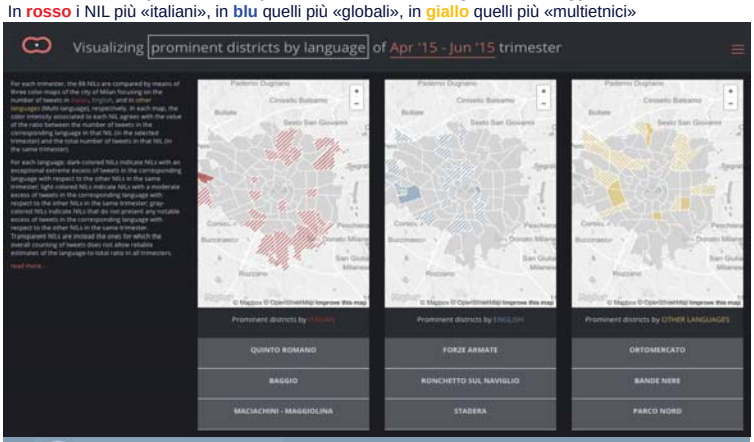
Quali "lingue" caratterizzano i diversi NIL di Milano?

Sono, per esempio, di interesse:

a) Le aree cittadine più rilevanti rispetto ai *tweet* scritti in italiano, oppure in inglese, oppure in altre lingue;
b) La profilazione delle aree cittadine rispetto alle lingue diverse dall'italiano e dall'inglese e come essa cambia nel tempo.

2. Cities into cities (2)

Le città di Twitter: **quali NIL sono più rilevanti** secondo la lingua dei messaggi.
In **rosso** i NIL più «italiani», in **blu** quelli più «globali», in **giallo** quelli più «multietnici»

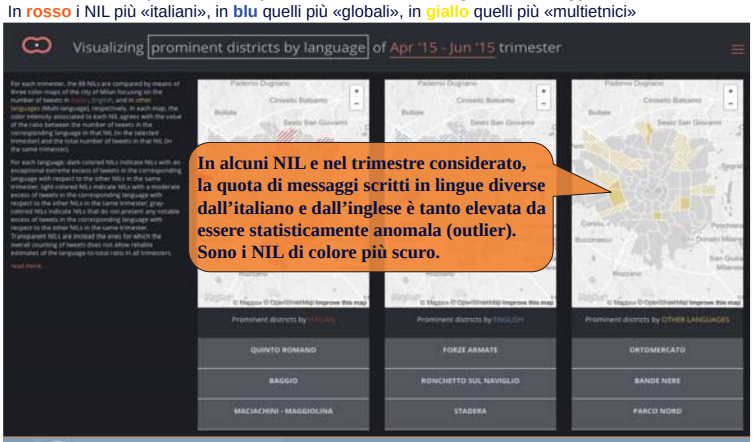


2. Cities into cities (2)

Le città di Twitter: **quali NIL sono più rilevanti** secondo la lingua dei messaggi.
In **rosso** i NIL più «italiani», in **blu** quelli più «globali», in **giallo** quelli più «multietnici»

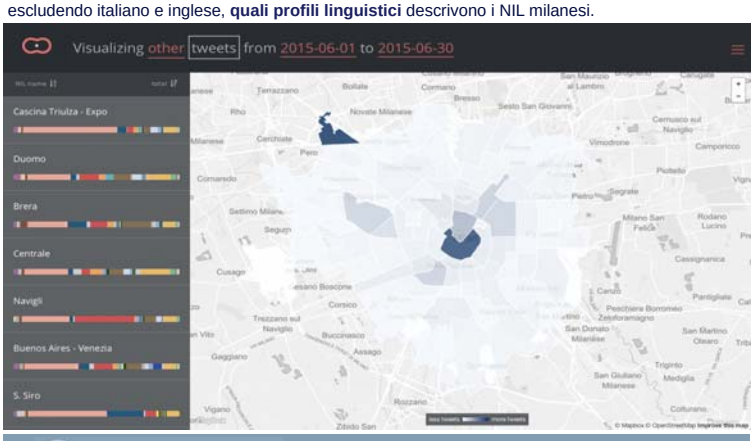
2. Cities into cities (2)

Le città di Twitter: **quali NIL sono più rilevanti** secondo la lingua dei messaggi.
In **rosso** i NIL più «italiani», in **blu** quelli più «globali», in **giallo** quelli più «multietnici»



2. Cities into cities (3)

Le città di Twitter:
escludendo italiano e inglese, **quali profili linguistici** descrivono i NIL milanesi.

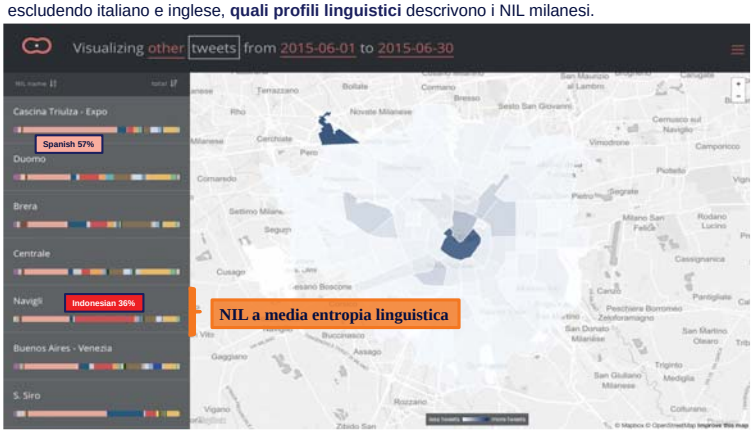


2. Cities into cities (3)

Le città di Twitter:
escludendo italiano e inglese, **quali profili linguistici** descrivono i NIL milanesi.

2. Cities into cities (3)

Le città di Twitter:
escludendo italiano e inglese, **quali profili linguistici** descrivono i NIL milanesi.



2. Cities into cities (3)

Le città di Twitter:
escludendo italiano e inglese, **quali profili linguistici** descrivono i NIL milanesi.

2. Cities into cities (3)

Le città di Twitter:
escludendo italiano e inglese, **quali profili linguistici** descrivono i NIL milanesi.

3. City Magnets

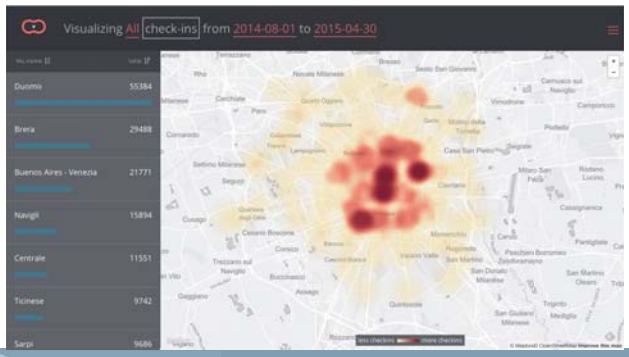
La terza lente:

Quali sono gli attrattori della città?

- La sezione City Magnets mostra quali sono i luoghi più frequentati e condivisi dagli utenti di *Foursquare* (oggi anche *Swarm*).

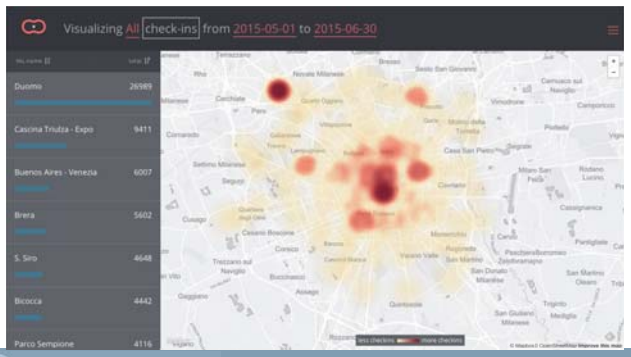
3. City Magnets e EXPO (1)

- La Mappa pre-Expo. Emergono gli attrattori “classici”: Duomo, Brera, Navigli…



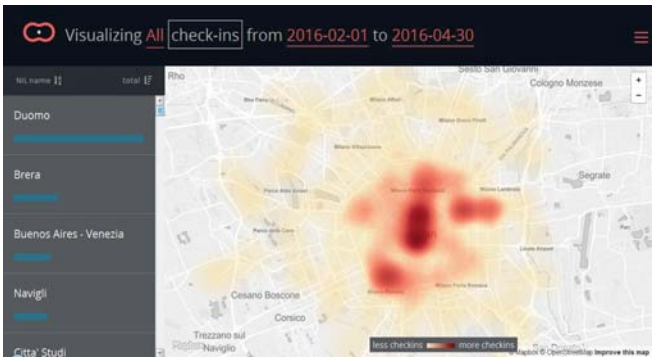
3. City Magnets e EXPO (2)

- La Mappa durante Expo. Expo come nuovo attrattore, che non cancella però quelli “classici”: Expo non svuota la città



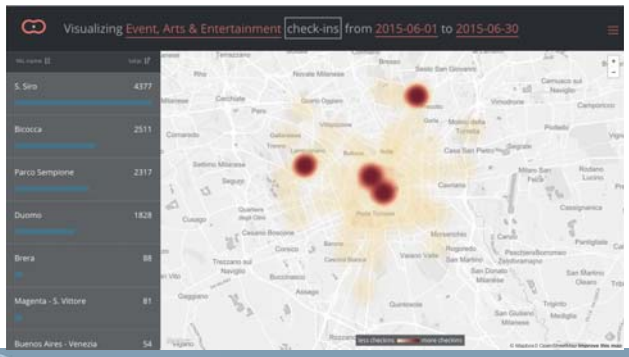
3. City Magnets e EXPO (3)

La Mappa dopo Expo: cosa succederà col Post Expo?



3. City Magnets e singoli eventi

- Mappa durante altri eventi
  (emerge S.Siro: Events, concerti di giugno 2015 – Vasco, Jovanotti…)



4. Top Venues

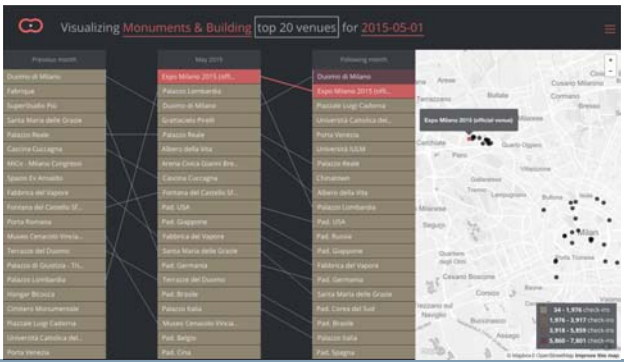
La quarta lente:

Quali sono i luoghi più “cool” di Milano?

- L’analisi dei *checkins* per tipologia di “venue” consente di vedere dove le persone vogliono far sapere di essere

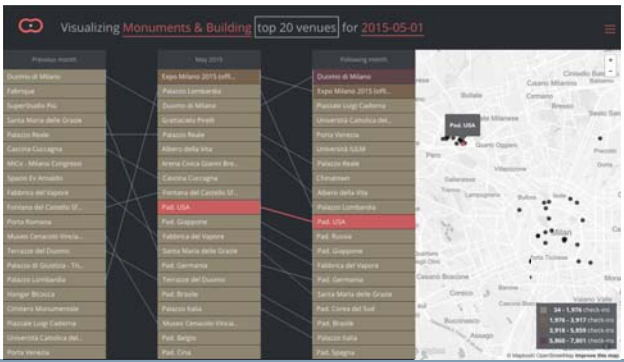
4. Top Venues: L’emergere di Expo (1)

- Le colonne rappresentano i luoghi con maggiori *checkins* nei mesi di aprile, maggio e giugno



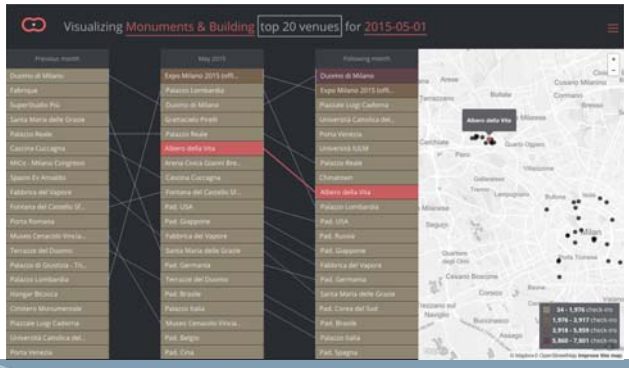
4. Top Venues: L’emergere di Expo (2)

- I Padiglioni



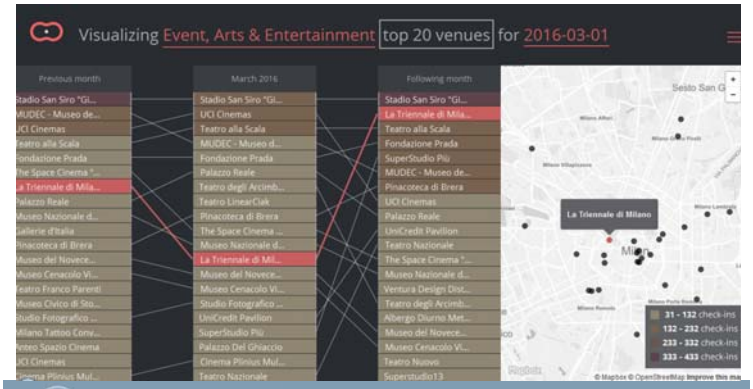
4. Top Venues: L'emergere di Expo (3)

- L'albero della vita



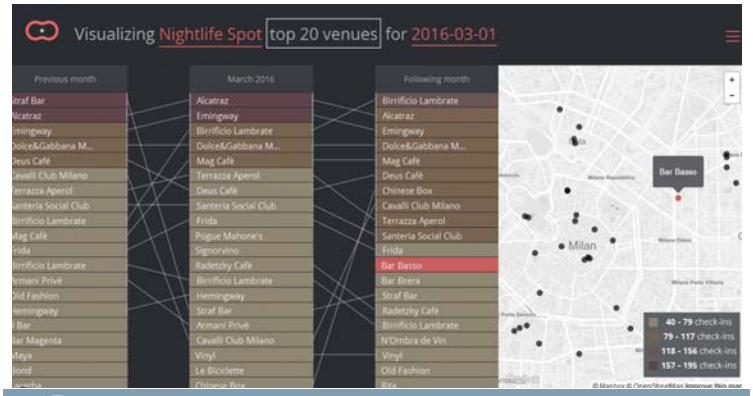
4. Top Venues: Gli eventi (1)

- La sensibilità dello strumento nel leggere gli eventi: Aprile 2016, mese del design



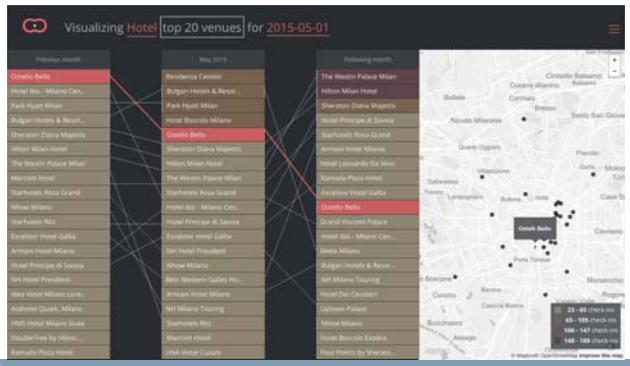
4. Top Venues: Nighlife

- I luoghi dei designer ad Aprile 2016 – Compare Bar Basso



4. Top Venues: Gli Hotel durante l'Expo (1)

- Decresce la frequenza in hotel di categoria minore

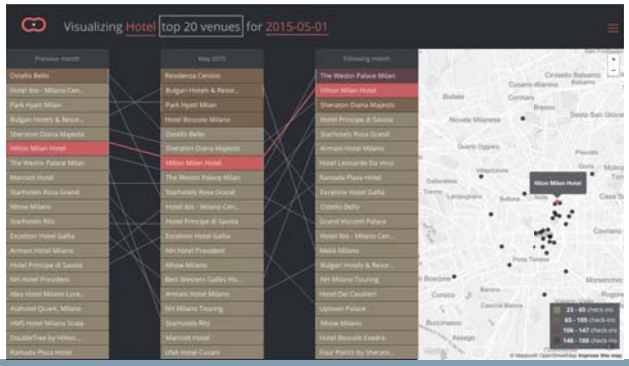


4. Top Venues: Gli Hotel durante l'Expo (2)

- Decresce la frequenza in hotel di categoria minore

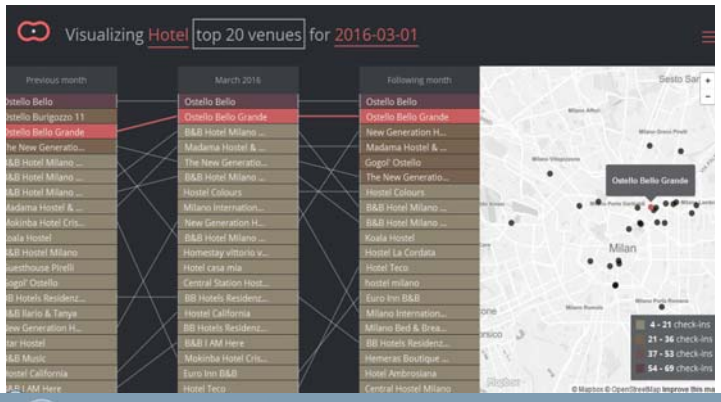
4. Top Venues: Gli Hotel durante l'Expo (3)

- Incremento della frequenza di Check-in negli alberghi di categoria elevata in coincidenza dell'apertura di Expo



4. Top Venues: Gli Hotel per altri eventi

- Tornano hotel di categoria inferiore



«The greatest value of a picture is when it forces us to notice what we never expected to see»

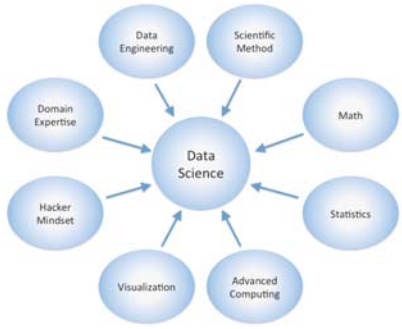
John W. Tukey. Exploratory Data Analysis. 1977

Il gruppo di ricerca: CompleXCity LaB

- CompleXCity Lab è un laboratorio interdipartimentale del Politecnico di Milano contribuisce a rinnovare il rapporto tra conoscenza e azione nei processi decisionali pubblici e privati che riguardano e producono la città e l'urbano, come processi spaziali, sociali, economici, istituzionali complessi.
- I dipartimenti coinvolti sono:
  - Dipartimento di Architettura e Studi Urbani (DASTU)
  - Dipartimento di Design (DESIGN)
  - Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)
  - Dipartimento di Ingegneria Gestionale (DIG)
  - Dipartimento di Matematica (DMAT)

Valorizzare i Big Data per DECISIONI COMPLESSE che insistono sull'urbano e per innovare la comunicazione agli STAKEHOLDER

A new profession: the data scientist



The key word in "Data Science" is not Data, it is Science

How to become a data scientist

http://www.mate.polimi.it/

Contatti

- **Piercesare Secchi:**
piercesare.secchi@polimi.it

- **Politecnico di Milano:**
http://www.polimi.it

- **Dipartimento di Matematica al Politecnico di Milano:**
https://www.mate.polimi.it

- **Laboratorio MOX:**
https://mox.polimi.it

- **Orientamento Politecnico di Milano:**
http://www.poliorientami.polimi.it

My references for this talk

- Abramowicz, K., Arnqvist, P., Secchi, P., Sjostedt de Luna, S., Vantini, S, Vitelli, V. (2016), Clustering misaligned dependent curves - applied to varved lake sediment for climate reconstruction", forthcoming in *Stochastic Environmental Research and Risk Assessment.*
- Menafoglio, A., Secchi, P., Dalla Rosa, M. (2013), A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics* 7, 2209–2240.
- Passamonti, F. (2016), *Spatio-temporal mobile phone data in Milan: Bagging-Voronoi exploration and modeling through soil use and land cover data*. MSc. Thesis, Politecnico di Milano
- D. Pigoli, A. Menafoglio, P. Secchi (2016). Kriging prediction for manifold valued random fields. *Journal of Multivariate Analysis*, 145, 117-131.
- Secchi, P., Vantini, S., Vitelli, V. (2013), Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, vol. 22, p. 53-64.
- Secchi, P., Vantini, S., Vitelli, V. (2015), Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Methods & Applications*, vol. 24 (2), p. 279-300.
- Secchi, P., Menafoglio A. (2016), Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics. *Manuscript.*