DMS StatLab

Data Methods and Systems Statistical Laboratory

DEPARTMENT OF ECONOMICS AND MANAGEMENT

UNIVERSITY OF BRESCIA

DSSR 2020

Dynamic crowding maps with mobile phone big data

Maurizio Carpita[1], Rodolfo Metulini[2]

1. Data Methods and Systems Statistical Laboratory - Department of Economics and Management, University of Brescia
2. Department of Economics and Statistics - University of Salerno

Third international conference on Data Science & Social Research

December 10-11, 2020

The project

- **Ongoing project ('till 06/2022):**

This talk describes the works conducted together with Prof. Roberto Ranzi and Dr. Matteo Balistrocchi (*Department of Civil, Environmental, Architectural Engineering and Mathematics, UNIBS*) in the context of **MoSoRe** project

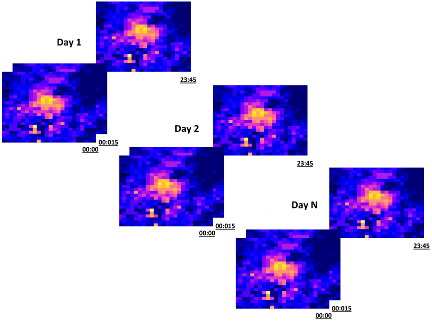
Regione Lombardia, Call HUB Research & Innovation: *Infrastrutture e servizi per la Mobilità Sostenibile e Resiliente* - *MoSoRe@Unibs*ID 1180965 - POR FESR 2014-2020

- **Scientific output:**
  1. Metulini, R., Carpita, M., (2020), A Spatio-Temporal Indicator for City Users based on Mobile Phone Signals and Administrative Data - Social Indicator Research, 1-21. DOI: 10.1007/s11205-020-02355-2
  2. Balistrocchi, M., Metulini, R., Carpita, M., and Ranzi, R.: Dynamic maps of people exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, 2020, in press. DOI: 10.5194/nhess-2020-201.

The context of application

- **Floods** are natural phenomena whose hazards afflict nearly 20 million people worldwide (Kellens et al., 2013), posing a serious challenge to the protection of human lives.
- **Urbanization** determines dramatic increases in **people exposure** and vulnerability to floods, since most of recent urbanizations are developed in **flood prone areas**.
- The development of effective **emergency management plans** are intended to provide communities with **early warnings**, reliable **real-time information**.
- We provide a detailed and reliable picture of the real-time spatiotemporal variability of the flood risk by proxying it with **dynamic crowding maps from mobile phone data** for reference groups of days.

Data



- ***Erlang mobile phone measures*** (Erlang, 1909): average number of mobile phone users (MPU) bearing a SIM connected to the network, recorded at constant time steps with reference to a georeferenced grid of square cells.

  Available for Telecom Ialia Mobile (TIM) in the period from 04/2014 to 08/2016 thanks to a collaboration with *Statistical Office* of *Comune di Brescia*.

- ***Census data*** from ISTAT, reporting residential population (01/01/2016) by age, for each *sezione di censimento* (SC)

The set-up

- To detect MPU spatiotemporal variability we define the subject of our analysis: the ***daily density profiles*** (DDP).
- Let e_{it} be the number of MPU in the $i-th$ grid cell in a generic time interval t,
- let $I_r = \{i_1, ..., i_m\}$ be the set of grid cells in region r of interest,
- let $T_d = \{t_1, ..., t_o\}$ be the set of intervals of time in a day d.
- DDP_{rd} can be defined as the vector of the sums of MPU (a sum for each considered time instant) in region r and day d (length $= o$)

$$DDP_{rd} = \left(\sum_{l=1}^{m} e_{il,t_1}, \sum_{l=1}^{m} e_{il,t_2}, ..., \sum_{l=1}^{m} e_{il,t_o} \right)'$$

- **Goal**: classifying the occurrences in the time series of DDP_{rd} related to the set $d = \{d_1, ..., d_n\}$ of n analyzed days. In other words, clustering similar DDP_{rd}.

Issues

- Our dataset amount to n observations (days) and $p = m * o$ features per day (cells$*$quarters).

  Let consider one year of data ($n = 365$): $o = 96$ (quarters per day), $m = 400$ (grid's cells of the sample area).
- Number of features is larger than number of observations, so we refer to an high-dimensional data setup (Donoho, 2000).
- Traditional techniques (Arabie and De Soete, 1996) may not return robust results in high-dimensional data, for example due to the presence of the curse of dimensionality (Keogh and Mueen, 2017).
- Bouveyron et al. (2007) addressed this issue with regard to clustering. However, as suggested by Jovi et al. (2015) , a suitable solution is represented by a preliminar data reduction strategy.
- ***Histogram of Oriented Gradients*** (HOG) approach is used for data reduction.

The Strategy ...

(... to take into account days' similarity)

Step	Type	Aim	Method	Features
1	Data reduction + clustering	find similar raster images	HOG + k-means cluster	HOG features
2	clustering	find similar functional curves	functional model-based clustering	DDP features
3	population assessment	estimate city users	spatial match of MPU and census data	DDP features + population
4	visualization	find reference daily profiles	functional box plots	DDP features

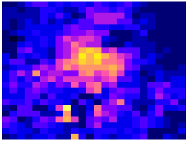
The Strategy

Step	Type	Aim	Method	Features
1	Data reduction + clustering	find similar raster images	HOG + k-means cluster	HOG features

HOG data reduction

- for a given t, let $\epsilon_{it} = \{e_{1,t}, e_{2,t}, ..., e_{i_m,t}\}'$ be the MPU vector of region r in time instant t (dimension m).
- ***Aim:*** to reduce ϵ_{it} to a smaller vector of values $\kappa_{1,t}$ $(m' < m)$, with the relevant information contained in ϵ_{it}.
- To do so, set ϵ_{it}, separately for each t, undergoes a histogram of oriented gradients (HOG) feature extraction (Dalal and Triggs, 2005).
- Vector $z_{it} = \{e_{i,t}/max_{i \in I_r}(e_{i,t})\}, \forall i \ in \ I_r$ undergoes HOG
- HOG method:
  1. split the m cells of the grid in S smaller grids $G_1, ..., G_S$ $(Gi \cap Gj = \emptyset, \forall i = 1, ..., S$ and $\forall j = 1, ..., S$ with $i \neq j)$ ($\sqrt{S}$ is a parameter to be chosen),
  2. for each grid G_i, *direction* and *magnitude* gradient matrices are computed (Dalal and Triggs, 2005).
  3. from the two gradient matrices, histogram of gradients is determined, with k equal bins (with k a parameter to be chosen).
- κ_{it} is stacked over the subscript t, in order to derive (for region r, day d) the vector of features κ_d (dimension $S * k * o$), $d = 1, ..., n$.

HOG data reduction ... explained

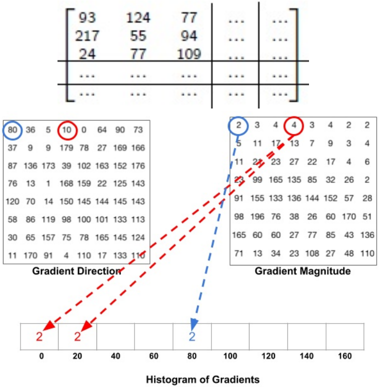


From a *nxn* raster data ....

$$\begin{bmatrix} 93 & 124 & 77 & \dots & \dots \\ 217 & 55 & 94 & \dots & \dots \\ 24 & 77 & 109 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

...to X_t, a matrix representing the number of people in that cell at time t

1. standardize MPU data;
2. split matrix in sub-matrices;
3. for each sub-matrix, compute the matrices of gradients (using the sobel operator);
4. assign each value of the direction matrix to one of the k bins of the histogram using its magnitude as weight, to produce the vector of features;
5. stack into a vector the features of all quarters of the day.



First step clustering

- Days are clustered in terms of how MPU are distributed over region r according to index i, i.e., according to similarity in the raster image.
- The objects to be clustered are the n days and κ_d contains the $S * k * o$ (with $S * k * o < m * o$) features for day d, $\forall d = d_1, ..., d_n$.
- **k-mean cluster** method (Hartigan and Wong, 1979) is adopted (after having tested against curse of dimensionality)
- According to Hartigan and Wong criterion, the clusters' number H is chosen by minimizing the ratio between the total within sum of squares and the total sum of squares for different values of H.

The Strategy

Step	Type	Aim	Method	Features
2	clustering	find similar functional curves	functional model-based clustering	DDP features

Second step clustering

- ***Aim***: at considering similarity in the functional form of the DDP_{rd}, if viewed as functional curves.
- We consider DDP_{rd} as the collection of functional observations $x_{rd}(T_d)$, $T_d \in (t_1, ..., t_o)$ (length o) (i.e. $\sum_{l=1}^{m} e_{il,t_1}$ in t_1), with d varying in $d = \{d_1, ..., d_n\}$.
- We adopt a **model-based functional data clustering** method (MB-FAC, Bouveyron et al., 2015), which provides estimated curve with specific parameters, to group days d (cluster's objects) in terms of the o DDP_{rd} values (cluster's variables)
- We adopt the following path:
  1. **functional data outlier detection by likelihood ratio test** (LRT) to remove anomalous DDP_{rd}, as proposed by Febrero-Bande et al. (2008);
  2. Bouveyron et al. (2015) clustering method, using `funFEM` package in `R`
- The method suits for high-dimensional data: it employs sub-space clustering criterion (Agrawal et al., 1998, it considers just the minimum number of variables for grouping objects)

The Strategy

Step	Type	Aim	Method	Features
3	population assessment	estimate city users	spatial match of MPU and census data	DDP features + population

Population assessment - I

- ***Aim***: to estimate the total amount of people (***city users***), while MPU availability regards just one mobile phone company.
- We compute an estimate of the ***market share*** of the provider company, to correct the DDP_{rd},
- by comparing the number of residents from archives with the number of TIM users on a residential area - in late evening hours (assuming that, in late evening hours, residential Sezione di Censimento (SC) are only populated by residents).
- MPU grid is made of square cells while SCs are irregular polygons $\rightarrow$ the number of TIM users belonging to each SC needs to be retrieved by intersecting the two sources.
- the portion of the cell belonging to the *SC* polygon were calculated in order to count how many TIM users are present in each polygon, by using the function extract in `raster` package, `R`.

Population assessment - II

- Let $Cell_j, \forall j = 1, 2, ..., J_{SC}$ be the cells of the sample area, the ratio

$$A_j = \frac{area(SC) \cap area(Cell_j)}{area(Cell_j)}$$

represents how much of $Cell_j$ is included in the chosen SC;
- let TUC_j be the MPU in $Cell_j$, the estimation of the number of MPU in SC is

$$ETU_{SC} = \sum_j TUC_j * A_j$$

.
- The estimated company market share in SC is given by

$$ETMS_{SC} = \frac{ETU_{SC}}{P_{SC}}$$

where P_{SC} is the resident number for that SC (children and elderly people excluded).
- The **median** (me(.)) of $ETMS_{SC}$ can be used as a proxy for the company market share at city level;
- the city users estimate is given by

$$D\hat{D}P_{rd} = \frac{DDP_{rd}}{me(ETMS_{SC})}$$

The Strategy

Step	Type	Aim	Method	Features
4	visualization	find reference daily profiles	functional box plots	DDP features

Visualization

- Let consider DDP_{rd} to be a functional curve $x_{rd}(T_d)$ displaying, in the y-axis, the sum of MPU in region r and day d with respect to, in the x-axis, time instants $T_d \in (t_1, ..., t_o)$.
- Functional box plots (FBP, Sun and Genton, 2011) can be used to display the profile for each final cluster.
- For cluster h, let $d_h = \{d_{1h}, ..., d_{nh}\}$ be the group of days belonging to cluster h, and let $D\hat{D}P_{rd,h} = [D\hat{D}P_{rd_1,h}, ..., D\hat{D}P_{rd_n,h}]$ be the matrix of dimension $o * n_h$ with a DDP_{rd} of cluster h in each column.
- By considering each DDP_{rd} a curve, the FBP representing the profile plot of the total number of people (that we call *city users*) in different hours (with DB), for cluster h, is computed using matrix $DDP_{rd,h}$

Case study description

- WGS 84 UTM 32 N coordinates: 5,040,920–5,049,980N, 585,970–592,970E (area about 64 km^2) centred on the Mandolossa-Gandovere network (grid of 20x20 150m^2 cells)
- at 15-minutes intervals (**quarters**) over the period **July 1st, 2015** - **August 10th, 2016**.

- After imputing missing quarters and removing the full day when they are too many, we ended up with a number of valid **360 days**.
- HOG parameters: $\sqrt{S} = 3$, $h = 4$.
- The interest is in residential and industrial part of 4 specific areas (Moie di Sotto, Villaggio Badia and Fantasina, southern Gandovere canal, Roncadelle)

First step clustering: results

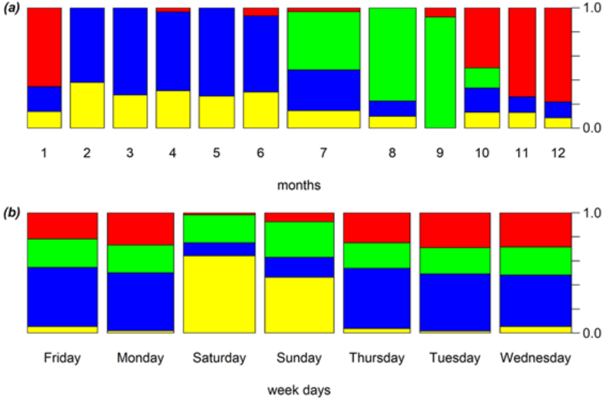


Figure: Spine-plots representing the first-step clustering of days along (a) months and (b) days of the week (green: all days mostly occurring in July, August and September; blue: working days mostly occurring from February to June; red: working days mostly occurring from October to January; yellow: weekends mostly occurring from October to June)

Representation: results

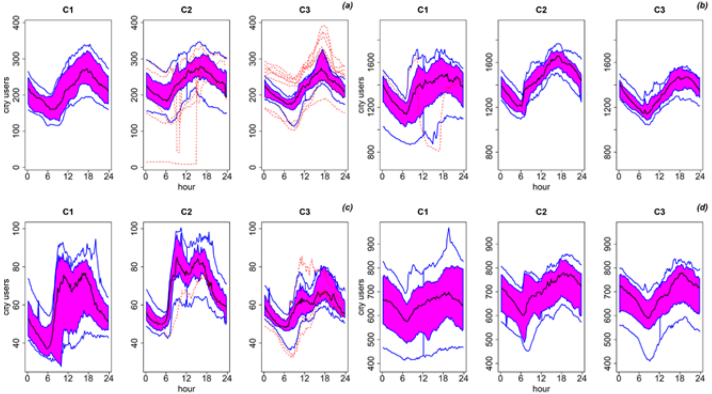


Figure: Functional box plots of exposed people ("city users") inside **residential areas**: (a) Moie di Sotto, (b) Villaggio Badia and Fantasina, (c) southern Gandovere canal, (d) Roncadelle . **Cluster 1 (July, August, September, C1)**, **Cluster 2 (working-days from October to June, C2)**, **Cluster 3 (week-ends from October to June, C3)**

Representation: results - II

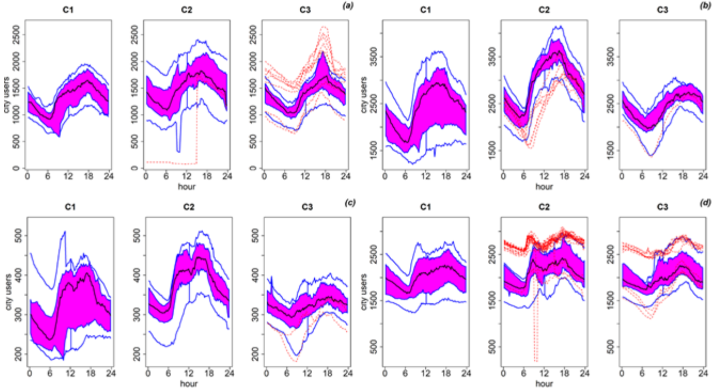


Figure: Functional box plots of exposed people ("city users") inside **industrial-commercial settlements**: (a) Moie di Sotto, (b) Villaggio Badia and Fantasina, (c) southern Gandovere canal, (d) Roncadelle. **Cluster 1 (July, August, September, C1)**, **Cluster 2 (working-days from October to June, C2)**, **Cluster 3 (week-ends from October to June, C3)**

Discussion

- The combination of:
  1. high spatial resolution (150 m^2) and short time step ($15'$) of data, and
  2. the application of the proposed statistical strategy thought for high dimensional data

  permits a
  1. reliable population assessments even for small area, and
  2. a precise evaluation of the temporal dynamic of city users in the sample area
- Functional box plot results are meaningful:
  1. working days and weekends show different temporal dynamics, when they belong to working months (October to June),
  2. daily dynamics in summer months (July, August and September, holydays in Italy), must be regarded as different from the others,
  3. working days and weekends feature more similar daily density profiles during such months.

References - I

1. Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, ACM Press, 94–105, doi:10.1145/276304.276314, 1998
2. Arabie, P., and De Soete, G.: Clustering and classification, World Scientific, doi:10.1142/1930, 1996
3. Bouveyron, C., Girard, S., and Schmid, C.: High-dimensional data clustering, Comput. Stat. Data An., 52(1), 502–519, doi:10.1016/j.csda.2007.02.009, 2007
4. Bouveyron, C., Come, E., & Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. The Annals of Applied Statistics, 9(4), 1726-1760.
5. Dalal, N., and Triggs, B.: Histograms of oriented gradients for human detection, in: Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPR '05), doi:10.1109/CVPR.2005.177, 2005.
6. Donoho, D. L.: High-dimensional data analysis: The curses and blessings of dimensionality, AMS Math Challenges Lecture, 1–32, 2000
7. Erlang, A. K. (1909). The theory of probabilities and telephone conversations. Nyt. Tidsskr. Mat. Ser. B, 20, 33-39.

References - II

1. Febrero-Bande, M., Galeano, P., & Gonzalez-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics: The official journal of the International Environmetrics Society, 19(4), 331-345.
2. Hartigan, J. A., and Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm, J. R. Stat. Soc., Series C (Applied Statistics), 28(1), 100–108, doi:10.2307/2346830, 1979.
3. Kellens, W., Terpstra, T., and De Maeyer, P.: Perception and communication of flood risks: A systematic review of empirical research, Risk Anal., 33/1, 24–49, doi:10.1111/j.1539-6924.2012.01844.x, 2013
4. Keogh, E., and Mueen, A.: Curse of dimensionality, in: Sammut, C., and Webb, G. I. (eds.), Encyclopedia of Machine Learning and Data Mining, 314–315, Springer, doi:10.1007/978-1-4899-7687-1-192, 2017.
5. Jovi, A., Brki, K., and Bogunovi, N.: A review of feature selection methods with applications, in: Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1200–1205, doi:10.1109/MIPRO.2015.7160458, 2015
6. Sun, Y., & Genton, M. G. (2011). Functional boxplots. Journal of Computational and Graphical Statistics, 20(2), 316-334.

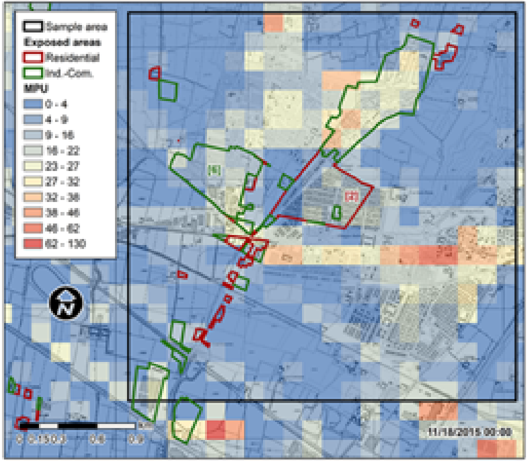


Figure: Snapshots of a dynamic map showing the spatiotemporal distribution of mobile phone users (MPU) occurred at 12pm, 17/11/2015 (Wednesday); base map Lombardy Regional Technical Map CTR 1:5000 provided by Lombardy Region (www.geoportale.regione.lombardia.it).

Back to slide

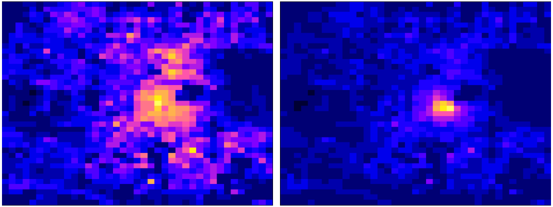


Figure: Example of dissimilarity among raster images.

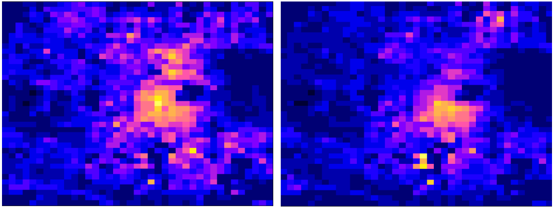


Figure: Example of similarity among raster images.

Back to slide

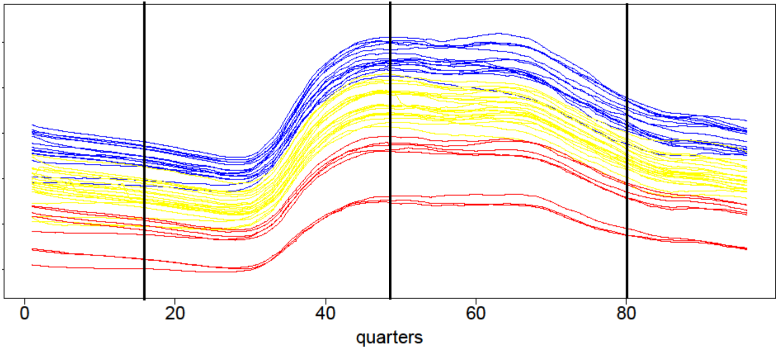


Figure: Example of similauty and dissimilaity in the functional form. Curves with the same colors are similar. On the contrary, curves with different colors are dissimilar.

Back to slide

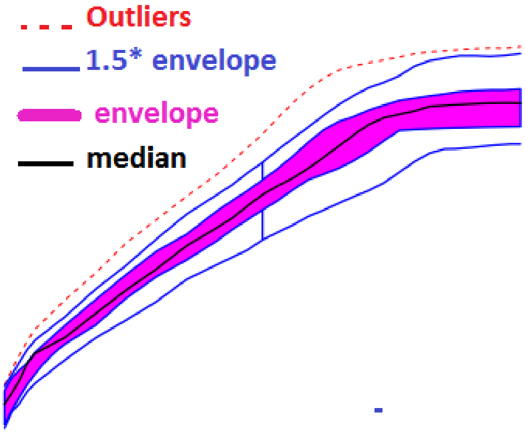


Figure:

Back to slide

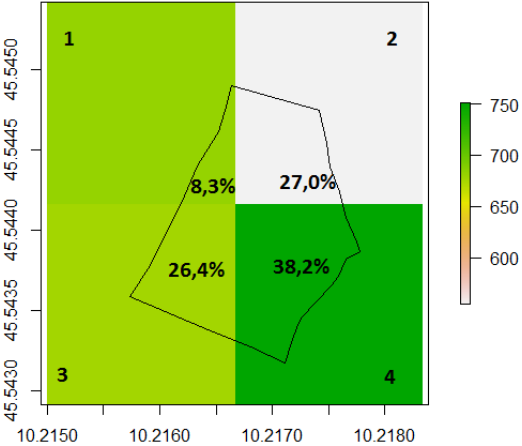


Figure: Example of weighting scheme to assign the number of TIM users to SC 110, located at latitude 45.544 N and longitude 10.217 N

Back to slide