Population Density Estimation from Mobile Data

Ghazaleh Khodabandelou

October 26, 2015

Outline

- Research focuses: inputs, goals, methods
- State of the art
- Preliminary results (Milano mobile phone data)

Research Focuses

Inputs: Mobile phone data + Census, land-use maps, users' activities/purpose, etc.

Goals:

- Quantifying crowd size: crowd disasters, events and anomaly detection, human mobility prediction, health, etc.
- Clustering: users' mobility patterns, type of infrastructures, activity patterns

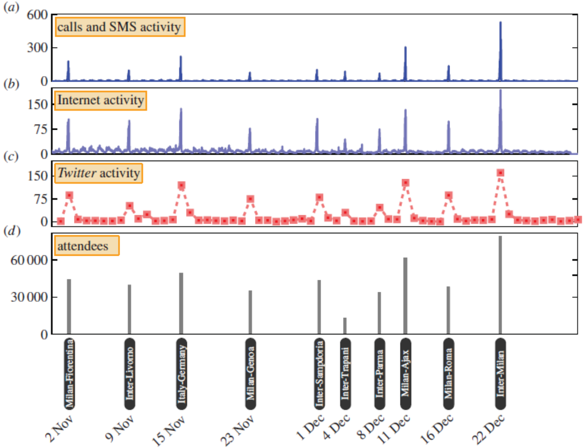
Methods:

- Linear regression (simple, multiple), more complex methods
- Markov chain, Hidden Markov Models, more complex models

Quantifying crowd size with mobile phone and Twitter data[1]

- Outline
- Research focuses
- St. of the Art
- Results
- Perspectives

- **Input:** Mobile phone and Twitter data
- **Goal:** Estimation of the number of people in a specific area at a given time



[1] Botta, Federico, Helen Susannah Moat, and Tobias Preis. *Royal Society Open Science* 2.5 (2015)

Quantifying crowd size with mobile phone and Twitter data

Outline

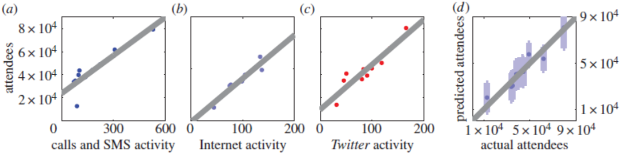
Research focuses

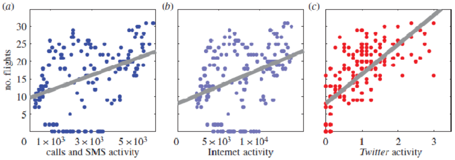
St. of the Art

Results

Perspectives

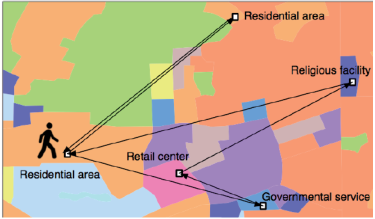
- Method: Ordinary least-squares regression
- Results: Linear relationship between the # of people attending the football matches and the volume of incoming and outgoing phone calls and SMS messages,
  Accurate estimation of the number of people in a given location and time





Relating land-use and human intra-city mobility[2]

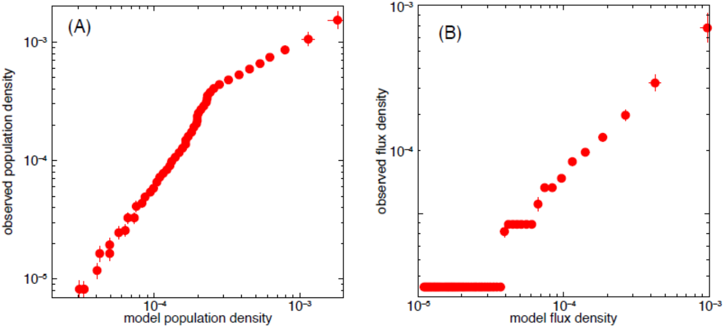
- Input: Land-use maps of the trip (origin-destination) + Purposes of people's trips + population density maps
- Goal: Prediction of mobility patterns
- Method: Markov process: Mobility model
  - Transition matrix between land-use types (capturing the trends)
  - Combining these observations with the gravity model to a model of human mobility



2 *Lee, Minjin, and Petter Holme. arXiv preprint arXiv:1505.07372 (2015)*

- Results: Predicting the trip lengths, the population density

(A) Average observed population density as a function of the population density predicted by the model
(B) Average observed flux density as a function of the flux density predicted by the model



Dynamic population mapping using mobile phone data[3]

- Goal: prediction of population density by MP & RS
- Method: Linear regression model with population-weighted least squares

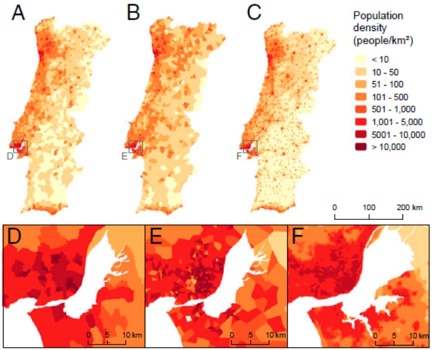


Figure: Comparison of predicted population density datasets with baseline data

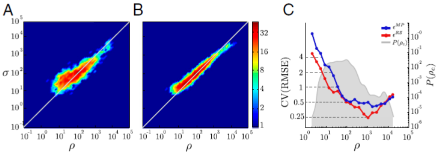
- MP relies on the density of towers (higher in urban areas)
- RS depends on geospatial datasets (cannot capture intraurban variations)

[3] *Deville, Pierre, et al., Proceedings of the National Academy of Sciences 111.45 (2014)*

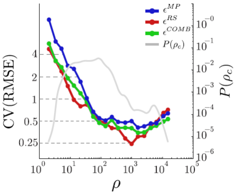
Dynamic population mapping using mobile phone data

- **MP** lower precision, especially in low-density areas
- **RS** higher precision but less accurate predictions (overestimation of ρ in $a_{low-density}$, underestimation of ρ in $a_{high-density}$)
- Normalized RMSE of both methods decreases with population density
- General increase of r values with population density
- RMSE (MP) $>$ RMSE (RS) (except in high-density areas)

Comparing the performance of the MP & RS (Precision and accuracy assessments)



RMSEs normalized by the average population density of intervals for MP, RS & COMB



A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data[4]

Signature: representation of the typical mobile traffic observed at a unit area
Signatures clustering ≈ subscribers' activities clustering

- Goals: Analyzing the different steps common to all urban fabric detection techniques:

	Soto	Cici	MWS
(1)Mobile traffic signature	$S_a(wd,t) \| S_a(we,t)$ Normalization (standard score)	$S_a(\mathbf{d},t)$,Filtering (FFT, IFFT), Seasonal Communication Series (SCS)	$s_a(mon,t) = \mu_{1/2}(\{v_a(d,t) \mid d \in \mathbf{d}\})$, standard score, daily normalization
(2) Distance between signatures	Euclidian ($d_{ab} = \hat{s}_a, \hat{s}_b$)	Pearson (d_{ab}= 1-p_{ab})	Pearson, Euclidian
(3) Clustering of signatures	k-means (k=5, 10)	linkage(a whole family of solutions)	linkage

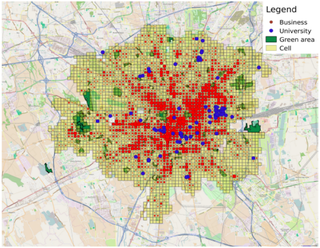
- Median Week Signature (MWS) is based on:
  - Mobile traffic activity within a one-week period $d^{mon} \bigcup ... d^{sun} = \mathbf{d}$
  - Median is more robust to outliers (not for average and absolute values)

Metrics: Density, coverage, Entropy, F-score → MWS-stdscr-pearsons

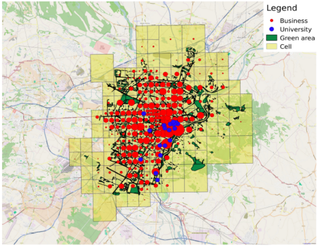
[4] *A. Furno, R. Stanica, M. Fiore, Proceedings of IEEE/ACM ASONAM 2015*

A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data

Ground-truth information

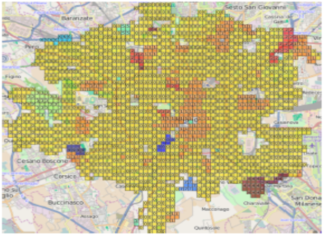


(a) Milan-2013

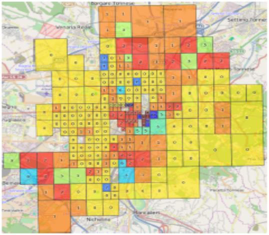


(b) Turin-2015

10 largest clusters identified identified through MWS-stdscr-pearson



(a) Milan-2013



(b) Turin-2015

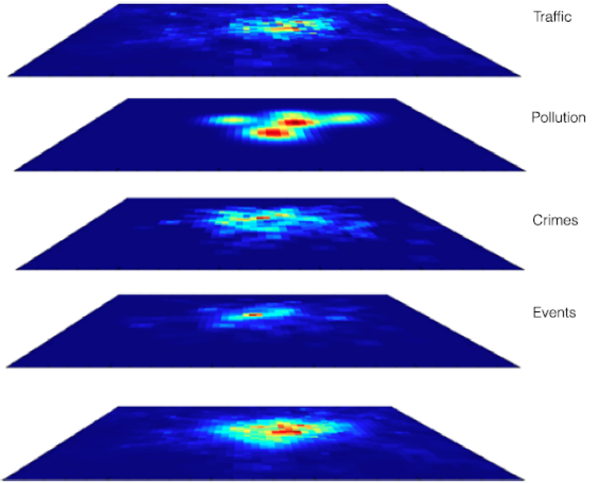
Personalized routing for multitudes in smart cities[5]

Capturing human mobility → understanding underlying patterns, designing intelligent systems

- **Input:** Mobile phone data
- **Goal:** An adaptive routing strategy considering constraints to recommend personalized routes
  - Constraints: Pollution layer, Event layer, Crime layer, Traffic layer
- **Method:** Defining a potential energy landscapes:
  - Constructing total constraint matrix by defining linear combination of constraints
  - Constructing a pairs of origin-destination probability matrix

[5] *De Domenico, Manlio, Antonio Lima, Marta C Gonzalez and Alex Arenas. EPJ Data Science 4, no. 1 (2015)*

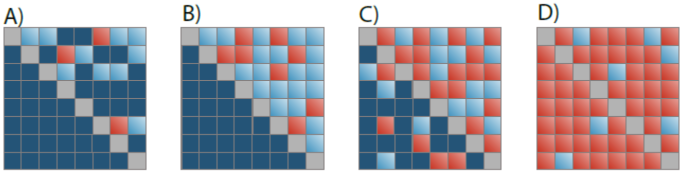
Personalized routing for multitudes in smart cities



Integration of diverse layers of constraints → Recommendations

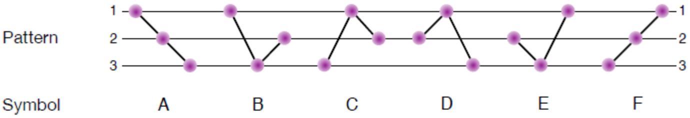
Symbolic Transfer Entropy[6]

- Input: Micro-blogging data (key words)
- Golas: Mapping of influence networks (extracting networks of causal influence among different geographical sub-units before, during, and after collective social phenomena)
- Method: Symbolic transfer entropy quantifies the directional flow of information between two time series X and Y:
  - categorizing the signals in a small set of symbols or alphabet
  - computing from the relative frequency of symbols in each sequence $\hat{X}$ and $\hat{Y}$ the joint and conditional probabilities of the sequences indices (Shannon's entropy)



[6] *Staniek, Matthus, and Klaus Lehnertz, Physical Review Letters 100.15 (2008)*

Symbolic Transfer Entropy

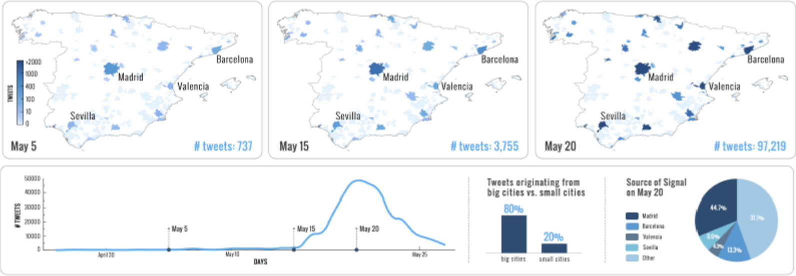


Sample: 120 74 203 167 92 148 174 47

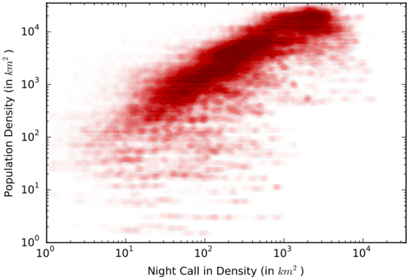
m-subsamples	120 74 203	74 203 167	203 167 92	167 92 148	92 148 174	148 174 47
indices	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
sorted *m*-subsamples	74 120 203	74 167 203	92 167 203	92 148 167	92 148 174	47 148 174
sorted indices	2 1 3	1 3 2	3 2 1	2 3 1	1 2 3	3 1 2
Symbolized sample	D	B	F	E	A	C

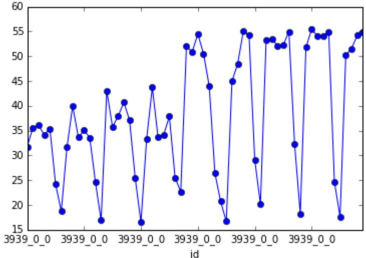
Symbolic Transfer Entropy

- Results:

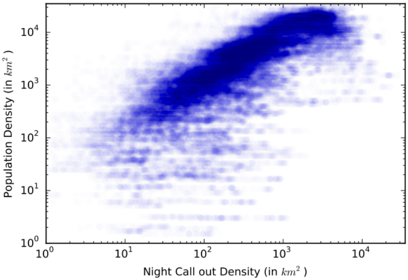


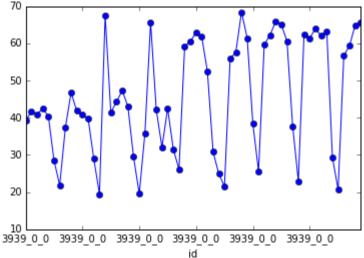
Results - Population Density & Number of Call-in Per Night



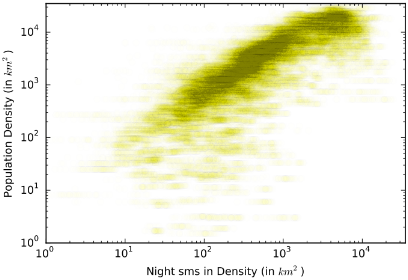


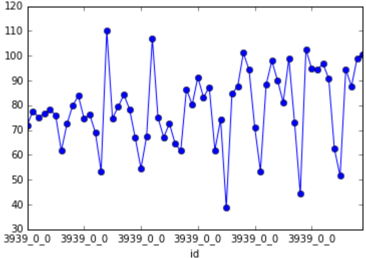
Results - Population Density & Number of Call-out Per Night



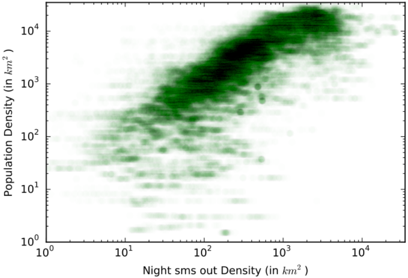


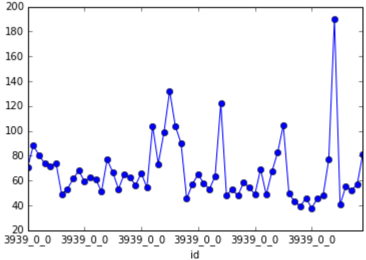
Results - Population Density & Number of SMS-in Per Night



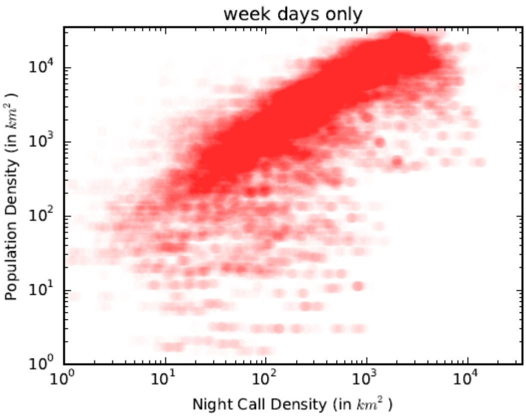


Results - Population Density & Number of SMS-out Per Night





Results - Population Density & Number of Call-in for Week Days



Perspectives

- Week-days, Week-end for whole dataset
- Applying a regression model fits better data distribution
- Mix of methods (*e.g.* Transfer Entropy + layers of constraints)
- ...