Mobility Prediction Advanced Techniques

Mirco Musolesi

Department of Geography, UCL

m.musolesi@ucl.ac.uk

Outline

- Introduction to prediction of human mobility
- Some examples:
  - Basic models
  - Linear time series analysis
  - Non linear time series analysis
  - Nextplace
  - Using mutual information for improving prediction of human mobility

Prediction of Human Mobility

- Hard problem but many potential applications:
  - Personalised advertising
  - Urban planning
  - Intelligent uploading of information
  - Intelligent handover in cellular networks
- Strong interest in the recent years in computer science but also in other communities such as complex systems, etc.

Different Types of Prediction

- Location can be:
  - Logical location (workplace, home)
  - Geographic location
    - Discrete areas (e.g., square in a grid)
    - GPS locations
- Not only *where* but also *when*
  - Spatio-temporal prediction is hard

Markovian Model

- Discretisation of the geographic space
- A matrix is used to describe *transitions* from a certain area of space to another one
- Memoryless random process
- Large body of theoretical results about these models

Spatial Transition Matrix

- You can build a Markov model of transitions between different geographic areas
- Then a transition matrix of activities can be associated to each location
- Problem: it is very hard to get enough statistics
- Solution: you can associate the most popular activity to each location (and use only the transition matrix of the movement)

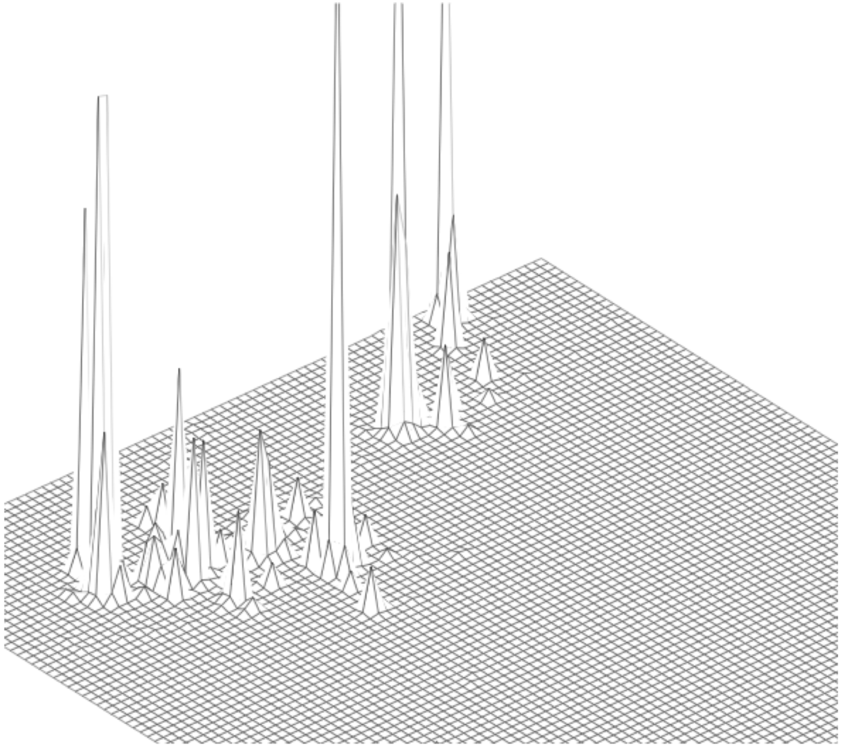
Predictability of Human Behavior

- Any prediction of future behaviour is based on the assumption of determinism.
- The intuition behind NextPlace is that a sequence of important locations that an individual visits every day is more or less fixed, with only minor variations.
- We focus on the analysis of **sequences of significant places**.

Extraction of Significant Places
GPS Logs

Methodology:

1. 2D Gaussian distribution at each GPS point.
2. Points over a threshold are selected.
3. Bordering areas are merged.
4. Consecutive short visits are merged together.



Extraction of Significant Places
WiFi Logs

Methodology

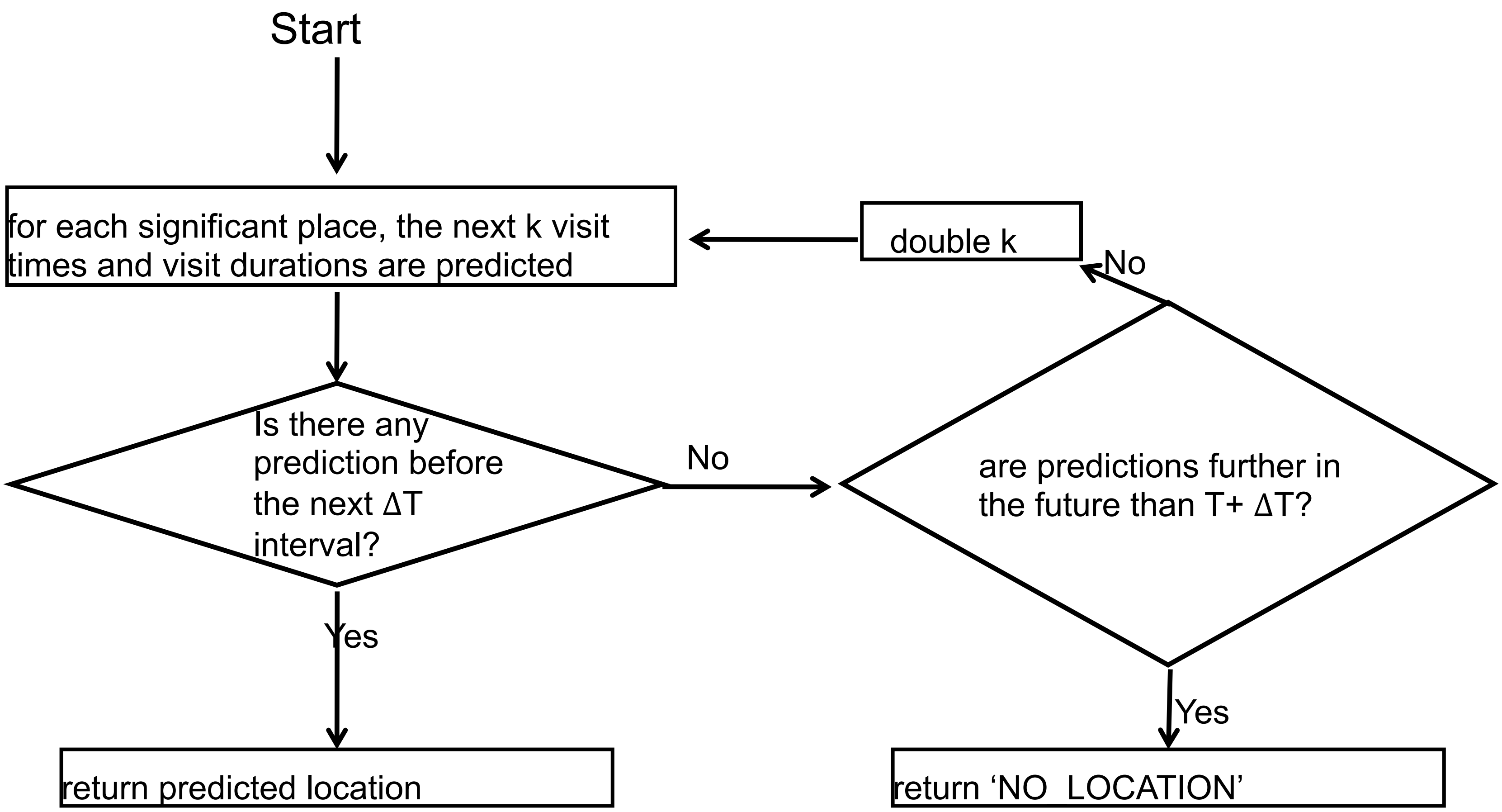
1. Most frequently seen Access Points are chosen as significant places.
2. Consecutive short association periods are merged together.

Predicting Future Visits to a Place

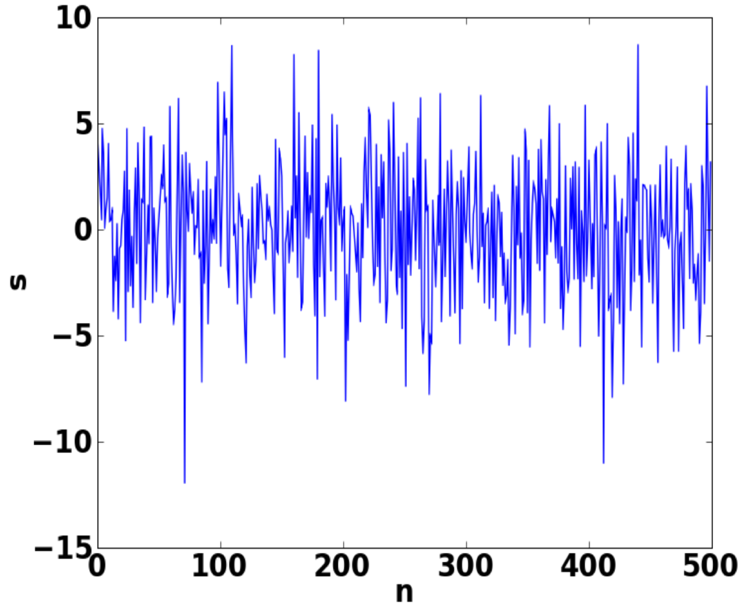
Given **a list of visits** $(t_1,d_1), (t_2,d_2),\ldots,(t_n,d_n)$ to a given significant place:

- Two time series $C = (c_1,c_2,\ldots c_n)$ with daily start times (24h format) and $D = (d_1,d_2,\ldots,d_n)$ with visit durations are created.
- We search in ***C*** **subsequences** similar to last m values.
- Estimate **next start time** c_{n+1} averaging all values which follow each found subsequence.
- Estimate **next visit duration** d_{n+1} in the same way but with the corresponding subsequences found in ***C.***

Predicting the Next Location



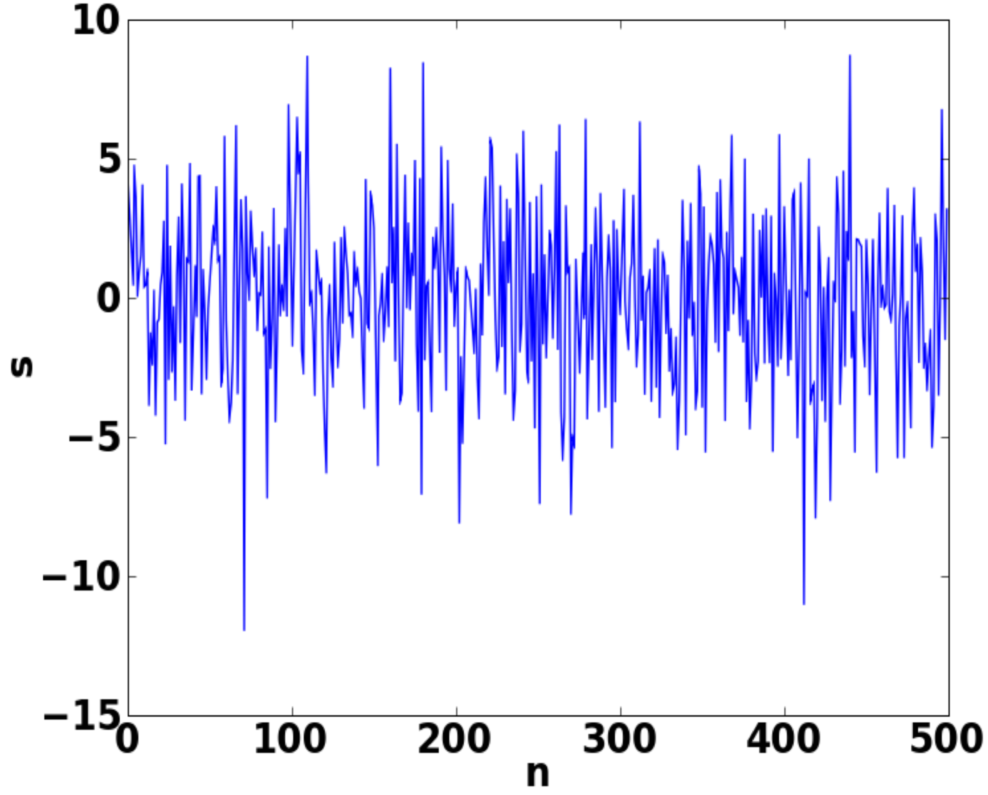
Nonlinear Prediction of Time Series



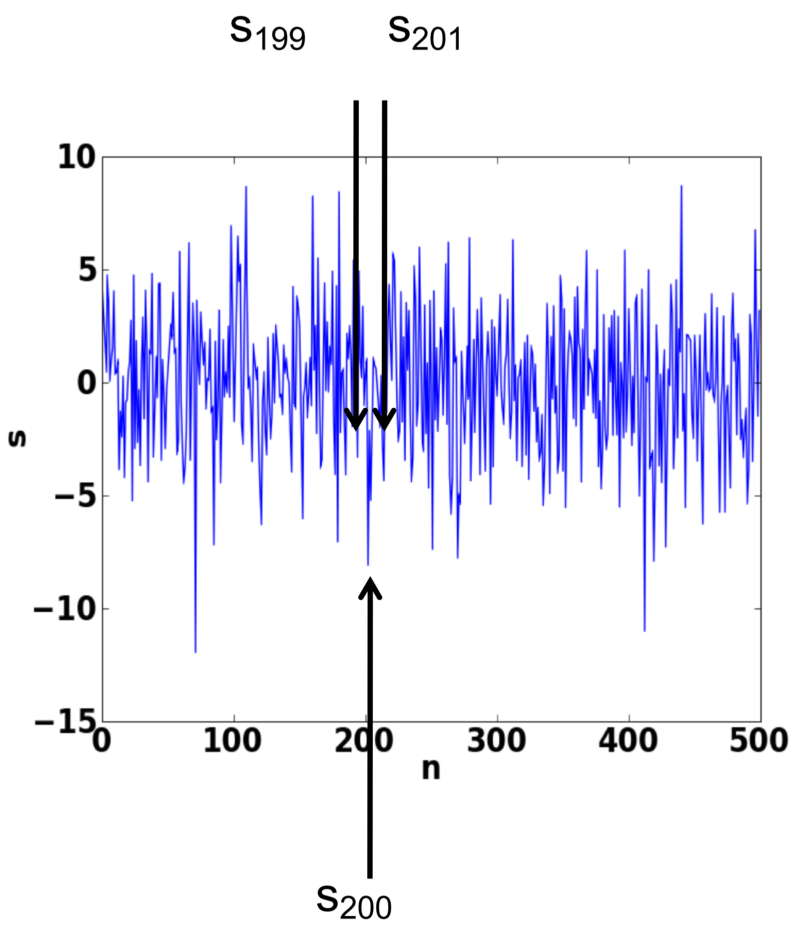
- Time series apparently random, but it is possible to uncover the characteristics of its dynamic evolution by analyzing subsequences of the time series itself.
- A time series **(s_0, s_1, s_2, s_3, … s_N)** can be embedded in a *m*-dimensional space by creating a delay vector reconstruction:

$$\beta_n = [s_{n-(m-1)}, s_{n-(m-2)}, \ldots, s_{n-1}, s_n]$$

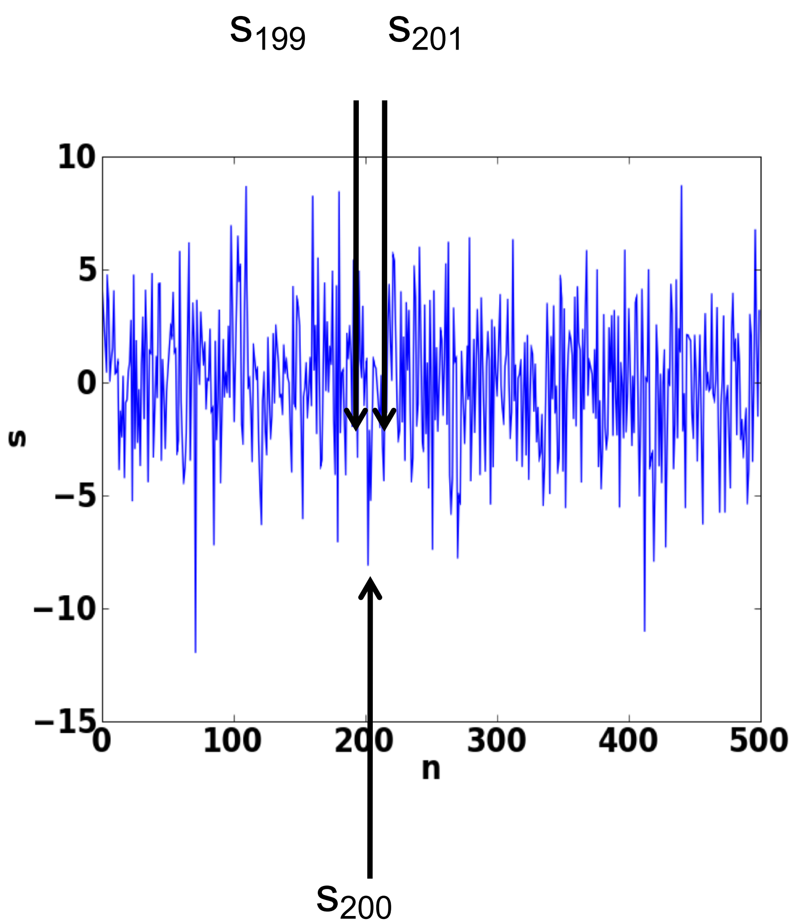
Nonlinear Prediction of Time Series

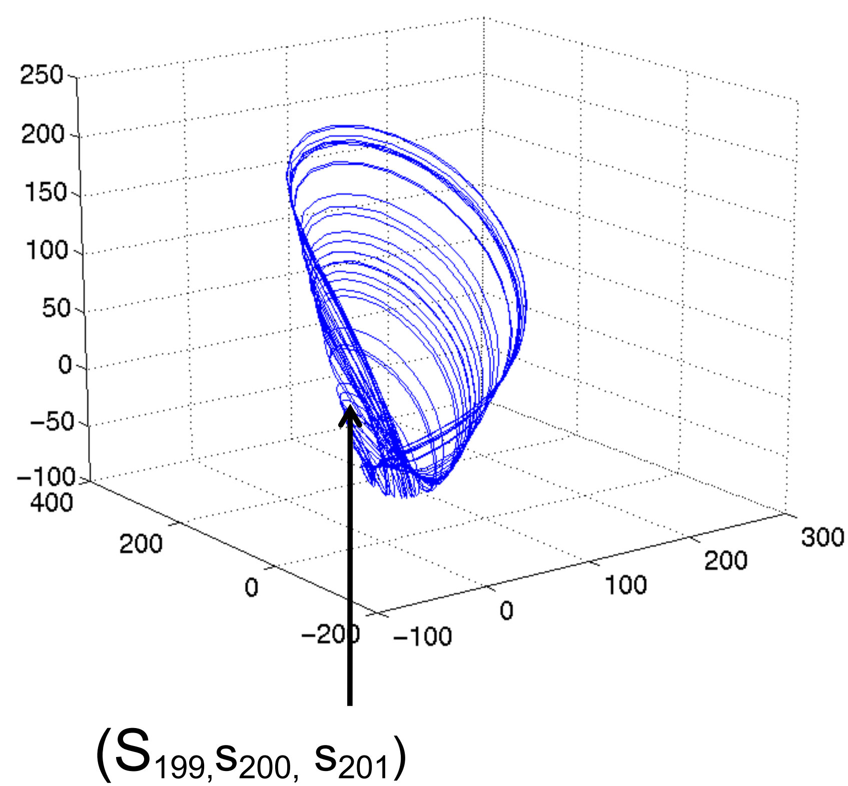


Nonlinear Prediction of Time Series



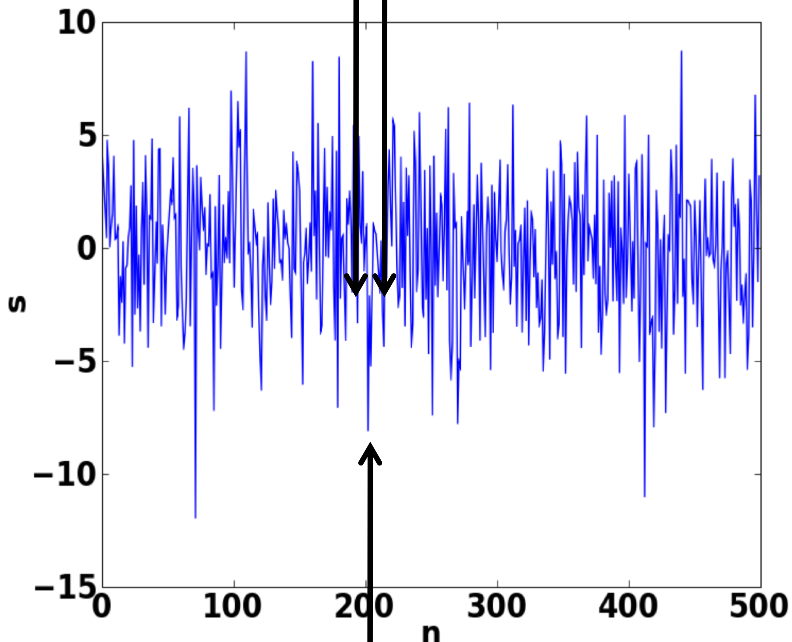
Nonlinear Prediction of Time Series





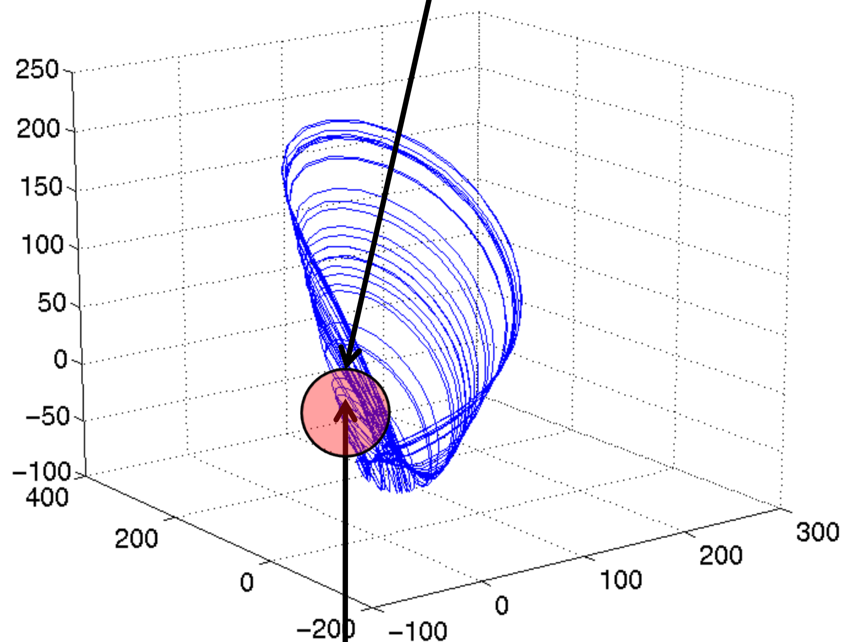
Nonlinear Prediction of Time Series

s_{199} s_{201}



s_{200}

We take the average of the values inside the neighborhood of $s_{n+\Delta n}$



$(S_{199}, s_{200}, s_{201})$

m=3

Datasets

Dataset	Users	Visits	Places	Average visit duration	Trace length
Cabspotting	252	150,612	6,122	231 s	23 days
CenceMe GPS	19	3,832	225	696 s	12 days
Dartmouth WiFi	2,043	772,217	539	2,094 s	60 days
Ile Sans Fils	804	142,407	173	5,296 s	370 days

Predictability Error

Real value

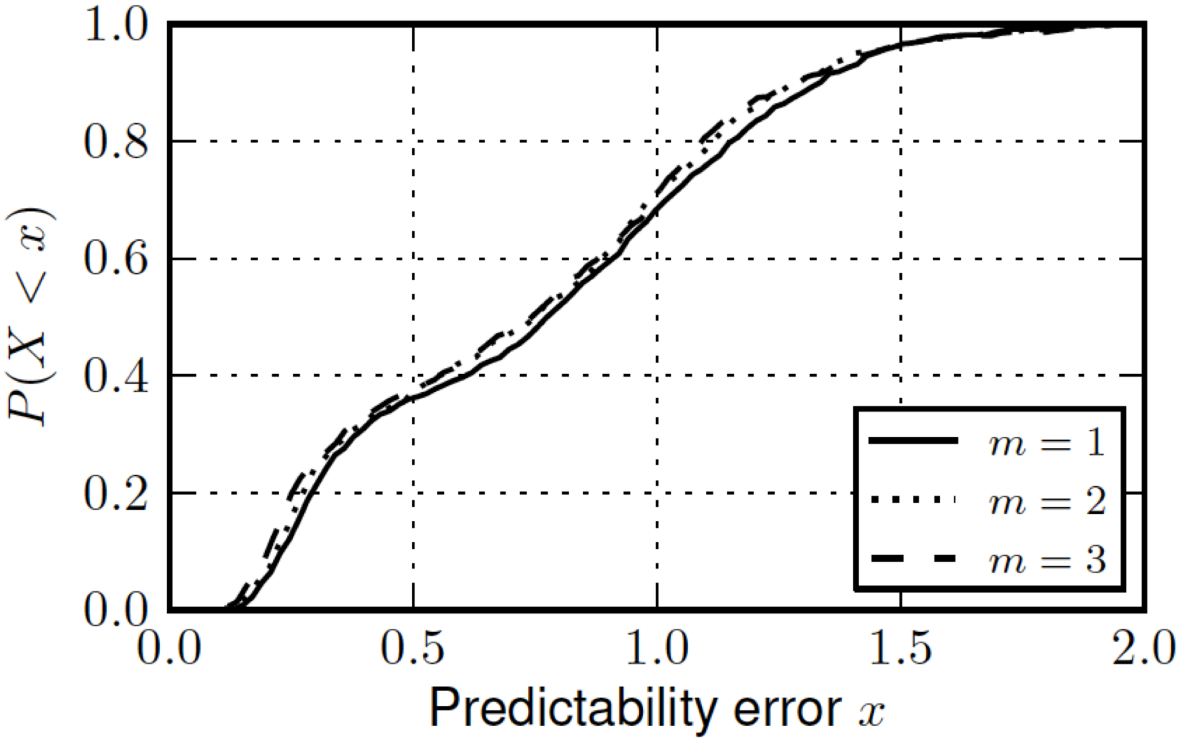
Predicted value

$$e = \frac{\sum_{i=1}^{N} (s_i - p_i)^2}{N}$$

Number of samples of the time series

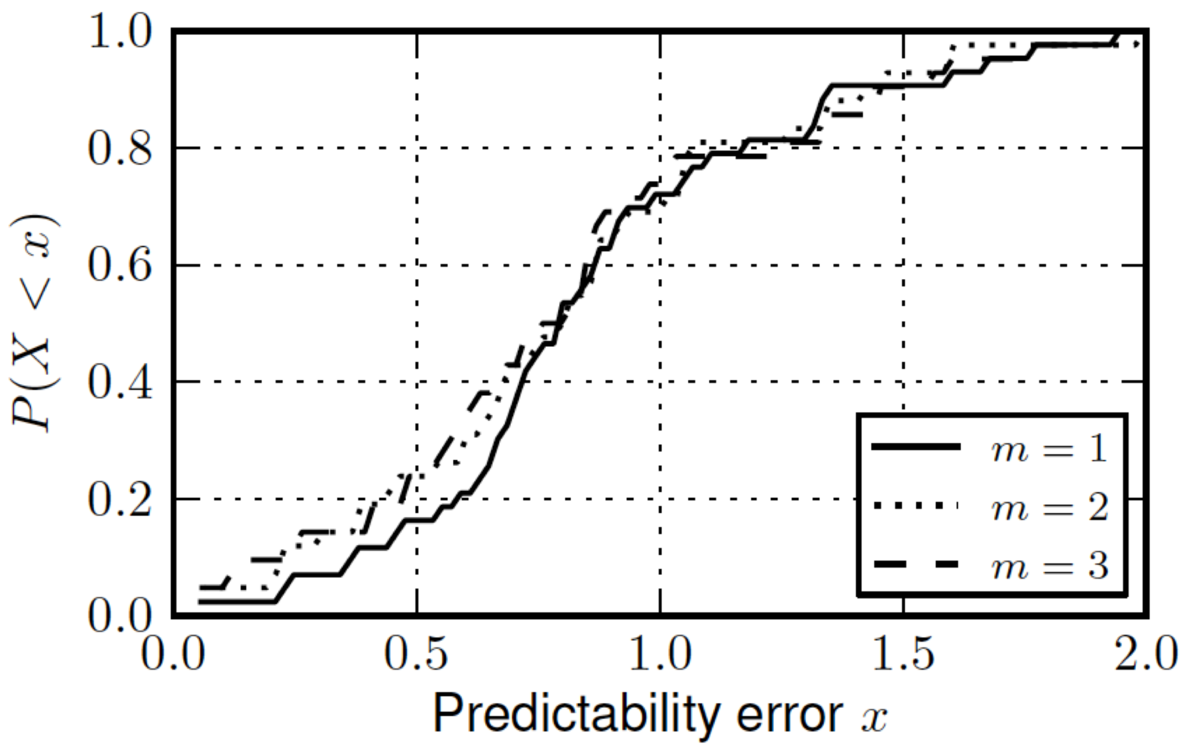
$$\text{Predictability error } x = \frac{e}{s^2}$$

Predictability Analysis - Cabspotting



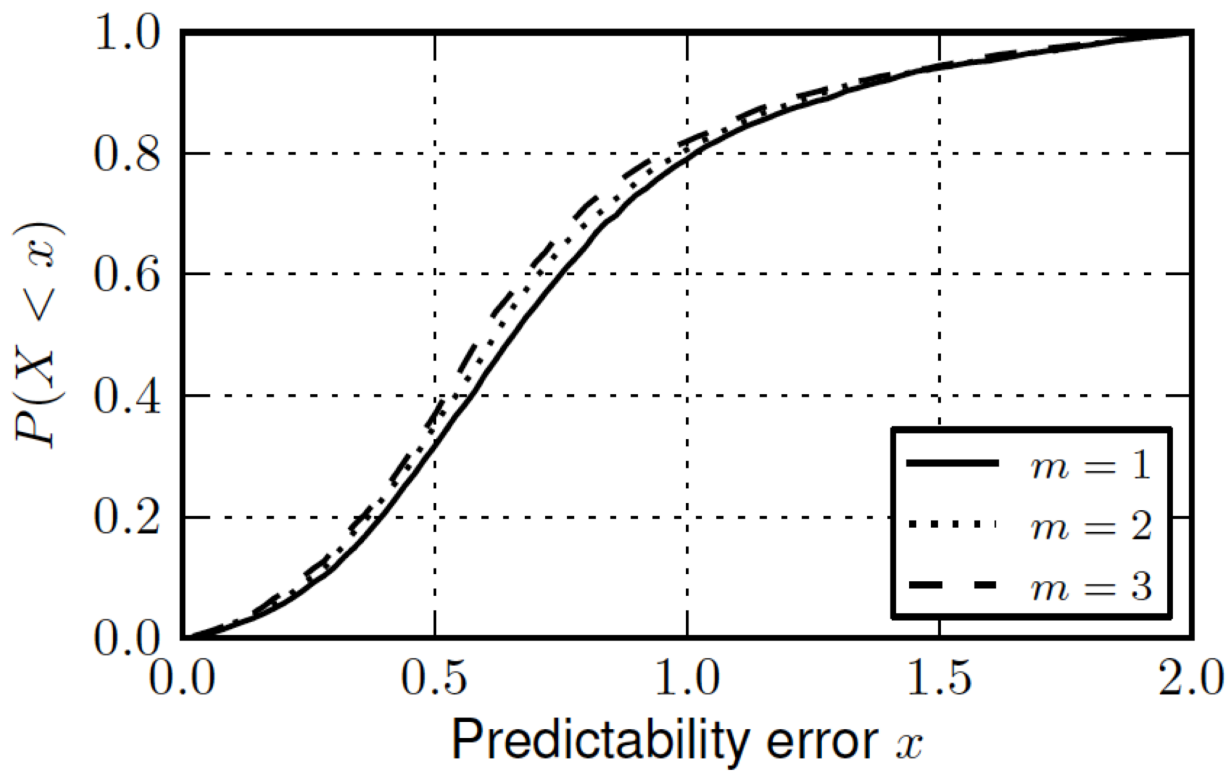
Cumulative Distribution of the **predictability error** on the start time series for nonlinear models with different embedding values ***m***.

Predictability Analysis - CenceMe GPS



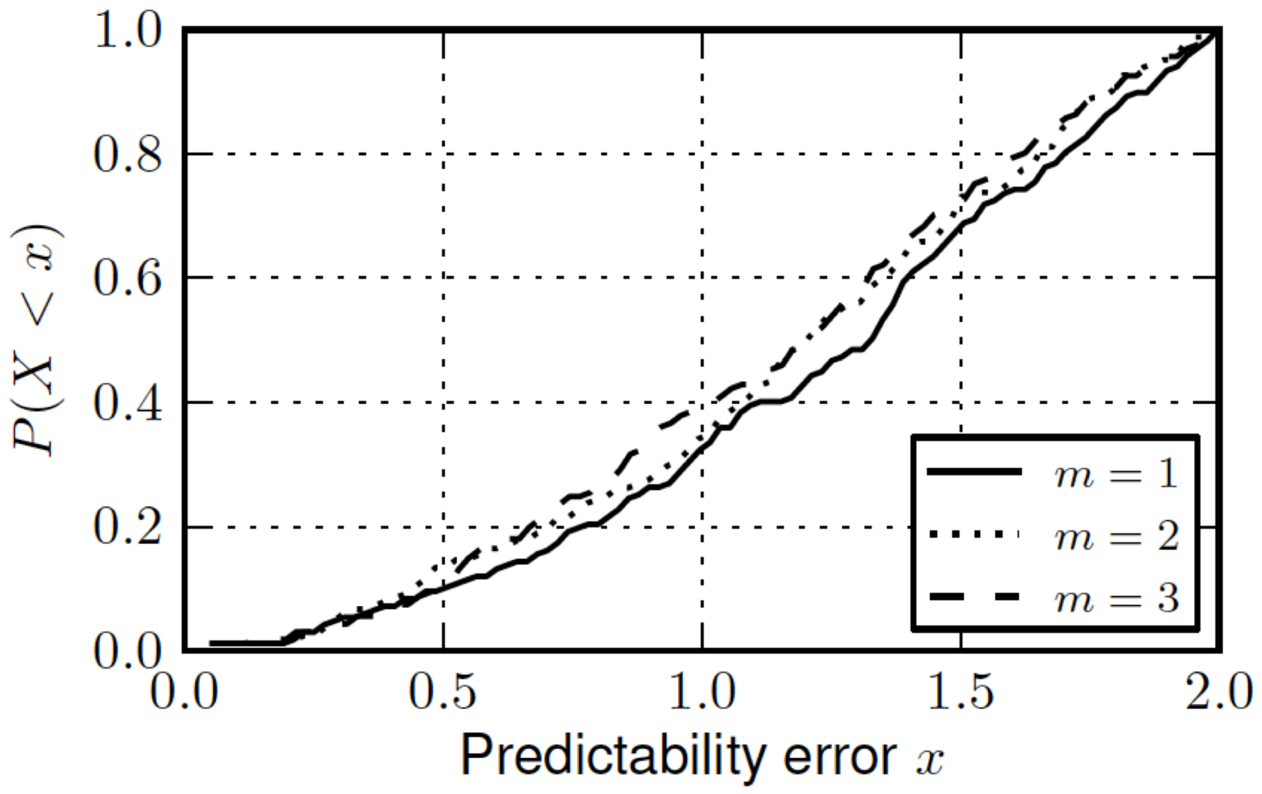
Cumulative Distribution of the **predictability error** on the start time series for nonlinear models with different embedding values ***m***.

Predictability Analysis - Dartmouth WiFi



Cumulative Distribution of the **predictability error** on the start time series for nonlinear models with different embedding values ***m***.

Predictability Analysis - Ile Sans Fils



Cumulative Distribution of the **predictability error** on the start time series for nonlinear models with different embedding values ***m***.

Prediction Performance: Comparative Evaluation

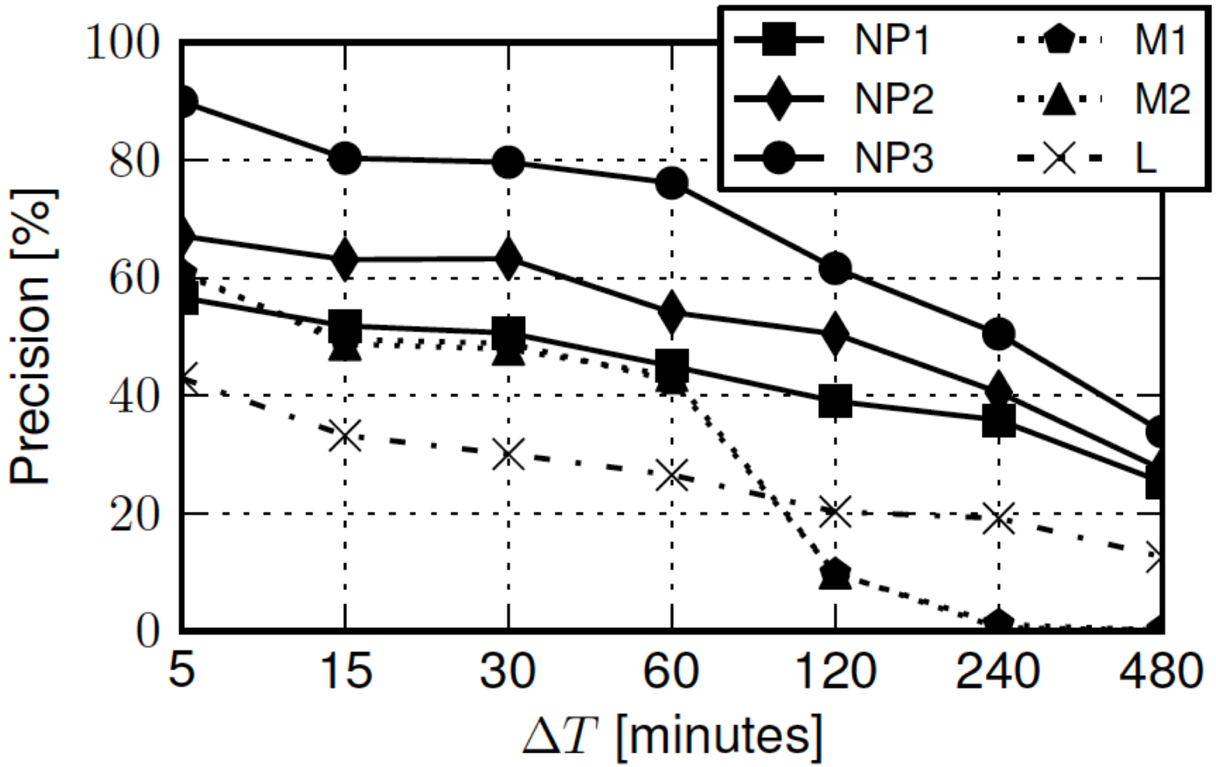
Comparative evaluation of prediction performance on different prediction models:

- **NextPlace** with **nonlinear** time series predictor (m=1,2,3)
- Spatio-temporal **Markov-based predictor** (1^{st} and 2^{nd} order)
- **NextPlace** with **linear** time series predictor (order-4 running average)

Experimental Evaluation: Metrics

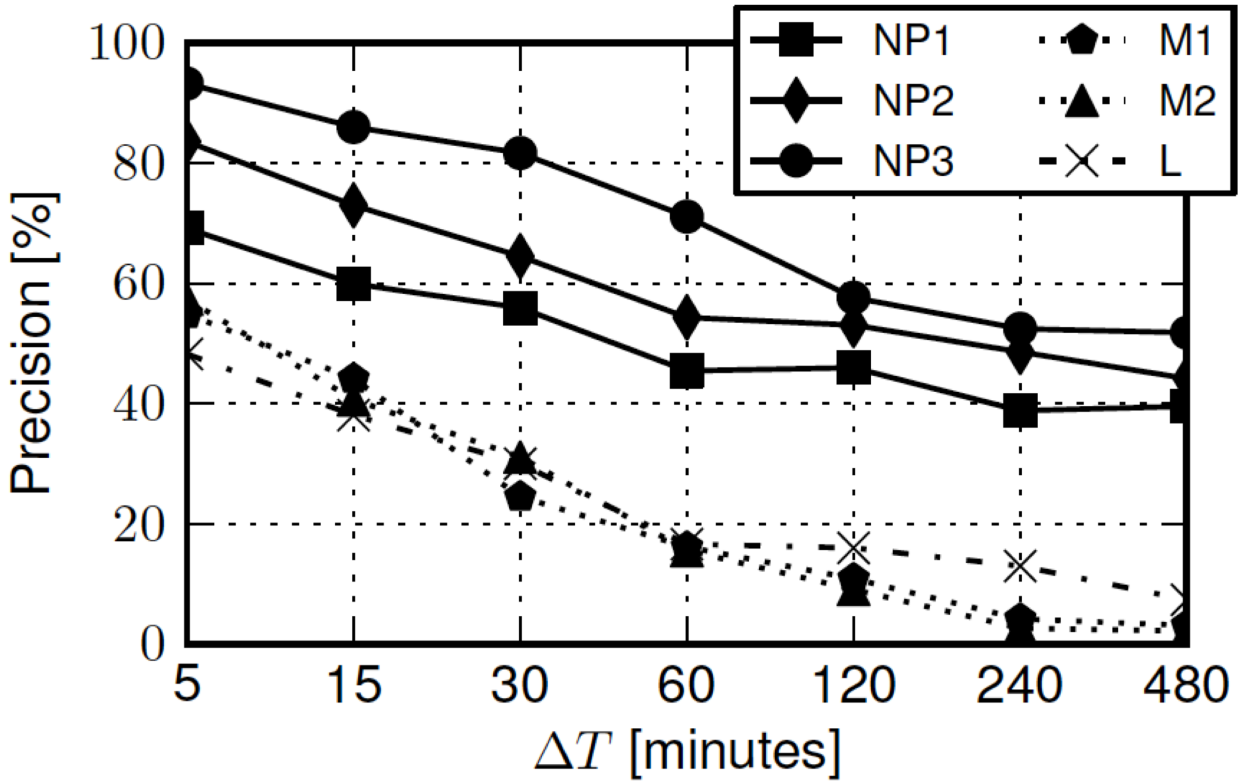
- **Correctness**: if we predict, at time $\boldsymbol{T}$, that the user $\boldsymbol{i}$ will be at location $\boldsymbol{p}$ at time $\boldsymbol{T_P} = \boldsymbol{T} + \Delta\boldsymbol{T}$, the prediction is considered correct only if the user is at $\boldsymbol{p}$ at any time during the interval $\boldsymbol{[T_P - \theta, T_P + \theta]}$, where $\boldsymbol{\theta}$ is the error margin.
- **Precision**: the ratio between the number of correct predictions and the number of all attempted predictions which forecast the user to be in a significant location. We do **not** consider prediction in NO_LOCATION.

Prediction Precision - Cabspotting



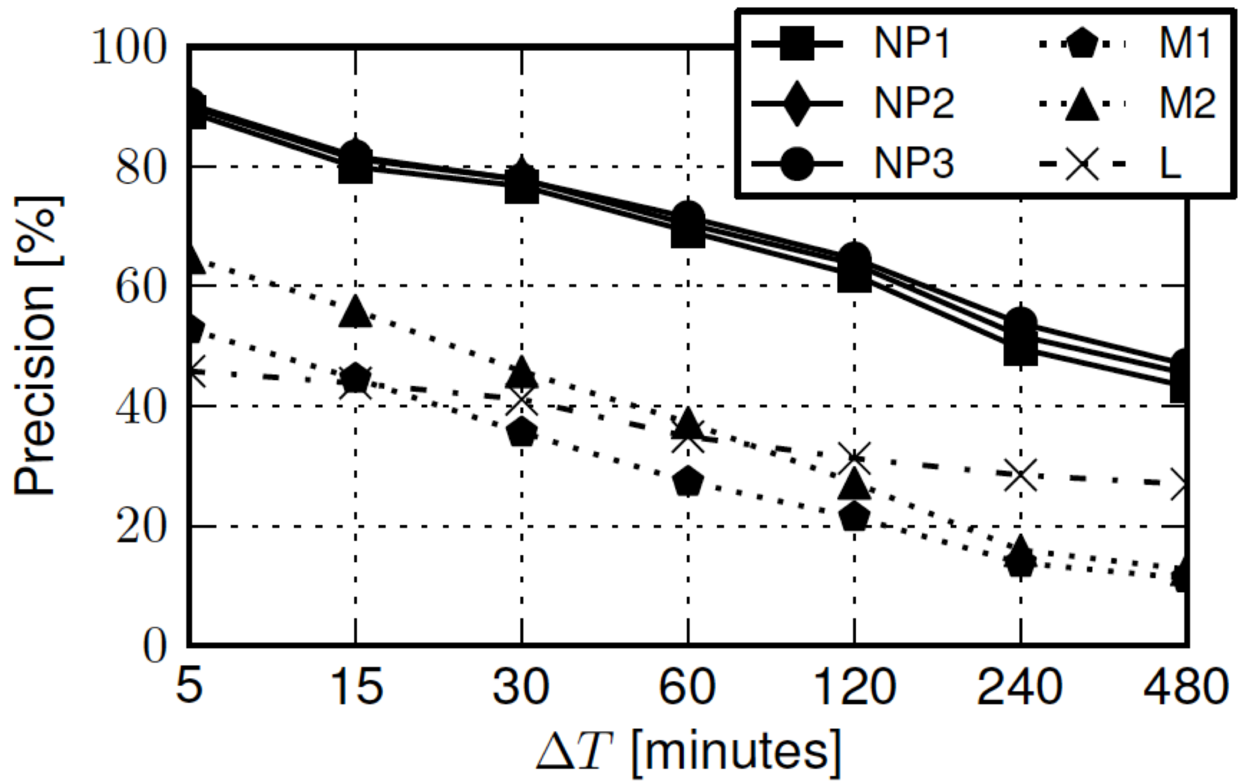
Prediction precision as a function of time interval ΔT for the different datasets and for different predictors. Error margin is θ = 900 seconds.

Prediction Precision - CenceMe GPS



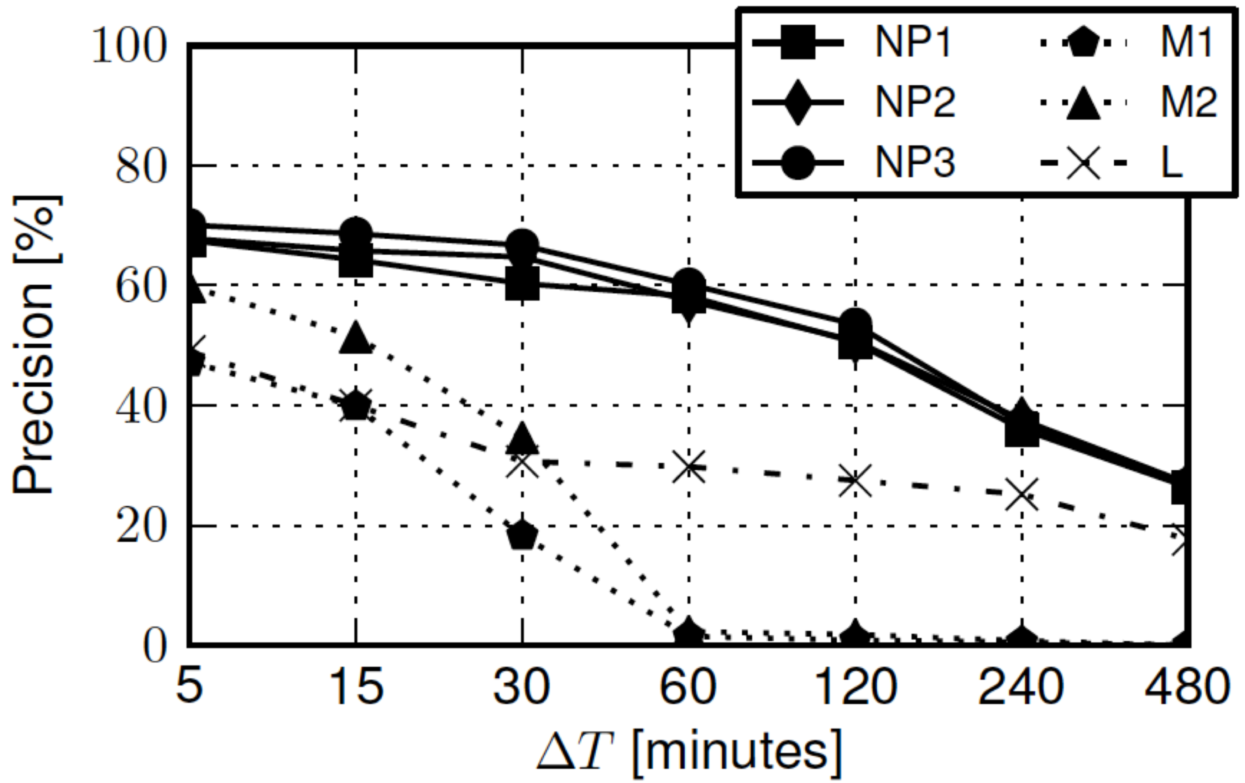
Prediction precision as a function of time interval ΔT for the different datasets and for different predictors. Error margin is θ = 900 seconds.

Prediction Precision - Dartmouth WiFi



Prediction precision as a function of time interval ΔT for the different datasets and for different predictors. Error margin is θ = 900 seconds.

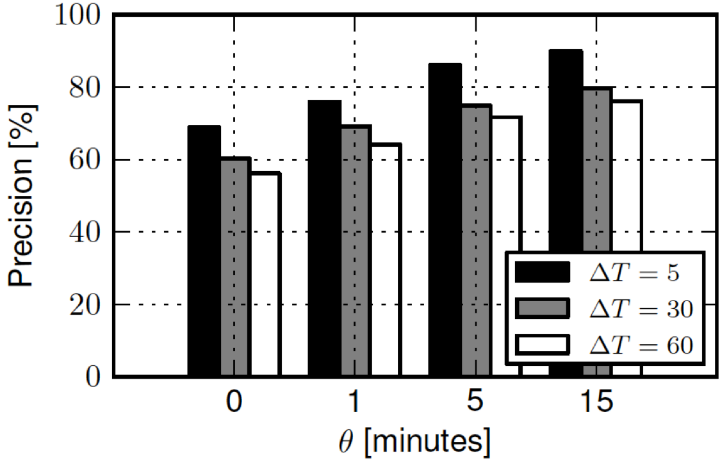
Prediction Precision - Ile Sans Fils



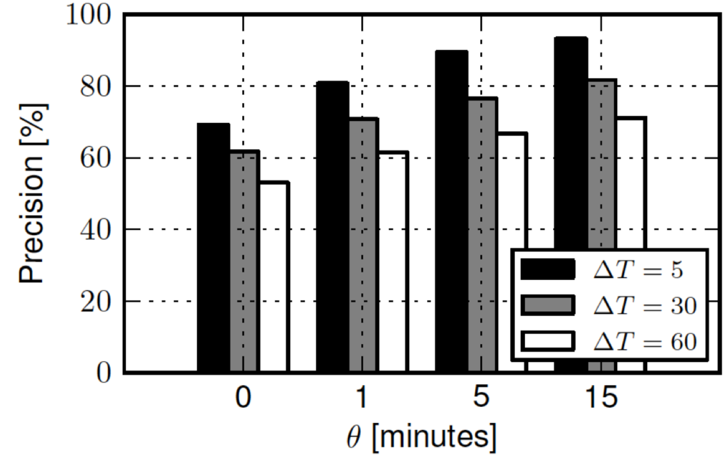
Prediction precision as a function of time interval ΔT for the different datasets and for different predictors. Error margin is θ = 900 seconds.

Prediction Precision – Error Margin

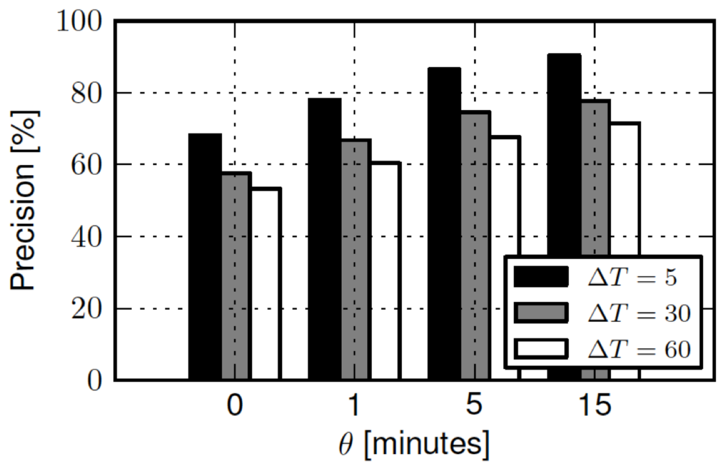
Cabspotting



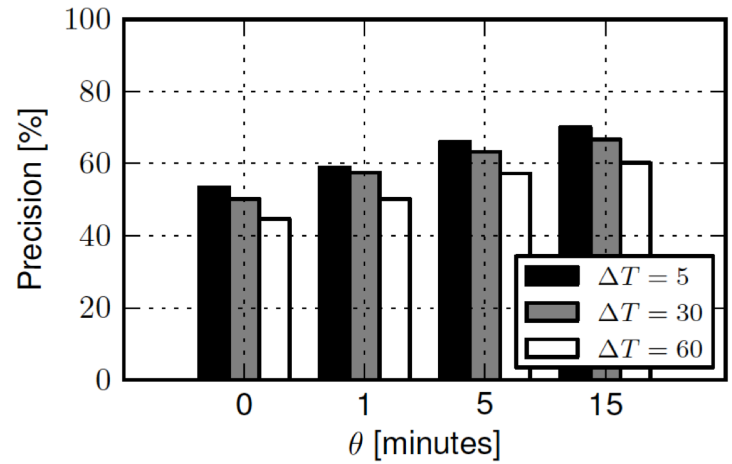
CenceMe GPS



Dartmouth WiFi



Ile Sans Fils



Reference

Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora and Andrew T. Campbell. NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems. In Proceedings of the 9th International Conference on Pervasive Computing (Pervasive 2011). San Francisco. California, CA. June 2011.

Exploiting Movement Correlation

- Is it possible to improve the accuracy of the prediction by considering traces of multiple users?
- If yes, who should we select for improving the prediction of the movements of a given user?
- Can mobility correlation be considered as a cue for inferring social ties?

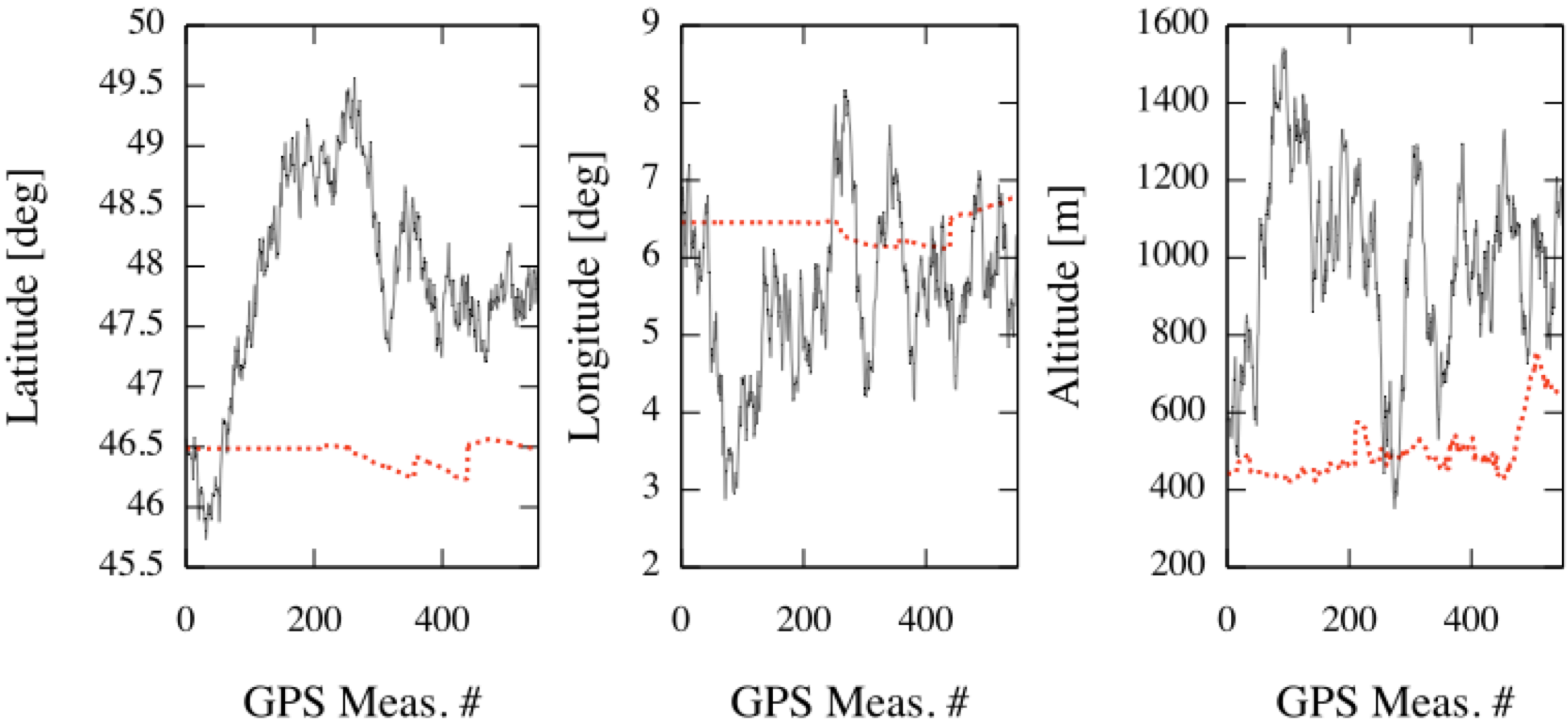

flickr: cubagallery

The Nokia MDC Dataset

- The complete dataset contains information from 152 smartphones (Nokia N95) for a year: address book, GPS, WLAN and Bluetooth traces, calls and SMS logs.
- We received data from 39 devices, 14 phone numbers were missing. We analysed a subset of the data related to 25 devices.

flickr: stevegarfield

Linear Predictor

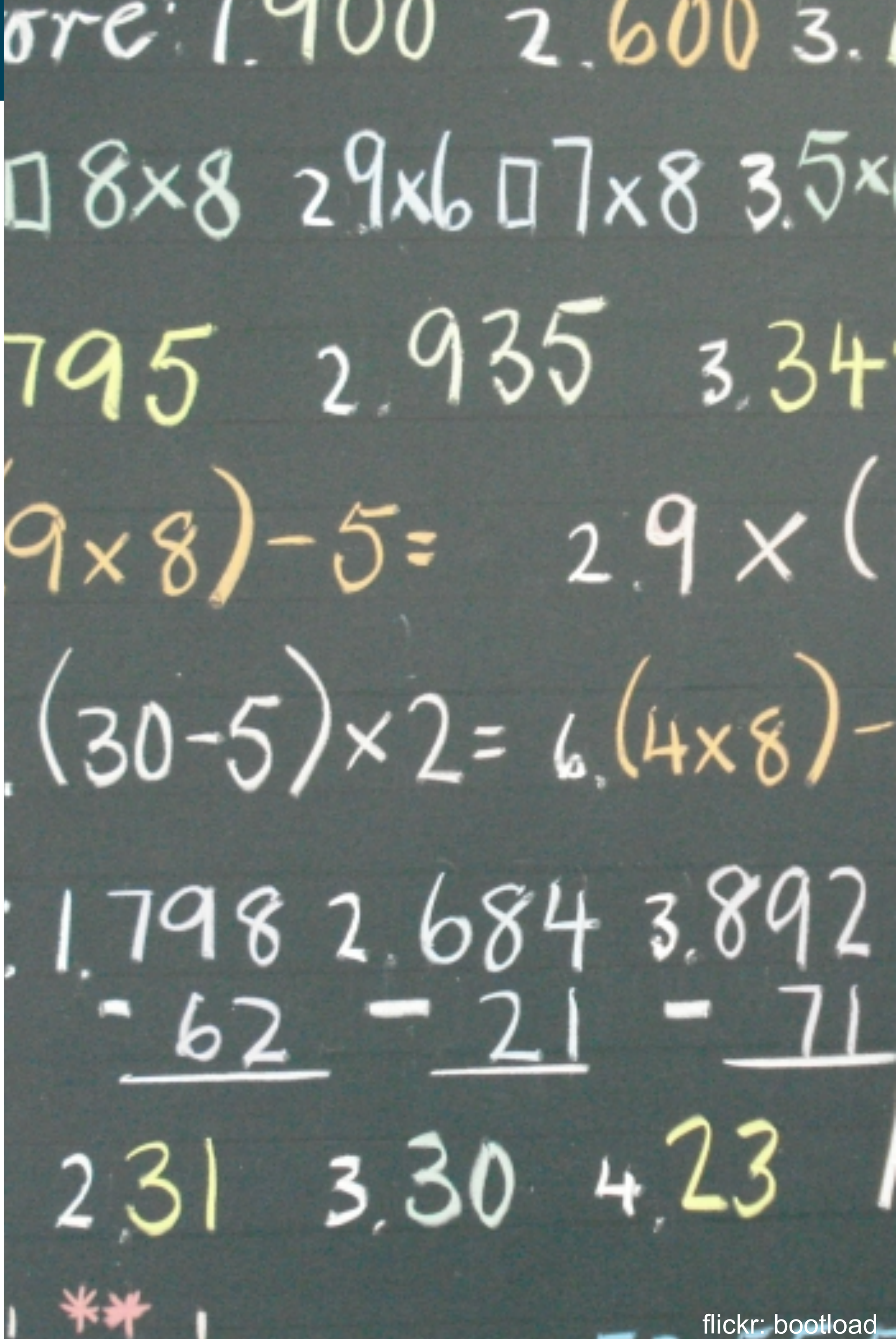


600 GPS (~60 hours) measurements (red) against forecast (black) for user 129

Global error (stdev) is of the order of 3 deg for lat-lng and 600 m for altitude.

Multivariate nonlinear time series prediction

- Extension to the case of multiple users.
- In particular pairs of users:
  - connected by a social link; and/or
  - with correlated mobility patterns.



flickr: bootload

The Mobility Model

User movements

$$\dot{\mathbf{x}}(t) = f(x,t) + \eta(t)$$

Noise

Hour of the day

Longitude

$$\mathbf{x_n} = (h_n, \phi_n, \lambda_n, \xi_n)$$

Latitude

Altitude

The Mobility Model

$$\dot{\mathbf{x}}(t) = f(x,t) + \eta(t) \qquad \mathbf{x_n} = (h_n, \phi_n, \lambda_n, \xi_n)$$

- We consider the mobility model as a nonlinear dynamical system of a deterministic signal with a stochastic noise.
- The position of a user is modeled with a 4-dim state vector.
- We cannot analyse the d-dimensional phase space of this chaotic system directly.

Takens' Embedding Theorem

$$\mathbf{x}_n \equiv \left(x_{n-(m-1)\tau}, \ldots, x_{n-\tau}, x_n\right)$$

$$\begin{aligned}\mathbf{v}_n \equiv \big(& y_{1,n-(m_1-1)\tau_1}, \cdots, y_{1,n}, \\ & y_{2,n-(m_2-1)\tau_2}, \cdots, y_{2,n}, \\ & y_{M,n-(m_M-1)\tau_M}, \cdots, y_{M,n}\big)\end{aligned}$$

- We can construct a space which conserves the dynamic properties of the system by using delayed measurements of the time-series.
- The theorem holds for time series of infinite length and without noise. We need a **multivariate** analysis to have a good precision on real-world time-limited noisy data.

Takens' Embedding Theorem

Embedding dimension = 8

$$\mathbf{x}_n \equiv (x_{n-(m-1)\tau}, \ldots, x_{n-\tau}, x_n)$$

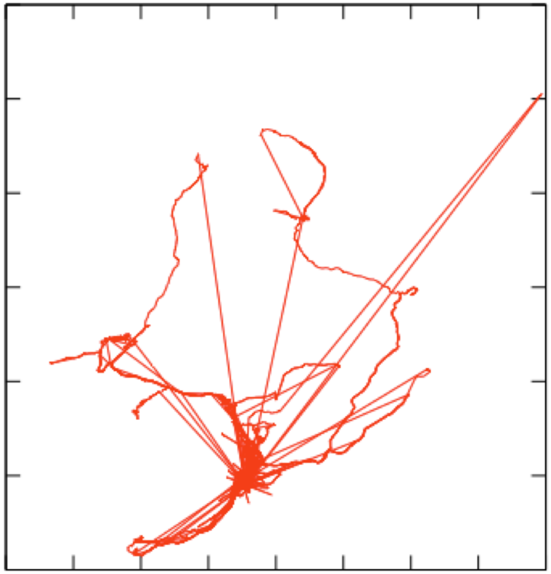
Delay time ~= 1000

We can construct a space which conserves the dynamic properties of the system by using delayed measurements of the time-series.

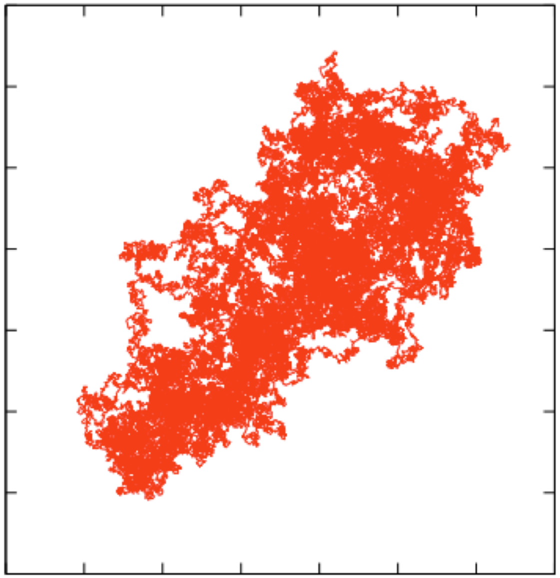
$$\begin{aligned}\mathbf{v}_n \equiv (&y_{1,n-(m_1-1)\tau_1}, \ldots, y_{1,n}, \\ &y_{2,n-(m_2-1)\tau_2}, \ldots, y_{2,n}, \\ &\ldots \\ &y_{M,n-(m_M-1)\tau_M}, \ldots, y_{M,n})\end{aligned}$$

The theorem holds for time series of infinite length and without noise. We need a **multivariate** analysis to have a good precision on real-world time-limited noisy data.

Delay Embedding Reconstructions



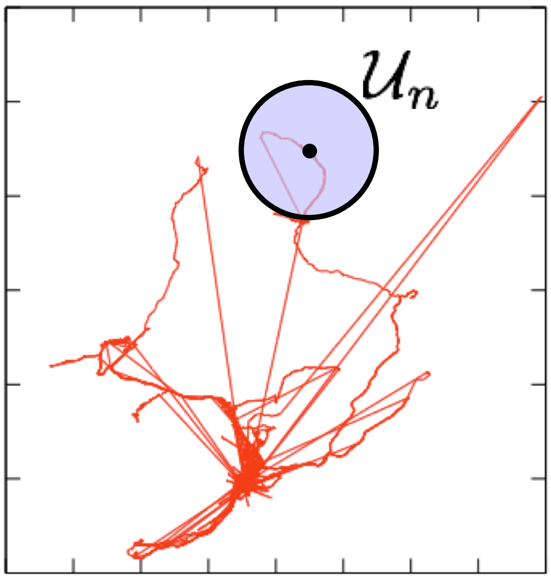
Reconstruction for user 129



Reconstruction for a Brownian motion

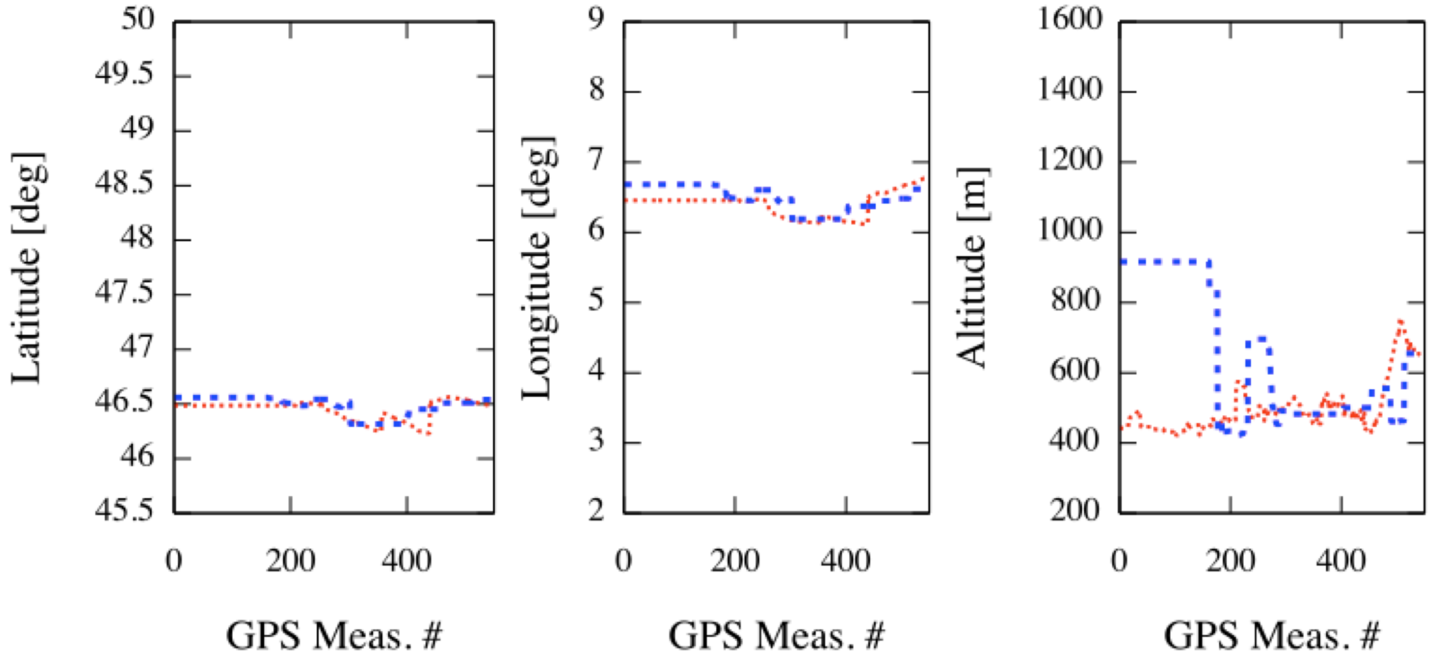
Multivariate Nonlinear Prediction

$$\hat{\mathbf{v}}_{n+k} = \frac{1}{|\mathcal{U}_n|} \sum_{\mathbf{v}_j \in \mathcal{U}_n} \mathbf{v}_{j+k}$$



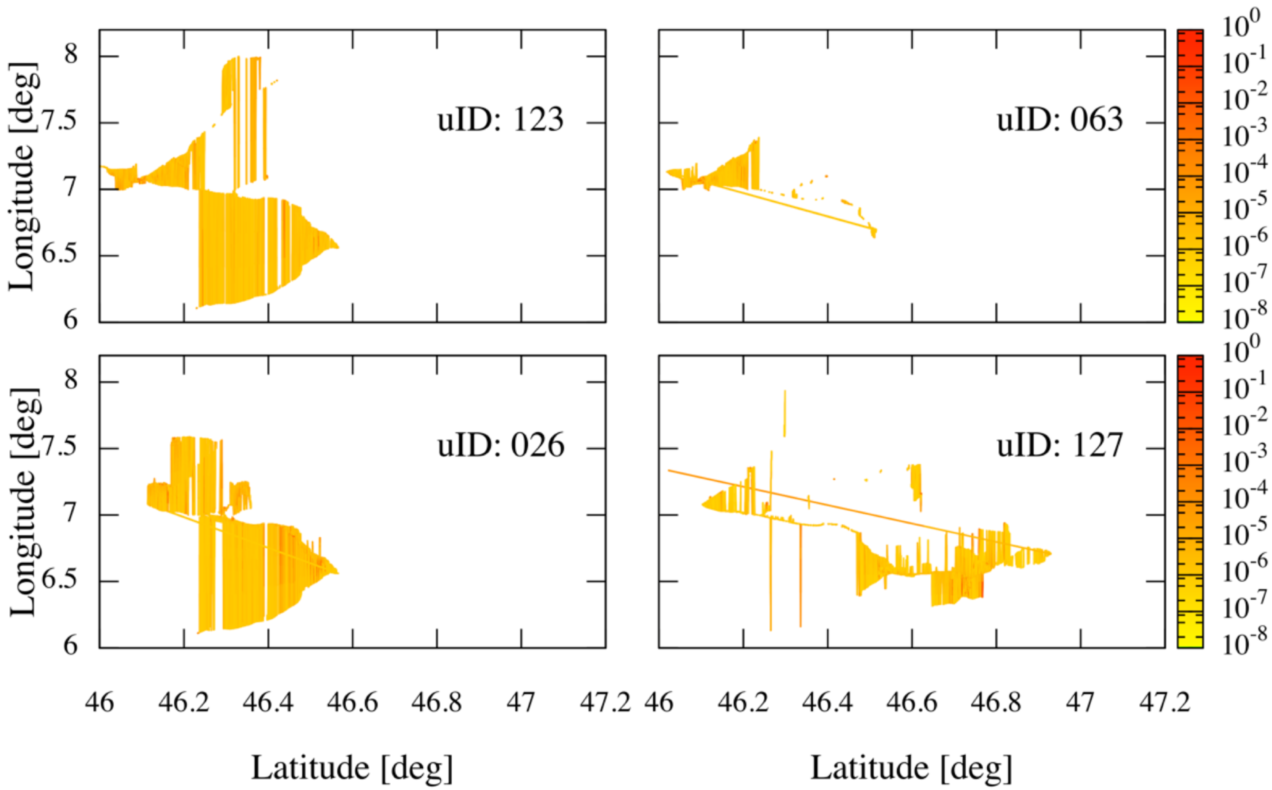
- The prediction is performed considering the average over the states which are *k* steps ahead of the neighbours states.
- In the reconstruction represented here for m=2, neighbours are inside the azure area.

Multivariate One-user Prediction



- Much better global prediction error of 0.19 for lat/lng and 219.43 m for the altitude.

Mobility Probability Density Function



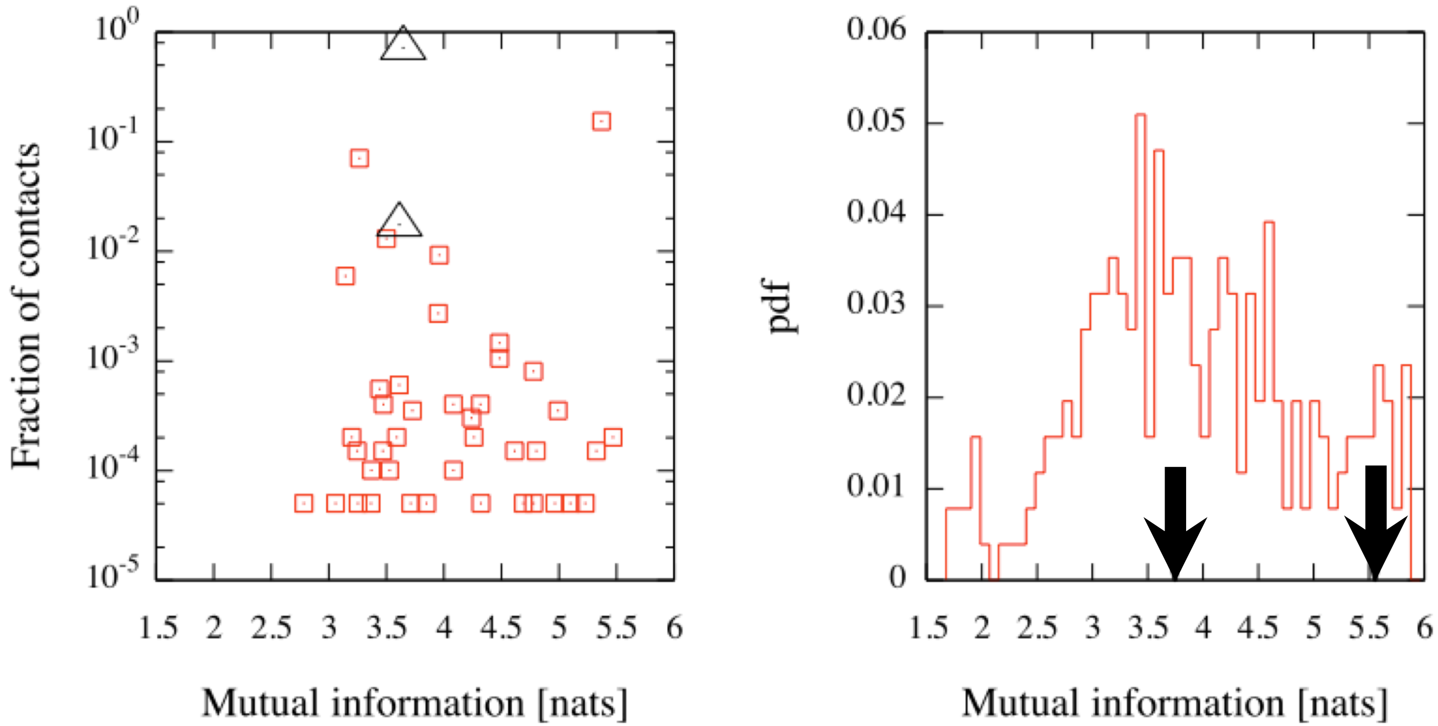
PDF of positions of users who are friends (top) and who are not friends (bottom)

Mutual Information

$$\mathcal{I}(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in \mathbf{Y}} P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \log \frac{P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{Y}}(\mathbf{y})}$$

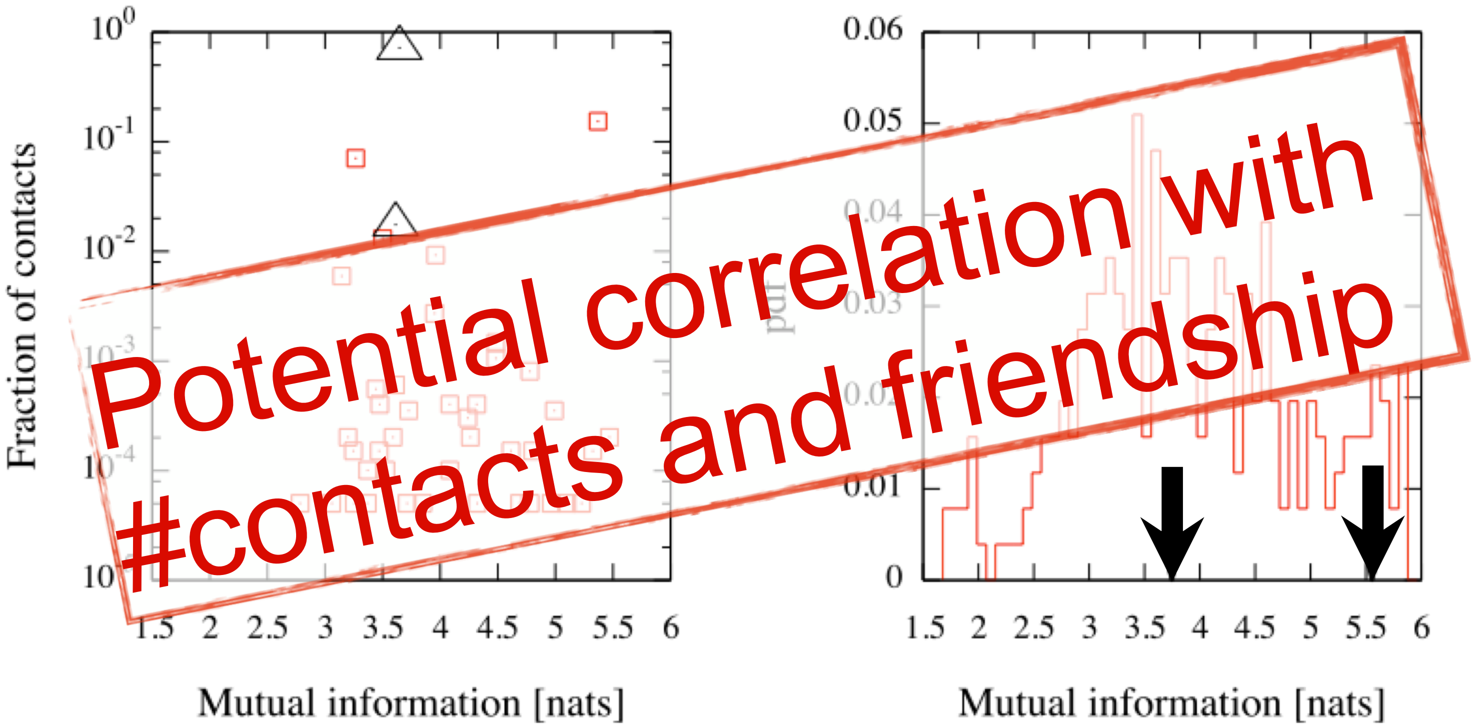
- The mutual information quantifies how much information a stochastic variable can provide about another stochastic variable. It can be used as an estimator of the amount of correlation between them.
- If they are uncorrelated, it is null.
- We use it to quantify how much the motion of a user can give us information about the motion of another.

Mutual Information, Contacts, Friendship



M.I. for people with at least one contact (left) and for people with no contacts at all (right).

Mutual Information, Contacts, Friendship



M.I. for people with at least one contact (left) and for people with no contacts at all (right).

Multivariate two-users prediction

Nodes	Social Link	Pos. Error [deg]	Alt. error [m]
026, 127	None	0.167	66.33
063, 123	Present	0.011	20.95
094, 009	Present	0.003	5.57

- The accuracy of the prediction improves by at least one order of magnitude (often two).

Reference

Manlio De Domenico, Antonio Lima and Mirco Musolesi. Interdependence and Predictability of Human Mobility and Social Interactions. In Pervasive and Mobile Computing. Volume 9. Issue 6. December 2013.

Current research directions

- Predicting long-term mobility
- Individual mobility vs group mobility
- Implications for the design of systems
  - Recommender systems
  - Navigation tools
- General scientific questions:
  - Human movement patterns from an evolutionary point of view

Prediction is difficult, especially about the future.

Niels Bohr

Questions?

Mirco Musolesi

Intelligent Social Systems Lab
Department of Geography
University College London

W: http://www.ucl.ac.uk/~ucfamus
E: m.musolesi@ucl.ac.uk
T: @mircomusolesi