Using Mobile Phones to Estimate Arterial Traffic through Statistical Learning

R. Herring, A. Hofleitner, S. Amin, T. Nasr, A. Khalek, P. Abbeel, A. Bayen

UC Berkeley – *Mobile Millennium* Project

1. Mobile Phones and Arterial Traffic

- Mobile phones enable **privacy-aware, participatory sensing**
  - *Virtual Trip Lines* (VTL) are virtual markers stored by the phone.
  - Provide **travel time** measurements between consecutive VTLs.
- Arterial traffic modeling **challenges**
  - **Flow discontinuities** due to signals, pedestrians, etc.
  - **Sparse** data arriving at **irregular** times.
- Our approach: **Statistical Learning**
  - Studied many techniques including **regression** and **belief propagation**.
  - Learn **traffic patterns** for each time interval of the week from past data.
  - Use belief propagation to estimate and predict traffic conditions in real-time, even for road segments with no current data.

Raw Data

Mobile Millennium Traffic Viewer

2. Statistical Learning

Learning – Use **past** data to learn statistical traffic dynamics based on traffic theory assumptions:

- **Training** – For given assumptions on traffic dynamics, use a learning technique to estimate model parameters: spatio-temporal dependencies of the network, travel time distributions.
- **Validation** – compute error, compare different hypothesis and models (tune meta-parameters) and re-train.
- **Testing** – only compute error on this set once finished with first 2 parts.

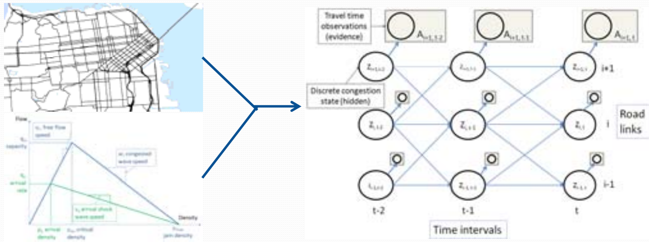
Inference – Use **learned patterns** and **current data** to do real-time estimation and prediction (also known as **Data Assimilation**)

Standard Techniques: **Regression**

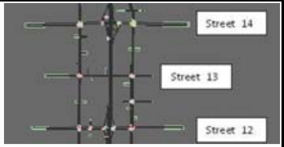
- Logistic Regression: estimate and predict Level Of Service,
- STARMA (Spatio-Temporal Auto Regressive Moving Average): estimate and predict mean value of traffic variable (travel time)

3. Real time estimation and prediction

- Graphical model with **hidden** Level Of Service (LoS) states:
  - **Expectation Maximization** (EM) Algorithm:
    - ➢ Learn travel time **distribution** with **sparse** (and/or) **missing** data.
    - ➢ Learn **spatio-temporal dependencies** between links of the network.
  - Discrete states represent level of congestion, each level of congestion is associated with a travel time distribution.
  - Real-time **estimation/prediction** via **particle filtering** using learned parameters (Sampling Importance Re-sampling).



4. Experiment design



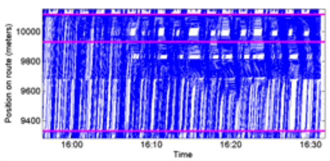
Paramics Simulation Network

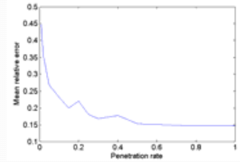
- **Data sources:**
  - Paramics software: extract travel time data from trajectories
  - New York: 3 field tests with 20 cars each
  - San Francisco Taxis: time sampling of the location of a fleet of taxis
- **Methodology:** Training, Validation, Test

Independent datasets to prevent over fitting and choose the optimal parameters.

- **Penetration rate study:**
  - How does the quality of our method vary with the percentage of vehicles probed?
  - Extract probe vehicles from all trajectories.

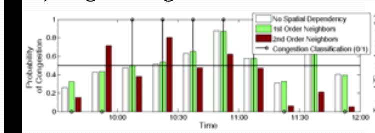
New York Experiment Network



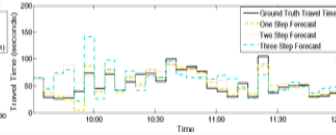


5. Results

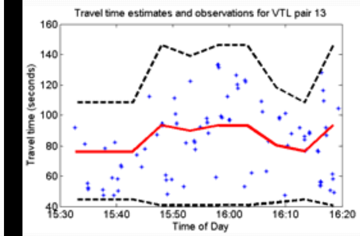
1) Logistic regression



2) STARMA



3) Graphical model



Red: estimate of the mean travel time

Black: estimate of the mean plus/minus 2 standard deviations

Blue: actual travel time values

Wasserstein error: distance between two probability distributions (estimated *f(x)* and empirical *g(y)*) defined as:

$$\left(\inf_{\gamma\in\Gamma}\int |x-y|^p \,\mathrm{d}\gamma(x,y)\right)^{1/p}$$

where γ is a joint distributions on *(x,y)* with marginals *f(x)* and *g(y)*

<u>*p=2*</u>: $W=\sqrt{\mu_x^2+\mu_y^2+\sigma_x^2+\sigma_y^2-2\mu_x\mu_y-2\sigma_x\sigma_y}$

Mean error of 14.5% on validation set for graphical model. (Logistic regression and STARMA don't estimate a distribution).

6. Conclusion and perspectives

- Statistical learning is an excellent tool for monitoring arterial traffic due to the **repetitive nature of traffic conditions**.
  - Identifies **patterns**
  - Robust to **missing data**
  - Adapts to changing conditions (re-learning)
  - Leverages **arterial traffic theory**
  - **Flexible** framework allows for incorporating many different data sources all into one model
- Future directions:
  - Base learning algorithm on specific **features**
    - ➢ School hours
    - ➢ Special **events** (sports, concerts)
    - ➢ **Weather**
  - **Share** learned parameters to speed up deployment in new locations
    - ➢ Parameters learned for San Francisco should be applicable with minor modifications to new areas (such as New York, Los Angeles,…)