Markov modeling and simulation of traffic flow

Márton Ispány & Péter Jeszenszky

Faculty of Informatics, University of Debrecen
Hungary

Joint work with
N. Bátfai, R. Besenczi, R. S. Major and F. Monori

Data Science, Statistics & Visualisation
TU Wien, Austria
9-11 July 2018

---

1 The work is supported by the EFOP-3.6.1-16-2016-00022 project. The project is co-financed by the European Union and the European Social Fund.
Outline

- Smart City and urban traffic background
- Previous work: OOCWC traffic simulator
- Overview on traffic simulation
- Markov modeling of the traffic
- Estimation in case of mobile sensors
- Real data analysis
- Conclusions and future work
Smart City background

- **Smart City** (idea came from the 1990s): connected cities, intelligent cities, digital cities, etc.
- Technological components: big data analytics, cloud computing, complex sensor systems complemented with participatory (or mobile) sensing (IoT).
- IoT is visioned to become true with 50 billion devices around 2020.
- Smart City applications, offered by city administrations, provide services to inhabitants:
  - intelligent city planning;
  - crowd-sourcing;
  - crisis and disaster management.
- By the year 2050, 70% of Earth’s population is expected to live in urban areas.
- New challenges for city infrastructures: e.g. urban traffic.
Developments in the automobile industry:
- Between 2000 and 2010, most cars were equipped with some sort of driving-aid system (e.g. pedestrian observers, lane support systems).
- Pure electric cars are on the market for a few years.
- In the 2020s, the widespread of driverless or autonomous cars are expected.
- These vehicles have advanced on-board computers for perception and controlling.

How can a (smart) city administration assist the widespread of these cars?

Hypothesis: the answers will come from a centralized IT infrastructure maintained by city administration.

The city may possess all the information necessary, such as traffic jams, accidents, detours, etc to produce optimal routes for driverless and electric cars.

Development and deployment of intelligent transportation systems (ITs).
rObOCar World Championship (OOCWC): a multi-agent oriented environment for creating urban traffic simulations. **Purpose**: offer a research and educational platform to investigate urban traffic control algorithms and the relationship between smart cities and robot cars.

**Architecture:**

- **Map**: OpenStreetMap
- **City**: connected part of OSM
- **ASA**: Automated Sensor Annotations
- **HSA**: Human controlled Sensor Annotations
- **Robocar City Emulator**: simulation
- **The competition**: e.g. the Police edition
- **Results**: for a competition or an analysis
- **Monitors**: visualisation
Visualisation
Crowd-sourcing based traffic simulation
Distribution of the roads during the simulation

Starting from 3 roads we have

(a) At the beginning of the simulation the number of roads is 78.

(b) After a minute the number of roads is 1085.

Purpose: find a distribution preserving simulation method!
Properties of transportation systems:
- defined on large scale, changing in time;
- complex, nonlinear and stochastic;
- multiple aggregation levels.

Traffic flow data: $X_t^i \in \mathbb{R}^p$, $i = 1, 2, \ldots, m$ (space), $t = 1, 2, \ldots, T$ (time).

Traffic flow modeling: find appropriate temporal and spatial process for describing some properties of the traffic.

Traffic flow prediction: predict the traffic flow at time interval $T + \Delta$ for some prediction horizon $\Delta$ at different places.

Traffic simulation:
- Macroscopic simulation: statistical dispersion models, freeway traffic models, etc.
- Microscopic simulation: cellular automata, multi-agent simulation, particle system simulation, etc.

Traffic flow prediction is a key functional component in ITSs.
Simulation packages

- **MatSim - Version 0.8.0**: Multi-Agent Transport Simulation is an open-source framework to implement large-scale agent-based transport simulations and transport planning models.

- **SUMO - Version 0.28.0**: Simulation of Urban Mobility is an open-source, highly portable, microscopic and continuous road traffic simulation package design for large road networks.

- **Aimsun - Version 8.1**: traffic modeling software environment, which integrates travel demand modeling, macroscopic functionalities and the meso- and microscopic hybrid simulator.

- **PTV Vissim - Version 9**: microscopic multi-modal traffic flow simulation package that allows you to simulate exact traffic patterns and displays all road users and their interactions in one model.
Box-Jenkins time-series analyses with ARIMA (Kohonen, S~, ~X) (Levin & Tsao Lee & Fambro, 1999; Stathopolous & Karlaftis, 2003; Ghosh et al., 2005; Chen et al., 2008; Xue & Shi, 2008),

Kalman filter theory (Wang et al., 2006; Ngoduy, 2008),

Non-parametric methods (k-NN, kernel, local regression) (Davis & Nihan, 1991; Smith et al., 2002; Turochy & Pierce, 2004, Smith & Demetsky, 1997),

Exponential smoothing (Messer, 1993; Gould et al., 2008; Castro-Neto et al., 2009),

Simulation models (Duncan & Littlejohn, 1997; Chrobok et al., 2001),

Spectral analysis (Nicholson & Swann, 1974),

Wavelet (Jiang & Adeli, 2005; Sun et al., 2006; Xie & Zhang, 2006; Cheng et al., 2007)
Machine learning and data mining techniques:
- Support vector regression, SVR (Jeong et al., 2013)
- Artificial neural network, ANN models (Chan et al., 2012; Park et al., 1998; Dia, 2001)
- Bayesian network (Sun et al. 2006)
- Deep learning (Yisheng Lv et al. 2015)

Computational intelligence techniques:
- Linear genetic programming, LGP (Brameier & Banzhaf, 2007)
- Fuzzy logic, FL (Iokibe et al., 1993; Li et al., 2006; Zhang & Ye, 2008; Srinivasan et al., 2009)

Markov models:
- Fuzzy partitioning (Chen et al., online 2017)
- Variable-order Markov models (Necula, 2014)
Road network

Notations:
\[ G = (V, E) \] directed graph (digraph)
\[ u, v, w \in V \] vertices
\[ e, f, g \in E \] edges,
\[ e = (v, w) = v \rightarrow w \]

Definition

A road network \( G \) is a simple directed graph, \( G = (V, E) \), where \( V \) is a set of nodes representing the terminal points of road segments, and \( E \) is a set of directed edges denoting road segments.

A road segment \( e = (v, w) \in E \) is a directed edge in the road network graphs, with two terminal points \( v \) and \( w \). The vehicle flow on this edge is from \( v \) to \( w \).

Second-order graph (line digraph):
\[ L(G) = (V', E') \] where \( V' = E \)

Assumption: the road network \( G \) is closed.
Consider the road graph $G$ below and its closure $\overline{G}$:

![Graph Image]

Then, the adjacency matrices are given as

$$A_G := 
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
1 & 0 & 1 & 0 & 0 \\
2 & 1 & 0 & 1 & 0 \\
3 & 0 & 0 & 0 & 1 \\
4 & 0 & 1 & 0 & 0 \\
\end{array}$$

$$A_{\overline{G}} := 
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
2 & 0 & 1 & 0 & 1 & 0 \\
3 & 0 & 0 & 0 & 0 & 1 \\
4 & 1 & 0 & 1 & 0 & 0 \\
\end{array}$$
Probability distribution (p.d.) on $G$:
\[
\pi : = (\pi_v)_{v \in V} : \quad \begin{align*}
(i) & \quad \pi_v \geq 0 \ \forall v \in V; \\
(ii) & \quad \sum_{v \in V} \pi_v = 1.
\end{align*}
\]

Markov kernel on $G$:
\[
P : = (P_{uv})_{u,v \in V} : \quad \begin{align*}
(i) & \quad P_{uv} \geq 0 \ \forall u, v \in V; \\
(ii) & \quad \sum_{v \in V} P_{uv} = 1 \ \forall u \in V; \\
(iii) & \quad P_{uv} > 0 \ \text{implies} \ u \rightarrow v \ \text{or} \ u = v \\
& \quad \text{(G-compatibility)}
\end{align*}
\]

Stationary distribution (s.d.) of $P$:
\[
\pi : = (\pi_v)_{v \in V} : \quad \sum_{u \in V} \pi_u P_{uv} = \pi_v \ \forall v \in V
\]
A toy example (2)

Markov modeling and simulation of traffic flow
Asymptotic behavior

Theorem
If the road network $G$ is strongly connected then there is a unique stationary distribution $\pi$ to any $G$-compatible Markov kernel $P$. Moreover, this distribution satisfies $\pi_v > 0$ for all $v \in V$.

Theorem
If a road network $G$ is strongly connected then any $G$-compatible Markov kernel $P$ is ergodic in the sense that the average Markov kernel $A_n := (n+1)^{-1}(I + P + \ldots + P^n)$, $n \in \mathbb{N}$, converges to a limiting Markov kernel $\Pi$ with all rows the same vector $\pi$, the unique stationary distribution of $P$. Moreover, if $G$ is an aperiodic, then the sequence of Markov kernels $P^n$, $n \in \mathbb{N}$, converges to the limiting Markov kernel $\Pi$. 
A macroscopic model for traffic flow.

**Definition**

Let the road network $G$ be strongly connected and let $P$ be a $G$-compatible Markov kernel on $V$ with unique s.d. $\pi$. Then, the triplet $(G, P, \pi)$ is called **Markov traffic** on the road network $G$. Let $\{X_t\}_{t \in \mathbb{Z}^+}$ be a Markov chain on $V$ such that $\pi X_0 = \pi$ and $X_t|X_{t-1} \sim P$ for all $t \in \mathbb{N}$. Then, $\{X_t\}_{t \in \mathbb{Z}^+}$ is called **Markov random walk** in the Markov traffic $(G, P, \pi)$.

**Definition**

A matrix $Q = (q_{uv})_{u,v \in V}$ is called **two-dimensional stationary distribution** on $G$ if (i) $q_{uv} \geq 0 \ \forall u, v \in V$ and $q_{uv} = 0$ provided $u \nrightarrow v$; (ii) $\sum_{u,v \in V} q_{uv} = 1$; and (iii) $\sum_{v \in V} q_{uv} = \sum_{v \in V} q_{vu} \ \forall u \in V$. 

DSSV2018 Márton Ispány & Péter Jeszenszky

Markov modeling and simulation of traffic flow 19/47
We estimate \((G, Q)\) using observed data instead of the triplet \((G, P, \pi)\).

- Either \(G\) is known apriori (by Google Maps or OpenStreetMap) or the road network should be reconstructed or fitted to an existing road network from observed data, e.g., GPS coordinates.

- We suppose that we have to estimate \(Q\).

There are two ways to gather traffic data:

- Using mobile sensors which can be vehicles, passengers etc.

- Using fixed sensors which can be cameras, sensors in the roads etc.

We investigate the mobile sensor case!
The data and basic statistics

Random sample of trajectories:

\[ X_1^i \rightarrow X_2^i \rightarrow \ldots \rightarrow X_{n_i}^i, \quad i = 1, \ldots, k, \]

- \( k \): number of sensors (trajectories);
- \( n_i \): length of the \( i \)-th trajectory;
- \( n := n_1 + \ldots + n_k \): cumulative length of trajectories;
- \( N_i = (n_{iuv})_{u,v \in V} \): two-dimensional frequencies of consecutive nodes in the \( i \)-th trajectory;
- \( N := N_1 + \ldots + N_k \): cumulative two-dimensional frequencies;
- \( s = (s_v)_{v \in V} \): starting frequencies of nodes in trajectories;
- \( e = (e_v)_{v \in V} \): ending frequencies of nodes in trajectories.
Frobenius distance on $G$: $A, B$ matrices of dimension $|V| \times |V|$

$$d_G^2(A, B) := \|A - B\|^2_G := \sum_{u,v:u\Rightarrow v} |a_{uv} - b_{uv}|^2$$

Squared error (SE) is defined as

$$\text{SE}(Q, n_{\text{eff}}|N) := n_{\text{eff}} \sum_{i=1}^{k} \left( n_{\text{eff}}^i \right)^{-1} d_G^2(N_i, n_{\text{eff}}^i Q)$$

where $N := (N_i)_{i=1,...,k}$, $n_{\text{eff}} := (n_{\text{eff}}^i)_{i=1,...,k}$ is the effective sample size for trajectories and $n_{\text{eff}} := \sum_{i=1}^{k} n_{\text{eff}}^i$.

We estimate the two-dimensional stationary distribution $Q$ and the effective sample size $n_{\text{eff}}$ by minimizing SE.
The symmetric unnormalized graph Laplacian matrix $L$ of a digraph $G$ is defined as

$$L := D - A - A^\top$$ (1)

where $D := \text{diag}\{d^+ + d^-\}$ where $d^\pm = (\text{deg}^\pm(v))_{v \in V}$.

Properties of the Laplacian

1. $L$ is symmetric and positive semi-definite.
2. The smallest eigenvalue of $L$ is 0, the corresponding eigenvector is the constant one vector $\mathbf{1}$.
3. $L$ has $|V|$ non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{|V|}$ and corresponding orthonormal eigenvectors $\mathbf{1} = \alpha_1, \alpha_2, \ldots, \alpha_{|V|}$ in $\mathcal{F}(V)$.
4. If $G$ is strongly connected then $\lambda_2 > 0$, i.e., the multiplicity of the eigenvalue $\lambda_1 = 0$ is 1, and $L$ is invertable on the invariant subspace $S := \{\alpha \in \mathcal{F}(V) | \alpha \perp \mathbf{1}\}$. 
Main theorem

There is a unique pair \((\hat{n}_{\text{eff}}, \hat{Q}_{\text{SE}})\) which minimize the squared error \(SE\). Moreover, the estimated effective sample sizes \(\hat{n}_{\text{eff}}, \hat{n}_{\text{eff}}\) and the estimated two-dimensional stationary distribution \(\hat{Q}_{\text{SE}}\) are given by

\[
\hat{n}_{\text{eff}}^i := w_i \hat{n}_{\text{eff}}, \quad \hat{w}_i := \frac{\|N_i\|_G}{\sum_{j=1}^k \|N_j\|_G},
\]

\[
\hat{n}_{\text{eff}} := (n - k) + (d^- - d^+)^\top \lambda,
\]

\[
\hat{Q}_{\text{SE}} := \hat{n}_{\text{eff}}^{-1} (N + (\lambda^\top - \lambda 1^\top) \circ A),
\]

where the Lagrange multiplicator \(\lambda \in S \subset \mathcal{F}(V)\) is defined as a unique solution to the linear equation \(L\lambda = s - e\) and \(\circ\) denotes the entrywise (Hadamard) product of matrices.
Consider our previous example with trajectories:

<table>
<thead>
<tr>
<th>Trajectory</th>
<th>Length</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 → 1 → 2 → 3 → 4</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>0 → 1 → 2 → 3</td>
<td>4</td>
<td>150</td>
</tr>
<tr>
<td>0 → 1 → 1 → 2</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>0 → 1 → 2 → 2 → 2 → 3</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>2 → 3 → 3 → 4 → 0</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td>2 → 1 → 1 → 1 → 0</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>3 → 4 → 4 → 0</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>4 → 2 → 1 → 0</td>
<td>4</td>
<td>150</td>
</tr>
<tr>
<td>3 → 3 → 4 → 4 → 0</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>1 → 2 → 3 → 4 → 4 → 0</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td>4 → 2 → 2 → 1</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>

We have:

\[
\lambda = \begin{bmatrix} -65.31 & -26.25 & 37.81 & -2.81 & 56.56 \end{bmatrix}^T
\]

\[
\pi = \begin{bmatrix} 0.11 & 0.22 & 0.29 & 0.19 & 0.19 \end{bmatrix}
\]

\[
Q = \begin{bmatrix}
0 & 0.11 & 0 & 0 & 0 \\
0.05 & 0.04 & 0.13 & 0 & 0 \\
0 & 0.07 & 0.09 & 0.13 & 0 \\
0 & 0 & 0 & 0.06 & 0.13 \\
0.06 & 0 & 0.07 & 0 & 0.06
\end{bmatrix}
\]

\[
P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0.21 & 0.19 & 0.6 & 0 & 0 \\
0 & 0.24 & 0.3 & 0.46 & 0 \\
0 & 0 & 0 & 0.3 & 0.7 \\
0.35 & 0 & 0.35 & 0 & 0.3
\end{bmatrix}
\]
For our experiments, we needed datasets of real-life traffic trajectory data.

**Requirements**

- Complete trajectories.
- High enough sampling rate.
- Sufficiently large size.
- Trajectories should cover a relatively small geographic area.
- Vehicles should not follow a fixed route.
- Public availability.
We found only a small number of suitable datasets, namely:

- Uber GPS Traces
  https://github.com/dima42/uber-gps-analysis

- ECML/PKDD 2015: Taxi Trajectory Prediction
  https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-

Because of its larger size we have chosen the second dataset to work with.
The Taxi Trajectory Prediction Dataset (1)

- Period: one year, from July 1, 2013 to June 30, 2014.
- Size: split into a training and a test set.
  - The training set contains 1,710,670 trajectories with 83,408,417 points.
  - The test set contains 320 trajectories with 14,438 points.
- Location: the city of Porto, in Portugal.
- Sampling period: 15 seconds.
Figure: Map of the area covered by the dataset (bounding box: (−15.630759, 36.886104, −3.930948, 51.037119)).
Figure: Hexbin plot of trajectory points.
Figure: Histogram of number of sample points per trajectories. The grey bar on the right represents trajectories with more than 200 sample points. On the average, a trajectory consists of 50 sample points and takes $\approx 12$ minutes.
Figure: Histogram of trajectory lengths. The grey bar on the right represents trajectories longer than 25,000 meters. The average trajectory length is 5630 meters.
Figure: Histogram of distance of consecutive trajectory points. The grey bar on the right represents trajectories longer than 500 meters. The average distance is $\approx 115$ meters.
Figure: Number of trajectories per day of week.
The Taxi Trajectory Prediction Dataset (8)

Figure: Number of trajectories per hour.
To be suitable for our experiments, we selected a subset of the dataset.

Subsetting is based on the geographic area defined by the bounding box \((-8.6518, 41.1129, -8.5771, 41.1756)\) (the city centre of Porto).

Trajectory points outside of this area are discarded.

The selected subset contains 1,550,044 trajectories with 57,663,357 points.

This means 91% of the total number of trajectories and 69% of the total number of points.
Figure: Map of the area covered by the selected subset of the dataset (bounding box \((-8.6518, 41.1129, -8.5771, 41.1756))\). The size of the area is about \(6,274\text{km} \times 6,963\text{km} = 43,68\text{km}^2\).
Figure: Hexbin plot of trajectory points of the selected subset.
Figure: 2D histogram of trajectory starting points (number of bins is $120 \times 120$).
Figure: 2D histogram of trajectory endpoints (number of bins is $120 \times 120$).
**Figure:** Plot of number of trajectory starting points minus trajectory endpoints per bin (number of bins is $120 \times 120$).
Figure: Plot of number of trajectory starting points minus trajectory endpoints per bin between 9:00–10:00 (number of bins is $120 \times 120$).
Figure: Plot of number of trajectory starting points minus trajectory endpoints per bin between 16:00–17:00 (number of bins is $120 \times 120$).
Figure: The Laplacian spectrum of Porto.
A Markov model is proposed for analyzing traffic flows which preserves its unique stationary distribution.

Currently, the model is verified by simulation using a small toy dataset.

The verification of the model and the simulation method based on the model for real large datasets is a work in progress.

Thank you for your attention!