Statistical Inference of Probabilistic Origin-Destination Demand Using Day-to-Day Traffic Data

Wei Ma, Sean Qian

Carnegie Mellon University, Civil and Environmental Engineering

MAC Mobility Data Analytics Center, Carnegie Mellon University

Carnegie Mellon University Civil and Environmental Engineering

Abstract

Motivation

Recent studies on transportation network uncertainty and reliability call for modeling the stochasticity of O-D demand and network flow. Few studies focus on estimating the mean and variance of O-D demand from day-to-day traffic data. The day-to-day variation of flow measurements stems not only from O-D demand but also from travelers' random route choices that vary from day to day.

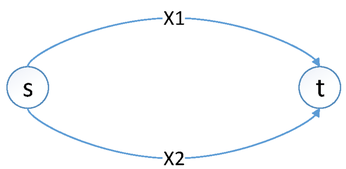
Objectives:

To develop a novel theoretical framework for estimating the mean and variance/covariance of O-D demand considering the variation induced by travelers' day-to-day random route choices.

Background

- ❑ Both classical traffic assignment models and O-D estimation methods overlook the variation of demand and link/path flow.
- ❑ Statistical traffic assignment models and corresponding probabilistic O-D estimation methods assume fixed portion of route choice.
- ❑ Travelers' route choice variation is an indispensable component of traffic variation.
- ❑ To our best knowledge, there is no method that estimates probabilistic O-D demand and considers the travelers' route choice variation simultaneously.

Illustrative example



- Link 1 and link 2 are exactly the same, observe the distribution of link 1 is $N(50, 100)$.
- If we do not consider the route choice variation, then demand follows $\mathrm{Q} \sim N(100{,}400)$ since $Q = 2X_1$.
- But if we consider the route choice variation:

$$\begin{aligned}(X_1, X_2)^T &\sim \text{Multinomial}(Q, (0.5, 0.5)^T)\\ Q &\sim \mathcal{N}(q, \text{Var}(Q))\end{aligned}$$

$$\begin{aligned}\text{Var}(X_1) &= \text{Var}(\mathbb{E}(X_1|Q)) + \mathbb{E}(\text{Var}(X_1|Q)) = p_1^2\text{Var}(Q) + p_1p_2\mathbb{E}(Q)\\ 100 &= 0.25\text{Var}(Q) + 0.25\mathbb{E}(Q)\\ 100 &= 0.25\text{Var}(Q) + 25\end{aligned}$$

- $Q \sim N(100, 300)$
- Overlooking the variation induced by travelers' route choice will lead to an overestimation of OD variance.

Model detail

Recurrent network flow

Level 1: $X_m \sim N(X + e, \Sigma_x + \Sigma_e)$ — Measurement error

Level 2: $X \sim N(\Delta p Q, \Sigma_x)$, $F \sim N(pQ, \Sigma_f)$ — Route choice variation

Level 3: $Q \sim N(q, \Sigma_q)$ — OD variation

Estimating joint probability distributions of O-D demand

- A disaggregated OD demand estimation algorithm is developed to estimate OD mean vector and variance-covariance matrix iteratively (IGLS).
- IGLS decomposes a complex estimation problem into two relatively simpler sub-problems, and therefore the entire solution algorithm is more friendly for practical use.

Estimate OD mean

- A statistical interpretation is given for estimate OD mean and the influence of number of data is analyzed.

$$\begin{aligned}\min_f \quad & n\left(\Delta^o f - \hat{x^o}\right)^T \Sigma_x^{o-1}\left(\Delta^o f - \hat{x^o}\right) + (q^H - Mf)^T \Sigma_q^{H-1}(q^H - Mf)\\ \text{s.t.} \quad & f \in \Phi^+\end{aligned}$$

Estimate OD variance

- MLE, convex relaxation, LASSO model selection.
- Solution algorithm: ISTA, FISTA

$$\begin{aligned}\min_{\Sigma_q} \quad & \|S_x^o - \Sigma_x^o\|_F^2 + \lambda\|\Sigma_q\|_1\\ \text{s.t.} \quad & \Sigma_x^o = \Delta^o\Sigma_{f|q}\Delta^{oT} + \Delta^o\tilde{p}\Sigma_q\tilde{p}^T\Delta^{oT}\\ & \Sigma_q \in \text{semidefinite}\left(\Re^{|K_q|\times|K_q|}\right)\end{aligned}$$

More issues

Model Observability

- The probabilistic OD demand is not guaranteed to be unique for given observed link flow data set x^o.
- It can be proved that under the proposed stochastic framework, the estimated O-D mean is no worse than that from deterministic O-D estimations.

Goodness of fit

- For probabilistic OD estimation method, the goodness of fit indicator should be built on the distance between the estimated distribution and true distribution.
- Hellinger distance or Kullback-Leibler distance can be adopted.

$$D_H((\mu_1,\Sigma_1)^T,(\mu_2,\Sigma_2)^T) = 1 - \frac{|\Sigma_1|^{\frac{1}{4}}|\Sigma_1|^{\frac{1}{4}}}{\left|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{8}(\mu_2-\mu_1)^T\left(\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right)^{-1}(\mu_2-\mu_1)\right) \quad (27)$$

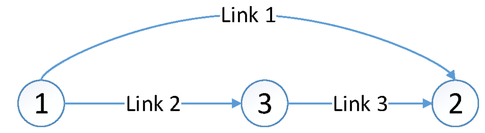
$$D_{KL}((\mu_1,\Sigma_1)^T,(\mu_2,\Sigma_2)^T) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{\Sigma_1} - d + \text{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right) \quad (28)$$

where d is the dimension of the vector μ_1.

Experiment results

Small network

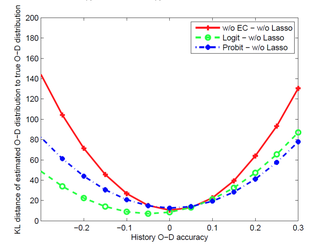
- We estimate probabilistic O-D using different methods on a simple network with three links and two O-D pairs as follows.

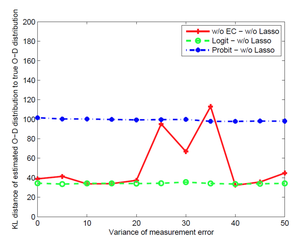


True ρ	Method	$\hat{q}_{1\to2}$	$\hat{q}_{3\to2}$	$\hat{\sigma}_{1\to2}$	$\hat{\sigma}_{3\to2}$	$\hat{\rho}$	RMPSE	KL-distance
	True	700	500	13.23	11.18	NA	NA	NA
0.5	w/o EC - w/o Lasso	611.02	588.01	14.18	11.46	0.37	14.64%	106.80
	Logit - w/o Lasso	728.55	588.85	12.81	11.07	0.55	11.00%	33.78
	Probit - w/o Lasso	618.63	590.17	15.34	11.36	0.45	14.31%	101.69
0	w/o EC - w/o Lasso	765, 43	588.78	11.95	10.25	0.03	12.90%	43.08
	Logit - w/o Lasso	727.94	588.05	13.05	11.00	0.05	10.89%	33.24
	Probit - w/o Lasso	618.58	587.79	13.79	11.16	0.07	14.02%	49.07
	Logit - w/ Lasso	728.27	588.48	12.87	11.53	0.00	10.95%	33.60
	Probit - w/ Lasso	621.39	588.56	13.71	11.67	0.00	13.94%	49.04
−0.5	w/o EC - w/o Lasso	780.12	586.04	7.80	11.18	−0.58	13.85%	95.26
	Logit - w/o Lasso	726.41	586.08	12.99	11.12	−0.58	10.61%	52.44
	Probit - w/o Lasso	621.40	588.66	13.29	10.68	−0.37	13.79%	32.90

w/o EC: without equilibrium constraint; Logit/Probit: using Logit/Probit based SUE constraint; w/o Lasso: without Lasso regularization; w/ Lasso: using Lasso regularization

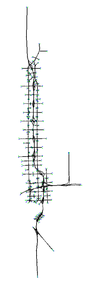
- Sensitivity analysis

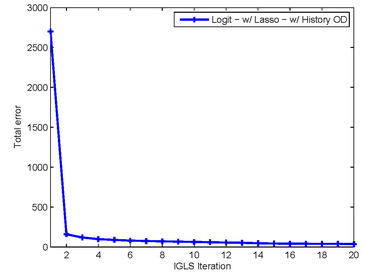




- History O-D information has fundamental influence on the quality of estimated O-D demand.
- The variance of measurement error and number of data has little influence on Probit and Logit based estimation methods, while the no equilibrium constrained method is unstable to these two factors.

Large network: SR41 corridor





- Logit - w/ Lasso - w/ History O-D method on a desktop computer (Inter(R) Core 529 i5-4460 3.20 GHz 2, RAM 8 GB)
- Average computation time for each IGLS iteration is 233.05s.
- Converge in 20 iterations.