View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 7 (2015) 192 – 211

Transportation Research Procedia

www.elsevier.com/locate/procedia

ELSEVIER

21st International Symposium on Transportation and Traffic Theory, ISTTT21 2015, 5-7 August 2015, Kobe, Japan

Estimation of mean and covariance of stochastic multi-class OD demands from classified traffic counts

Hu Shao[a,b], William H. K. Lam[a,c,*] , Agachai Sumalee[a,d], Martin L. Hazelton[e]

[a]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[b]Department of Mathematics, School of Sciences, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China
[c]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
[d]Department of Civil Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand.
[e]Institute of Fundamental Sciences, Massey University, Palmerston North 4442, New Zealand.

Abstract

This paper proposes a new model to estimate the mean and covariance of stochastic multi-class (multiple vehicle classes) origin-destination (OD) demands from hourly classified traffic counts throughout the whole year. It is usually assumed in the conventional OD demand estimation models that the OD demand by vehicle class is deterministic. Little attention is given on the estimation of the statistical properties of stochastic OD demands as well as their covariance between different vehicle classes. Also, the interactions between different vehicle classes in OD demand are ignored such as the change of modes between private car and taxi during a particular hourly period over the year. To fill these two gaps, the mean and covariance matrix of stochastic multi-class OD demands for the same hourly period over the year are simultaneously estimated by a modified lasso (least absolute shrinkage and selection operator) method. The estimated covariance matrix of stochastic multi-class OD demands can be used to capture the statistical dependency of traffic demands between different vehicle classes. In this paper, the proposed model is formulated as a non-linear constrained optimization problem. An exterior penalty algorithm is adapted to solve the proposed model. Numerical examples are presented to illustrate the applications of the proposed model together with some insightful findings on the importance of covariance of OD demand between difference vehicle classes.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Selection and peer-review under responsibility of Kobe University

Keywords: OD demand estimation; multiple vehicle class; covariance matrix; classified traffic counts.

* Corresponding author. Tel.: +852-27666045; Fax: +852- 23659291.
E-mail address: william.lam@polyu.edu.hk

2352-1465 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Selection and peer-review under responsibility of Kobe University

doi:10.1016/j.trpro.2015.06.011

1. Introduction

Origin-destination (OD) traffic demand is one of the fundamental input data for transportation planning and traffic management. In the past decades, OD demand estimation from traffic counts has been an important topic in the field of transportation research so as to minimize the cost for data collection. However, most of the existing OD demand estimation models ignore two important features of the OD demands as follows.

- The interactions between different vehicle classes (or types) in OD demand, such as taxis, private cars and goods vehicles.
- The statistical characteristics of multi-class OD demands, such as the covariance of traffic demands between different vehicle classes.

This paper proposes a new model for estimation of the mean and covariance of stochastic multi-class (i.e. multiple vehicle classes) OD demands from hourly classified traffic counts throughout the whole year.

1.1. Covariance of OD demands

Due to daily and seasonal variations in activity patterns, the OD traffic demands of different vehicle classes during the same hourly period (e.g. morning peak, 8:00 am - 9:00 am) are stochastically varied from day to day over the whole year. This type of varying traffic demands is referred to as stochastic multi-class OD demands in this paper. Statistically, the random characteristics of the stochastic multi-class OD demands can be reflected by their mean and covariance. In the conventional OD demand estimation models, focus is usually put on the mean of the OD demands while the covariances of OD demands by vehicle class have not been considered. The covariance of stochastic multi-class OD demands would however reflect the correlations between OD demands by vehicle class. For instance, for the traffic demand of the same OD pair, the higher the private car usage, the less the taxis usage.

It should be pointed out that there are generally three categories of OD demand covariances, i.e. the spatial, temporal and vehicle class covariances. Firstly, spatial OD demand covariance refers to the correlation (or dependency to some extent) of the OD demands during the same hourly periods between different OD pairs in a spatial manner (Shao et al., 2014). Secondly, temporal OD demand covariance represents the correlation of the OD demands for the same OD pair between different time periods (e.g. 8:00 am - 9:00 am and 9:00 am - 10:00 am). Thirdly, the vehicle class OD demand covariance relates to the statistical dependency of traffic demand between different vehicle classes of the same OD pair. These three categories of OD demand covariances simultaneously exist and contribute to the stochasticity of OD demand in reality. However, in order to facilitate the essential ideas on correlations between OD demands by vehicle class, this paper ignores first and second categories of the covariances. Specifically, this paper aims to estimate the third category of OD demand covariance using classified traffic counts for the same hourly period over the year.

In road transportation networks, the covariance of OD demands should not be ignored particularly for OD demand estimation from traffic counts. The ignorance of the correlation between random variables may lead to very different output of the models (Haas, 1999). For example, Waller et al. (2001) found that the correlation level of OD demands plays a major role in determining the degree of error in relation to the expected total network travel time. Zhao and Kockelman (2002) discussed the propagation of errors through the four-step traffic demand forecasting model. They stated that neglecting the correlation of data (e.g., OD demands) would ultimately reduce the reliability of the traffic forecasts, and in turn affect the policy-making and infrastructure decisions. Duthie et al. (2011) found that the assumption of independent demands when correlations do in fact exist could lead to errors in the estimation of system performance and result in poor policy decisions; for instance, building highways may not be able to meet a higher-than-expected demand. Shao et al. (2014) found that the spatial covariance has significant impact of network performance evaluation. In view of the above studies, it is shown that the correlation between OD demands should not be overlooked. This has a great effect on OD demand estimation problem particularly with the use of traffic counts.

The mean and covariance of the stochastic multi-class OD demands to be estimated in this paper can be used in the reliability-based traffic assignment models which have recently been developed for multi-modal transportation networks with uncertainty (Chen et al., 2002; Nakayama and Takayama, 2003; Clark and Watling, 2005; Shao et al., 2006; Lam et al., 2008; Chen and Zhou, 2010; Chen et al., 2011; Sumalee et al., 2011). These relevant studies demonstrated that increasing attention has been given on development of reliability-based network equilibrium models but it was assumed that the probability distributions of stochastic multi-class OD demands are known and given. However, to the best of our knowledge, less attention has been paid to estimation of the probability distribution of the stochastic multi-class OD demands, which is a necessary input for the application of the above models in networks with uncertainty. It is known that the mean and covariance matrix are the two key parameters to characterize the OD demand probability distribution. Therefore, the estimation of mean and covariance of the stochastic multi-class OD demands is an important extension of the current research work on reliability-based network equilibrium models.

1.2. Interactions between multiple vehicle classes in OD demand

Conventionally, the OD demands of different vehicle classes can be estimated by the combined model, which combines two steps of the conventional four-step trip-based model, i.e. mode and path (or route) choices. The combined model can be regarded as a “top-down” approach to reflect the interactions between multiple vehicle classes as shown in Figure 1. In the combined model, travel modes are classified by different vehicle classes, such as cars, taxis, buses, heavy trucks, light trucks, etc. The sum of OD demands for all modes is available, and the choice of each mode is a function of travel times between an OD pair using alternative travel modes. At equilibrium, no user can change his route to lessen his travel cost, but may do so by changing his mode (Lam and Huang, 1992a). The corresponding combined model is a process of integration of the modal split and traffic assignment, which can be used to estimate the OD demand of each travel mode (vehicle class). However, the accuracy of the combined model results depends on the value of model parameters which need to be calibrated by expensive household interview surveys (Frank, 1978; Lam and Huang, 1992b).

In comparison with the combined model, this paper proposes a new model which is a “bottom-up” approach to estimate the mean and covariance of stochastic multi-class OD demands so as to account for the interactions between multiple vehicle classes in OD demand. In practice, the observed link traffic counts (flows) can be classified by vehicle class with the use of advanced technologies. For example, the automatic vehicle identification (AVI) reader, which is being used in Hong Kong for electronic toll collection and journey time estimation purposes (Tam and Lam, 2008), can recognize the vehicle class. The observed classified link traffic counts during the same hourly period (e.g. morning peak, 8:00 am - 9:00 am) are stochastically fluctuated from day to day over the whole year. Thus, classified link traffic counts are considered as random variables. The sample mean and covariance of traffic counts for all vehicle classes can then be calculated using the observed classified link traffic counts. It should be noted that the sample covariance of classified link traffic counts would capture the statistical dependence between different vehicle classes. It is a result of the interactions between multiple vehicle classes in traffic demand as well as the effect of the route or path choices. Such information and/or relationships are adopted in this paper to estimate the covariance of stochastic multi-class OD demands.

Apart from the covariance between different vehicle classes in OD demands, the mean OD demand for each vehicle class is also a kind of important information for transportation planning and traffic management. The mean OD demand by vehicle class can be used to assess the link choice proportion by vehicle class (i.e. the proportion of each vehicle class on links of the transportation network). Such information is useful for road traffic operation and control, traffic accident management and transportation policy evaluation. Although traffic flows of different vehicle classes can be uniformly converted into the passenger car units (pcu) for evaluation purpose, vehicle class is obviously an important factor that influences the operating capability and traffic accident severity in road transportation networks. For example, heavy vehicles have lower operating capabilities than private cars, particularly with respect to acceleration, deceleration, and the ability to maintain speed on hilly roads (Transportation Research Board, 2000). Involvement of motor cycles in multiple-vehicle accidents has a higher proportion of fatal or serious injuries than that of private cars and taxis (Yau, 2006). Therefore, estimation of mean value for different vehicle classes in OD demands can help the traffic manager effectively operate and evaluate the road transportation system particularly in congested network with uncertainty.

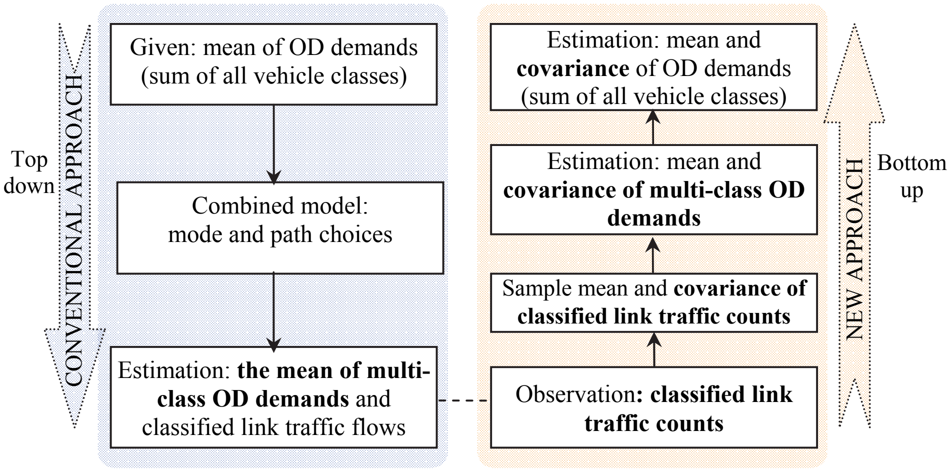

Top down
CONVENTIONAL APPROACH
Given: mean of OD demands (sum of all vehicle classes)
Combined model: mode and path choices
Estimation: **the mean of multi-class OD demands** and classified link traffic flows
Estimation: mean and **covariance** of OD demands (sum of all vehicle classes)
Estimation: mean and **covariance of multi-class OD demands**
Sample mean and **covariance of classified link traffic counts**
Observation**: classified link traffic counts**
NEW APPROACH
Bottom up

Fig. 1. Basic motivation and framework for the proposed model

1.3. OD demand estimation models

Although various methods for OD demand estimation from traffic counts have been widely investigated in the past decades, the existing estimation models can not fully capture the statistical characteristics as well as the interactions between different vehicle classes of OD demands as shown in Table 1. These OD demand estimation models include entropy maximizing model (Van Zuylen and Willumsen, 1980), maximum likelihood model (Spiess, 1987; Watling, 1994), generalized least squares (GLS) model (Cascetta, 1984; Bell, 1991), stochastic mapping method (Ashok and Ben-Akiva, 2000 and 2002), Bayesian inference estimation model (Maher, 1983; Sun et al., 2006; Castillo et al., 2008a,b) and Markov chain model (Li, 2009). It is assumed in these estimation models that the OD demands are either deterministic variables or mutually independent random variables. The first outstanding feature of the existing OD demand estimation models is that some in-depth statistical properties were generally ignored. For example, few of them addressed the OD demand variations, such as the correlation between random traffic demands of different vehicle classes (covariance). In general, only the mean of the OD demand was estimated in most of the existing models. An exception was the research work by Shao et al. (2014), which considered the estimation of spatial covariance of OD demands. However, this work may suffer from the difficulty of the “overfitting” problem as there are too many paramters to be estimated in the covarince matrix of stochastic OD demands. Also, the path choice proportion and/or the link choice proportion is assumed to be a deterministic variable. Such an assumption can not account for the stochastic path choice behaviours under condition with uncertainty. The second outstanding feature of the existing models is that most of them only considers the OD demand estimation problem for single vehicle class except the work by Wong et al. (2005). In the latter, only the mean of multi-class OD demands were estimated. In view of the above two outstanding features of the exisiting OD demand estimation models, this paper proposes a new model to estimate simutaneously both the mean and covariance of multi-class OD demands while explicitly considering the stochastic mode and path choice behaviours.

Table 1. Classfications of OD demand estimation models

		Statistical characteristics	
		Mean	Covariance
Vehicle classes	Single	Most of the existing OD demand estimation models with the use of traffic counts	Shao et al. (2014)
	Multiple	Wong et al. (2005)	**This paper**

1.4. Contribution statement

It is well-known that a common difficulty of the existing OD demand estimation models is the identifiability problem, i.e. it is impossible to identify the unique OD demand matrix from the observed traffic counts as the number of observed links (from which the traffic counts are available) is usually less than the number of parameters (or unknowns) to be estimated in the OD demand matrix (Hazelton, 2003). For the problem of estimating the mean and covariance matrix of the stochastic multi-class OD demands, the identifiability problem may be exacerbated as more parameters (say for instance, the covariance matrix) are required to be estimated. In view of this, the shrinkage estimator, a technique that is useful for estimating large-dimensional parameters with comparatively fewer observations, has the potential to overcome this difficulty in practice.

The widely-used shrinkage estimator, lasso (least absolute shrinkage and selection operator) method (Tibshirani, 1996), is adopted to address the identifiability problem in this paper. There are a large number of elements (or parameters) in the covariance matrix of stochastic multi-class OD demands. For example, in a simple network with two OD pairs and three vehicle classes, there are 21 parameters needed to be estimated in the 6×6 covariance matrix of multi-class OD demands. It should be noted that the covariance matrix is symmetrical. Thus, for a $n\times n$ covariance matrix, there are actually $(n+1)\times n/2$ elements needed to be estimated. However, in reality, some elements in the multi-class OD demand covariance matrix may be zero. For instance, the traffic demands between private cars and goods vehicles of the same OD pair may be independent with each other. And the corresponding elements in the covariance matrix should be zero. Due to the identifiability difficulty, the reasonable solutions of these covariances (i.e. zero) may not be identified as multiple solutions may exist. To overcome this difficulty, the lasso method does variable selection and shrinkage by solving the L1-penalized least squares (or linear regression) (Tibshirani, 1996 and 2011). The lasso method is equivalent to minimization of the sum of squares with a constraint of L1-norm of estimated parameters. For the covariance between independent multi-class OD demands (e.g. private cars and goods vehicles), the adoption of the lasso method can make it shrunk towards zero. As such, it is expected that covariance matrix of the OD demands could be uniquely identified.

This paper proposes a new model for estimating the mean and covariance of stochastic OD demands from classified traffic counts for the same hourly period over the year. The proposed model extends the existing works with the following new features.

- Not only the mean but also covariance matrix of stochastic multi-class OD demands is estimated. Particularly, the estimated covariance of stochastic multi-class OD demands can statistically reflect the traffic demand interactions between vehicle classes under network uncertainty over the year. Meanwhile, the interactions of stochastic path and mode choices are explicitly considered in the proposed model, which may has great potential to help the transportation planner and/or traffic manager understand the complexity and randomness of the mode and path choice behaviors in congested network with uncertainty.
- Lasso method is firstly incorporated into the OD demand estimation problem for variable selection and shrinkage so as to address the exacerbated identifiability issue for estimation of stochastic multi-class OD demand covariance matrix.

An equivalent non-linear constrained optimization model is proposed and formulated for the multi-class stochastic OD demand estimation problem. A n-fold cross-validation procedure is adapted to determine the lasso parameter in the proposed model. A heuristic solution algorithm based on the penalty method is adapted to solve the proposed model. Numerical examples are shown below to demonstrate the applications of the proposed model and solution algorithm together with some insightful discussion on the problem concerned.

1.5. Paper orginzation

The rest of the paper is organized as follows. In Section 2, the model formulation for estimating the mean and covariance of multi-class OD demands is presented. Then, a heuristic solution algorithm is proposed in Section 3. Numerical examples are discussed in Section 4. Finally, conclusions and further studies are given in Section 5.

2. Model formulation

2.1. Notations and basic assumptions

The notations used throughout the paper are listed as follows unless otherwise specified. For notational consistency, the italic capital letters are used to denote random variables and the italic lower-case letters are used to denote deterministic variables throughout the paper.

Nomenclature	
Indices:	
a,a'	Link index, $a,a' \in \mathbf{A}$.
i,i'	Vehicle class index, $i,i' \in \mathbf{D}$.
k,k'	Path index, $k,k' \in \mathbf{K}_{rs}$.
rs,rs'	OD pair index, $rs,rs' \in \mathbf{R}$.
Sets:	
$\mathbf{A}$	Set of links in the network.
$\tilde{\mathbf{A}}$	Observed link set, which is a subset of link set $\mathbf{A}$. The traffic flows on link $a \in \tilde{\mathbf{A}}$ can be observed by the traffic sensor during the observed time period.
$\mathbf{D}$	The set of vehicle classes, $i \in \mathbf{D}$.
$\mathbf{G}=(\mathbf{N},\mathbf{A})$	A road network, with $\mathbf{N}$ being the set of nodes and $\mathbf{A}$ being the set of links, respectively.
$\mathbf{K}$	Total path set of the network, $\mathbf{K}=\bigcup_{rs\in\mathbf{R}}\mathbf{K}_{rs}$.
$\mathbf{K}_{rs}$	Path set between OD pair rs.
$\mathbf{H}$	Sample of $V_{a,i}$, $\mathbf{H}=\{\tilde{v}^{(1)}_{a,i,o},\tilde{v}^{(2)}_{a,i,o},\cdots,\tilde{v}^{(h)}_{a,i,o}\}$.
$\mathbf{H}_n$	A subset of $\mathbf{H}$. $\mathbf{H}$ is randomly divided into 10 subsets with the same sample size $\frac{1}{10}h$, $\mathbf{H}=\bigcup_{n=1}^{10}\mathbf{H}_n$.
$\mathbf{R}$	Set of OD pairs. $\mathbf{R}$ is a subset of $\mathbf{N}\times\mathbf{N}$.
Variables:	
$F^k_{rs,i}$	Random traffic flow of vehicle class $i \in \mathbf{D}$ on path (or route) $k \in \mathbf{K}_{rs}$ between OD $rs \in \mathbf{R}$.
$\mathbf{F}$	$\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert$-vector of random multi-class path flows $(\cdots,F^k_{rs,i},\cdots)^T$ for all $k \in \mathbf{K}_{rs}$, $rs \in \mathbf{R}$ and $i \in \mathbf{D}$.
$f^k_{rs,i}$	Mean traffic flow of vehicle class $i \in \mathbf{D}$ on path $k \in \mathbf{K}_{rs}$ between OD $rs \in \mathbf{R}$, $f^k_{rs,i}=E[F^k_{rs,i}]$.
$\mathbf{f}$	$\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert$-vector of mean multi-class path flows $(\cdots,f^k_{rs,i},\cdots)^T$ for all $k \in \mathbf{K}_{rs}$, $rs \in \mathbf{R}$ and $i \in \mathbf{D}$.
$P^k_{rs,i}$	Random mode-path choice proportion of the traffic flow on path $k \in \mathbf{K}_{rs}$ of vehicle class i between OD pair rs.
$p^k_{rs,i}$	Mean path choice proportion of the traffic flow on path $k \in \mathbf{K}_{rs}$ of vehicle class i between OD pair rs, $p^k_{rs,i}=E[P^k_{rs,i}]$.
$\Sigma^{\mathbf{P}}$	$\Sigma^{\mathbf{P}}=\left\{\sigma^{p,k,k'}_{rs,i,rs',i'}\right\}_{\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert\times\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert}=\left\{\mathrm{cov}[P^k_{rs,i},P^{k'}_{rs',i'}]\right\}_{\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert\times\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert}$ is the covariance matrix of mode-path choice proportion.
$Q_{rs,i}$	Random traffic demand of vehicle class i between OD pair rs.
$\mathbf{Q}$	$\lvert\mathbf{R}\rvert\lvert\mathbf{D}\rvert$-vector of $(\cdots,Q_{rs,i},\cdots)^T$ for all $rs \in \mathbf{R}$ and $i \in \mathbf{D}$.
$q_{rs,i}$	Mean OD demand between OD pair rs of vehicle type i, $E\lfloor Q_{rs,i}\rfloor = q_{rs,i}$.

$\mathbf{q}$	$\lvert\mathbf{R}\rvert\lvert\mathbf{D}\rvert$-vector of $(\cdots,q_{rs,i},\cdots)^T$ for all $rs\in\mathbf{R}$ and $i\in\mathbf{D}$, $\mathbf{q}=E[\mathbf{Q}]$.
$\mathbf{q}^-$	The lower boundary of the estimated mean OD demand.
$\mathbf{q}^+$	The upper boundary of the estimated mean OD demand.
$\hat{q}_{rs,i}$	Actual mean OD demand of vehicle class i between OD pair rs.
t	Lasso parameter.
$V_{a,i}$	Random traffic flow on link a of vehicle class i.
$v_{a,i}$	Mean traffic flow on link a, $v_{a,i}=E[V_{a,i}]$.
$\tilde{v}_{a,i,o}^{(l)}$	Observed traffic flow of vehicle class i on link $a\in\tilde{\mathbf{A}}$ during the observed time period on day l $(l=1,2,\cdots,h)$.
$\tilde{v}_{a,i,o}^{\mathbf{H}}$	Sample mean of multi-class link flows, which is calculated using sample $\mathbf{H}$.
$\tilde{v}_{a,i,o}^{\mathbf{H}_n}$	Sample mean of multi-class link flows, which is calculated using sample $\mathbf{H}_n$.
$\tilde{\mathbf{v}}_o^{(l)}$	$\lvert\tilde{\mathbf{A}}\rvert\lvert\mathbf{D}\rvert$-vector of $(\cdots,\tilde{v}_{a,i,o}^{(l)},\cdots)^T$ for $a\in\tilde{\mathbf{A}}$ and $i\in\mathbf{D}$ on day l.
$\tilde{\mathbf{v}}_o$	Sample mean of observed multi-class link flows , $\tilde{\mathbf{v}}_o=(\cdots,\tilde{v}_{a,i,o},\cdots)^T=\dfrac{1}{h}\sum_{l=1}^{h}\tilde{\mathbf{v}}_o^{(l)}$.
$\lambda_{\min}$	Minimal eigenvalue of $\Sigma^{\mathbf{q}}$.
$\Sigma^{\mathbf{f}}$	Covariance matrix of multi-class path flows, $\Sigma^{\mathbf{f}}=\left\{\sigma_{rs,i,rs',i'}^{f,k,k'}\right\}_{\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert\times\lvert\mathbf{K}\rvert\lvert\mathbf{D}\rvert}$.
$\Sigma^{\mathbf{q}}$	Covariance matrix of traffic demands of all OD pairs for all vehicle classes, $\Sigma^{\mathbf{q}}=\left\{\sigma_{rs,i,rs',i'}^{q}\right\}_{\lvert\mathbf{R}\rvert\lvert\mathbf{D}\rvert\times\lvert\mathbf{R}\rvert\lvert\mathbf{D}\rvert}$.
$\tilde{\mathbf{\Sigma}}_o^{\mathbf{v}}$	Sample covariance matrix of observed multi-class link flows, $\tilde{\mathbf{\Sigma}}_o^{\mathbf{v}}=\left\{\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}}\right\}_{\lvert\tilde{\mathbf{A}}\rvert\lvert\mathbf{D}\rvert\times\lvert\tilde{\mathbf{A}}\rvert\lvert\mathbf{D}\rvert}$.
$\sigma_{rs,i,rs',i'}^{f,k,k'}$	Covariance between $F_{rs,i}^{k}$ and $F_{rs',i'}^{k'}$ $\sigma_{rs,i,rs',i'}^{f,k,k'}=\mathrm{cov}[F_{rs,i}^{k},F_{rs',i'}^{k'}]$.
$\sigma_{rs,i,rs',i'}^{q}$	Covariance between OD demand $Q_{rs,i}$ and $Q_{rs',i'}$, $\sigma_{rs,i,rs',i'}^{q}=\mathrm{cov}\left[Q_{rs,i},Q_{rs',i'}\right]$.
$\sigma_{a,i,a',i'}^{v}$	Covariance between link flows $V_{a,i}$ and $V_{a',i'}$, $\sigma_{a,i,a',i'}^{v}=\mathrm{cov}\left[V_{a,i},V_{a',i'}\right]$.
$\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}}$	Sample covariance of multi-class link flows, which is calculated using sample $\mathbf{H}$.
$\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}_n}$	Sample covariance of multi-class link flows, which is calculated using sample $\mathbf{H}_n$.
Parameters:	
m	A large integer used in cross-validation. In this paper, it is assumed that $m=100$.
u	Iteration number.
$\delta_{rs}^{k,a}$	Element of link-path incidence matrix Δ.
$\mu_1^{(u)},\cdots,\mu_5^{(u)}$	Four positive penalty coefficients at iteration u.
ξ	Enlarge parameter in the adapted penalty function solution algorithm, $\xi>1$.
τ	Stopping tolerance for the proposed solution algorithm.

To facilitate the presentation of the essential ideas without loss of generality, the following basic assumptions are made in this paper.

A1. It is assumed that the daily total OD demand of all vehicle classes in the same OD pair are identical across all working days. That is to say the weekend days and public holidays are excluded in this paper. The randomness or stochasticity of observed day-to-day multi-class traffic counts are only resulted from the random individual path and mode choices of the travelers. To explicitly model the interactions between travel modes (in terms of vehicle classes) as well as path choices, other sources of randomness resulting in randomness of multi-class traffics are ignored in this paper. These sources include the day-to-day variations of the total OD demand, travelers' departure times, measurement errors and so on.

A2. Following assumption A1, it is assumed that the multi-class traffic demands between different OD pairs are independent. This assumption would be reasonable. For example, the higher usage of private may only result in lower usage of taxi of the same OD pair. And it may not influence usage of any traffic mode of other OD pairs. It should be noted that the proposed model could still work without this assumption. And this assumption is made for simplicity.

A3. To facilitate the presentation of essential idea, the mode-path choice proportion is defined as $P_{rs,i}^{k}$, which is a random variable to account for the stochastic mode choice and path choice. The mean of mode-path choice proportion is assumed to be known and fixed, as such information could be available from some probe-vehicle data (e.g. taxi GPS data). The covariance of between different $P_{rs,i}^{k}$ is set as the decision variable in this paper.

2.2. Covariance of stochastic multi-class OD demands

According to assumption A1, the daily total traffic demands of all vehicle classes during the same hourly period (e.g. morning peak, 8:00 am - 9:00 am) between OD pair rs is denoted as q_{rs}, which is identical every day. q_{rs} is the summation of traffic demands of all vehicle classes as follows.

$$q_{rs} = \sum_{i} Q_{rs,i} \quad \forall \; rs \in \mathbf{R} \tag{1}$$

where $Q_{rs,i}$ the stochastic traffic demand of vehicle class i, and $E[Q_{rs,i}] = q_{rs,i}$. It should be noted that although $Q_{rs,i}$ is random variable its summation q_{rs} is deterministic variable. For convenience, denote $\mathbf{Q}$ and $\mathbf{q}$ as the $|\mathbf{R}||\mathbf{D}|$-vectors of $(\cdots, Q_{rs,i}, \cdots)^T$ and $(\cdots, q_{rs,i}, \cdots)^T$ for all $rs \in \mathbf{R}$ and $i \in \mathbf{D}$, respectively. The covariance between OD demand $Q_{rs,i}$ and $Q_{rs',i'}$ is denoted as:

$$\sigma_{rs,i,rs',i'}^{q} = \operatorname{cov}\left[Q_{rs,i}, Q_{rs',i'}\right] \; \forall rs, rs' \in \mathbf{R} \;\; , \; i, i' \in \mathbf{D} \tag{2}$$

The corresponding covariance matrix of stochastic multi-class OD demand can be expressed as:

$$\Sigma^{\mathbf{q}} = \left\{\sigma_{rs,i,rs',i'}^{q}\right\}_{|\mathbf{R}||\mathbf{D}|\times|\mathbf{R}||\mathbf{D}|} \tag{3}$$

Remark 1: It should be pointed out that although the traffic demand of each vehicle class is sotchastic, the summation total OD demand of all vehicle classes is deterministic and fixed. Such feature differs the proposed model from the that of Shao et al. (2014). In their paper, the total OD demands is defined as random variables. The spatial covariance of OD demands between different OD pairs is to be estimated without consideration of stochastic traffic demand iteractions between vehicle classes.

Remark 2: In line with assumption A2, it follows that:

$$\sigma_{rs,i,rs',i'}^{q} = 0 \, , \; \text{if} \;\; rs \neq rs' \tag{4}$$

This is because the assumption that multi-class traffic demands of different OD pairs are independent with each other. Thus, $\Sigma^{\mathbf{q}}$ is a block diagnal matrix.

Remark 3: The covariance matrix $\Sigma^{\mathbf{q}}$ describes the interactions of traffic demand between different vehicle classes of the same OD pair. It is generally believed that the higher usage of the one class of vehicle may not accompany with higher usage of the other classes of vehicle for the same OD pair. Thus, the value of $\sigma_{rs,i,rs,i'}^{q}$ can be zero or negative but not positive. It should be noted that this assumption is made for illustrative purpose in this paper, emperical studies need to be carried out to support this assumption.

Remark 4: The covariance matrix $\Sigma^{\mathbf{q}}$ has a special structural character. In view of assumption A1 and Equation (1), the variance of the total OD demand of all vehicle classes for each OD pair is zero. Then, the following condition should hold for $\Sigma^{\mathbf{q}}$.

$$\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sigma^{q}_{rs,rs,i,i'}=0\,,\ \ \forall rs\in\mathbf{R} \tag{5}$$

2.3. Conservation conditions between traffic flows in terms of mean and covariance

Let $F^k_{rs,i}$ be the random traffic flow on path $k\in\mathbf{K}_{rs}$ of vehicle class i with its mean $f^k_{rs,i}=E[F^k_{rs,i}]$. For convenience, $\mathbf{F}$ and $\mathbf{f}$ are denoted as the $|\mathbf{K}||\mathbf{D}|$-vectors of $(\cdots,F^k_{rs,i},\cdots)^T$ and $(\cdots,f^k_{rs,i},\cdots)^T$ for all $k\in\mathbf{K}_{rs}$, $rs\in\mathbf{R}$ and $i\in\mathbf{D}$, respectively. The path flows and OD demands satisfy the following flow conservation condition:

$$\mathbf{Q}=\Lambda\mathbf{F} \tag{6}$$

where Λ is the multi-class OD-path incidence matrix. Then, it follows that:

$$\mathbf{q}=E[\mathbf{Q}]=E[\Lambda\mathbf{F}]=\Lambda E[\mathbf{F}]=\Lambda\mathbf{f} \tag{7}$$

Equation (7) can be rewritten as:

$$q_{rs,i}=\sum_{k\in\mathbf{K}_{rs}}f^k_{rs,i}\quad \forall rs\in\mathbf{R}\,,i\in\mathbf{D} \tag{8}$$

According to assumption A3, the path flow of vehicle class i is a product of the corresponding mode-path choice proportion and the OD demand as follows.

$$F^k_{rs,i}=P^k_{rs,i}q_{rs}\quad \forall k\in\mathbf{K}_{rs},rs\in\mathbf{R}\,,i\in\mathbf{D} \tag{9}$$

where $P^k_{rs,i}$ is the random mode-path choice proportion of the traffic flow on path $k\in\mathbf{K}_{rs}$ of vehicle class i. Then, it follows from Equation (9) and assumption A3 that:

$$f^k_{rs,i}=p^k_{rs,i}q_{rs}\ \forall k\in\mathbf{K}_{rs},rs\in\mathbf{R}\,,i\in\mathbf{D} \tag{10}$$

According to assumption A3, the covariance between $F^k_{rs,i}$ and $F^{k'}_{rs',i'}$ can be deduced as:

$$\sigma^{f,k,k'}_{rs,i,rs',i'}=\mathrm{cov}[F^k_{rs,i},F^{k'}_{rs',i'}]=\mathrm{cov}[P^k_{rs,i}q_{rs},P^{k'}_{rs',i'}q_{rs'}]=q_{rs}q_{rs'}\,\mathrm{cov}[P^k_{rs,i},P^{k'}_{rs',i'}]$$
$$\forall k\in\mathbf{K}_{rs},k'\in\mathbf{K}_{rs'},rs,rs'\in\mathbf{R}\,,i,i'\in\mathbf{D} \tag{11}$$

The corresponding covariance matrix of multi-class path flows can be expressed as:

$$\Sigma^{\mathbf{f}}=\left\{\sigma^{f,k,k'}_{rs,i,rs',i'}\right\}_{|\mathbf{K}||\mathbf{D}|\times|\mathbf{K}||\mathbf{D}|} \tag{12}$$

According to Equation (6), the covariance conservation condition between multi-class path flows and multi-class OD demands is expressed as:

$$\sigma^{q}_{rs,rs',i,i'}=\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}\sigma^{f,k,k'}_{rs,i,rs',i'}=\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}q_{rs}q_{rs'}\,\mathrm{cov}[P^k_{rs,i},P^{k'}_{rs',i'}]\ \ \forall rs,rs'\in\mathbf{R}\ \ ,\ i,i'\in\mathbf{D} \tag{13}$$

Denote $\delta^{k,a}_{rs}$ as the element of link-path incidence matrix Δ. If path k uses link a, $\delta^{k,a}_{rs}=1$. Otherwise, $\delta^{k,a}_{rs}=0$. Then, the conservation condition of the estimated link and path flows for vehicle class i is expressed as:

$$V_{a,i}=\sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}_{rs}}\delta^{k,a}_{rs}F^k_{rs,i}\quad \forall\ a\in\mathbf{A}\,,i\in\mathbf{D} \tag{14}$$

where $V_{a,i}$ is the random traffic flow on link a of vehicle class i. The mean link flow is denoted as $v_{a,i}=E[V_{a,i}]$. It follows from Equation (14) that:

$$v_{a,i}=\sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}_{rs}}\delta^{k,a}_{rs}f^k_{rs,i}=\sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}_{rs}}\delta^{k,a}_{rs}p^k_{rs,i}q_{rs}=\sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}_{rs}}\delta^{k,a}_{rs}p^k_{rs,i}\left(\sum_{i'\in\mathbf{D}}q_{rs,i'}\right)\ \forall\ a\in\mathbf{A}\,,i\in\mathbf{D} \tag{15}$$

where $p^k_{rs,i}=E[P^k_{rs,i}]$, which is assumed to be known in this paper. It can be seen from Equation (15) that the mean of multi-class link flows is a linear function with respect to the mean of multi-class OD demands, which can be expressed as follows.

$$v_{a,i}=v_{a,i}(\mathbf{q})\ \ \forall\ a\in\mathbf{A}\,,i\in\mathbf{D} \tag{16}$$

Also, the conservation condition of the estimated multi-class link and path flow covariances can be obtained as:

$$\begin{aligned}\sigma^{v}_{a,i,a',i'} &= \mathrm{cov}\left[V_{a,i}, V_{a',i'}\right] = \mathrm{cov}\left[\sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}}\delta^{k,a}_{rs}F^{k}_{rs,i}, \sum_{rs'\in\mathbf{R}}\sum_{k'\in\mathbf{K}}\delta^{k',a'}_{rs'}F^{k'}_{rs',i'}\right] \\ &= \sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}}\sum_{rs'\in\mathbf{R}}\sum_{k'\in\mathbf{K}}\delta^{k,a}_{rs}\delta^{k',a'}_{rs'}\,\mathrm{cov}[F^{k}_{rs,i}, F^{k'}_{rs',i'}] = \sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}}\sum_{rs'\in\mathbf{R}}\sum_{k'\in\mathbf{K}}\delta^{k,a}_{rs}\delta^{k',a'}_{rs'}\sigma^{f,k,k'}_{rs,i,rs',i'} \\ &= \sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}}\sum_{rs'\in\mathbf{R}}\sum_{k'\in\mathbf{K}}\delta^{k,a}_{rs}\delta^{k',a'}_{rs'}q_{rs}q_{rs'}\,\mathrm{cov}[P^{k}_{rs,i}, P^{k'}_{rs',i'}]\ \forall\ a,a'\in\mathbf{A}\ i,i'\in\mathbf{D} \\ &= \sum_{rs\in\mathbf{R}}\sum_{k\in\mathbf{K}}\sum_{rs'\in\mathbf{R}}\sum_{k'\in\mathbf{K}}\delta^{k,a}_{rs}\delta^{k',a'}_{rs'}\left(\sum_{j\in\mathbf{D}}q_{rs,j}\right)\left(\sum_{j'\in\mathbf{D}}q_{rs',j'}\right)\mathrm{cov}[P^{k}_{rs,i}, P^{k'}_{rs',i'}]\ \ \forall\ a,a'\in\mathbf{A}\ i,i'\in\mathbf{D}\end{aligned} \tag{17}$$

where $\sigma^{v}_{a,i,a',i'}$ is the covariance between link flows $V_{a,i}$ and $V_{a',i'}$, $a,a'\in\mathbf{A}$ and $i,i'\in\mathbf{D}$. It can be seen from Equation (17) that the covariance of multi-class link flows is a linear function with respect to the covariance matrix of multi-class OD demands, which is expressed as follows.

$$\sigma^{v}_{a,i,a',i'} = \sigma^{v}_{a,i,a',i'}(\mathbf{q},\Sigma^{\mathbf{p}})\ \forall\ a,a'\in\mathbf{A}\ i,i'\in\mathbf{D} \tag{18}$$

where $\Sigma^{\mathbf{p}} = \left\{\sigma^{p,k,k'}_{rs,i,rs',i'}\right\}_{|\mathbf{K}||\mathbf{D}|\times|\mathbf{K}||\mathbf{D}|} = \left\{\mathrm{cov}[P^{k}_{rs,i}, P^{k'}_{rs',i'}]\right\}_{|\mathbf{K}||\mathbf{D}|\times|\mathbf{K}||\mathbf{D}|}$ is the covariance matrix of mode-path choice proportion, which is a decision variable in the proposed model.

2.4. Observed classified traffic counts

Denote $\tilde{\mathbf{A}}$ as a subset of link set $\mathbf{A}$. The traffic flows on link $a\in\tilde{\mathbf{A}}$ can be observed by the traffic sensor during the observed time period. The link with (without) traffic sensor is called "observed link" ("unobserved link") throughout this paper. Due to daily demand fluctuations, link flows of vehicle class i during the observed time periods vary from day to day. $\tilde{v}^{(l)}_{a,i,o}$ is denoted as the observed traffic flow of vehicle class i on link $a\in\tilde{\mathbf{A}}$ during the observed time period on day l $(l=1,2,\cdots,h)$. Then, $\mathbf{H}=\{\tilde{v}^{(1)}_{a,i,o},\tilde{v}^{(2)}_{a,i,o},\cdots,\tilde{v}^{(h)}_{a,i,o}\}$ is a sample of $V_{a,i}$ with sample size $h(h>1)$. For convenience, $\tilde{\mathbf{v}}^{(l)}_{o}$ is denoted as the $|\tilde{\mathbf{A}}||\mathbf{D}|$-vector of $(\cdots,\tilde{v}^{(l)}_{a,i,o},\cdots)^{T}$ for all $a\in\tilde{\mathbf{A}}$ and $i\in\mathbf{D}$ on day l . Then, the sample mean of the observed multi-class link flows using sample $\mathbf{H}$ can be calculated as:

$$\tilde{\mathbf{v}}_{o} = (\cdots,\tilde{v}^{\mathbf{H}}_{a,i,o},\cdots)^{T} = \frac{1}{h}\sum_{l=1}^{h}\tilde{\mathbf{v}}^{(l)}_{o} \tag{19}$$

The sample covariance matrix of the observed multi-class link flows using sample $\mathbf{H}$ can be calculated as:

$$\tilde{\Sigma}^{\mathbf{v}}_{o} = \left\{\tilde{\sigma}^{v,\mathbf{H}}_{a,i,a',i',o}\right\}_{|\tilde{\mathbf{A}}||\mathbf{D}|\times|\tilde{\mathbf{A}}||\mathbf{D}|} = \frac{1}{h-1}\sum_{l=1}^{h}\left\{\left(\tilde{\mathbf{v}}^{(l)}_{o}-\tilde{\mathbf{v}}_{o}\right)\left(\tilde{\mathbf{v}}^{(l)}_{o}-\tilde{\mathbf{v}}_{o}\right)^{T}\right\} \tag{20}$$

where $\tilde{\sigma}^{v,\mathbf{H}}_{a,i,a',i',o}$ is the sample covariance between observed traffic flows $V_{a,i}$ and $V_{a',i'}$, $a,a'\in\mathbf{A}$ and $i,i'\in\mathbf{D}$.

2.5. Lasso method for multi-class stochastic OD demand estimation model

The lasso method for the estimation of mean and covariance of multi-class OD demands is formulated as follows.

$$\min_{\mathbf{q},\Sigma^{\mathbf{p}}}\left\{z = \sum_{a\in\tilde{\mathbf{A}}}\sum_{i\in\mathbf{D}}\left(\tilde{v}^{\mathbf{H}}_{a,i,o} - v_{a,i}(\mathbf{q})\right)^{2} + \sum_{a\in\tilde{\mathbf{A}}}\sum_{a'\in\tilde{\mathbf{A}}}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left(\tilde{\sigma}^{v,\mathbf{H}}_{a,i,a',i',o} - \sigma^{v}_{a,i,a',i'}(\mathbf{q},\Sigma^{\mathbf{p}})\right)^{2}\right\} \tag{21a}$$

s.t.

$$\sum_{rs\in\mathbf{R}}\ \sum_{rs'\in\mathbf{R},rs\neq rs'}\ \sum_{k\in\mathbf{K}_{rs}}\ \sum_{k\in\mathbf{K}_{rs'}}\ \sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left|\sigma^{p,k,k'}_{rs,i,rs',i'}\right| \leq t \tag{21b}$$

$$\mathbf{q}\geq\mathbf{q}^{-} \tag{21c}$$

$$\mathbf{q}\leq\mathbf{q}^{+} \tag{21d}$$

$$\Sigma^{\mathbf{p}}\succeq 0 \tag{21e}$$

$$\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sigma^{q}_{rs,rs,i,i'}=\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}\left(\sum_{j\in\mathbf{D}}q_{rs,j}\right)^{2}\sigma^{p,k,k'}_{rs,i,rs',i'}=0\,,\quad\forall rs\in\mathbf{R}\tag{21f}$$

where t is the lasso parameter, which can be determined by a n-fold cross-validation procedure. The lower and upper boundaries are attached to restrict the mean OD demands within an interval of [$\mathbf{q}^{-}$, $\mathbf{q}^{+}$]. The non-negative lower and upper boundaries of the mean OD demands, $\mathbf{q}^{-}$ and $\mathbf{q}^{+}$, can be determined by prior or historical mean OD demands. The boundary constraints mean that resultant mean multi-class OD demands should not be significantly changed in comparison with the prior one. $\Sigma^{\mathbf{p}}\succeq 0$ represents that the covariance matrix of $\Sigma^{\mathbf{p}}$ is a symmetric and positive semi-definite matrix. It can be seen that the optimization problem (21) is an extended model formulation of the conventional least squares (LS) method. The objective function is to minimize the squared difference between observed and estimated means and covariance matrices of link flows (Hazelton, 2003). The purpose of Equation (21a) is to make the estimated mean and covariance of link flows match (or approximate) the observed ones as closely as possible. It should be pointed out that Equation (21a) is basically suited for normally distributed variables that are widely used in the literature. For other distributions (especially the skewed distribution) the objective function (21a) may need to be revised.

As mentioned above, there are too many elements in the covariance matrix to be estimated. However, the mode-path choice proportion covariance matrix has a special character. Some of the elements (say the covariances) in this matrix are actually zero. For example, it is assumed that the choices of private cars and goods vehicles are independent with each other, which means the corresponding covariance is zero. It should be noted that the lasso method is just applicable to the concerned covariance matrix estimation problem as a number of the covariances (the elements of the covariance matrix) are zero or close to zero.

Due to the identifiability difficulty, the actual value (i.e. zero) of the zero covariance may not be obtained by the conventional OD estimation models (e.g. LS model). The outstanding feature of lasso method is exactly capable of overcoming this difficulty. The lasso method does variable selection and shrinkage by employing the L1-norm constraint in the conventional LS model. It causes shrinkage of estimated coefficients towards zero (Tibshiranti, 1996). As such, lasso method shrinks some coefficients and set others to zero. Regarding the multi-class OD demand covariance estimation herein, lasso method is adopted by the constraint (21b). The constraint on the sum of the absolute values of the parameters to be estimated (referred to as the L1-norm, constraint (21b)) has the effect of forcing some parameters to zero, making the remaining parameters more identifiable (although identifiability is not guaranteed). That is to say in constraint (21b) the covariances (excluding variances in the covariance matrix) are shrunk by the L1-norm constraint. It shrinks all the non-diagonal elements in the covariance matrix towards zero. The variances (diagonal elements in the covariance matrix) are not taken into account in the L1-norm constraint by setting $rs\neq rs'$ in Equation (21b). This is because the actual multi-class demand varies form day to day, which verifies that the corresponding variances are not zero. As a result, the variances should be not shrunk towards zero. An important issue of the lasso method is that the value of parameter t controls the amount of shrinkage that is applied to the estimates (i.e. the covariances). Thus, it is important to determine the appropriate value of t. The cross-validation procedure is adopted and discussed in the following section to determine the value of lasso parameter t.

2.6. Cross-validation for determination of lasso parameter

To illustrate how to determine the lasso parameter t, a n-fold cross-validation procedure is proposed in this paper, where n (>1) is an integer. For simplicity, it is assumed that $n=10$ (i.e. tenfold) in the cross-validation. The observed multi-class link traffic flow data set $\mathbf{H}$ is randomly divided into 10 subsets with the same sample size $\frac{1}{10}h$, $\mathbf{H}=\bigcup_{n=1}^{10}\mathbf{H}_n$. It is noted that for each subset $\mathbf{H}_n$ is also a sample of the multi-class link flows. The corresponding sample mean and sample covariance of multi-class link flow can be calculated, which are denoted as $\tilde{v}^{\mathbf{H}_n}_{a,i,o}$ and $\tilde{\sigma}^{v,\mathbf{H}_n}_{a,i,a',i',o}$, respectively. The lasso parameter t can be determined by the following procedure.

Step 0: Initialization.

Set $t=0$. Let m be a large integer (In is paper, it is assumed that $m=100$). The conventional LS estimates of $\mathbf{q}$ and $\Sigma^{\mathbf{p}}$ of can be obtained using the observed data set $\mathbf{H}$, which are denoted as follows.

$$(\mathbf{q}^{\mathbf{H}}, \Sigma^{\mathbf{p},\mathbf{H}}) = \arg \min_{\mathbf{q}^{\mathbf{H}}, \Sigma^{\mathbf{p},\mathbf{H}}} \left\{ \sum_{a \in \tilde{\mathbf{A}}} \sum_{i \in \mathbf{D}} \left(\tilde{v}_{a,i,o}^{\mathbf{H}} - v_{a,i}\left(\mathbf{q}^{\mathbf{H}}\right) \right)^2 + \sum_{a \in \tilde{\mathbf{A}}} \sum_{a' \in \tilde{\mathbf{A}}} \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \left(\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}} - \sigma_{a,i,a',i'}^{v}(\Sigma^{\mathbf{p},\mathbf{H}}) \right)^2 \right\} \tag{22a}$$

s.t.

$$\mathbf{q}^{\mathbf{H}} \geq \mathbf{q}^{-} \tag{22b}$$

$$\mathbf{q}^{\mathbf{H}} \leq \mathbf{q}^{+} \tag{22c}$$

$$\Sigma^{\mathbf{p},\mathbf{H}} \succeq 0 \tag{22d}$$

$$\sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \sigma_{rs,rs,i,i'}^{q,\mathbf{H}} = \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \sum_{k' \in \mathbf{K}_{rs'}} \sum_{k \in \mathbf{K}_{rs}} \left(\sum_{j \in \mathbf{D}} q_{rs,j} \right)^2 \sigma_{rs,i,rs',i'}^{p,k,k',\mathbf{H}} = 0, \quad \forall rs \in \mathbf{R} \tag{21f}$$

where $\mathbf{q}^{\mathbf{H}}$ is denoted as a vector $(\cdots, q_{rs,i}^{\mathbf{H}}, \cdots)^T$ all $rs \in \mathbf{R}$ and $i \in \mathbf{D}$, $\Sigma^{\mathbf{p},\mathbf{H}}$ is denoted as the matrix $\Sigma^{\mathbf{p},\mathbf{H}} = \left\{ \sigma_{rs,i,rs',i'}^{p,k,k',\mathbf{H}} \right\}_{|\mathbf{K}||\mathbf{D}| \times |\mathbf{K}||\mathbf{D}|}$. Denote $s = \sum_{rs \in \mathbf{R}} \sum_{rs' \in \mathbf{R}, rs \neq rs'} \sum_{k \in \mathbf{K}_{rs}} \sum_{k \in \mathbf{K}_{rs'}} \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \left| \sigma_{rs,i,rs',i'}^{p,k,k',\mathbf{H}} \right|$.

Step 1: Estimation errors for possible values of t.

For each $t_j = \frac{1}{m}s, \frac{2}{m}s, \cdots, \frac{m-1}{m}s, s$ (i.e. $t_j = j \times \Delta t$, where $\Delta t = \frac{s}{m}$), repeat Steps 1.1 and 1.2.

Step 1.1 For each $\mathbf{H}_n$, repeat Steps 1.1.1 and 1.1.2.

Step 1.1.1 Use data $\mathbf{H} - \mathbf{H}_n (\mathbf{H}_1, \mathbf{H}_2, \cdots \mathbf{H}_{n-1}, \mathbf{H}_{n+1}, \cdots, \mathbf{H}_{10})$, to solve the following optimization problem.

$$(\mathbf{q}^{\mathbf{H}-\mathbf{H}_n}, \Sigma^{\mathbf{p},\mathbf{H}-\mathbf{H}_n}) = \arg \min_{\mathbf{q}^{\mathbf{H}-\mathbf{H}_n}, \Sigma^{\mathbf{p},\mathbf{H}-\mathbf{H}_n}} \left\{ \sum_{a \in \tilde{\mathbf{A}}} \sum_{i \in \mathbf{D}} \left(\tilde{v}_{a,i,o}^{\mathbf{H}-\mathbf{H}_n} - v_{a,i}\left(\mathbf{q}^{\mathbf{H}-\mathbf{H}_n}\right) \right)^2 + \sum_{a \in \tilde{\mathbf{A}}} \sum_{a' \in \tilde{\mathbf{A}}} \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \left(\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}-\mathbf{H}_n} - \sigma_{a,i,a',i'}^{v}(\mathbf{q}^{\mathbf{H}-\mathbf{H}_n}, \Sigma^{\mathbf{p},\mathbf{H}-\mathbf{H}_n}) \right)^2 \right\} \tag{23a}$$

s.t.

$$\sum_{rs \in \mathbf{R}} \sum_{rs' \in \mathbf{R}, rs \neq rs'} \sum_{k \in \mathbf{K}_{rs}} \sum_{k \in \mathbf{K}_{rs'}} \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \left| \sigma_{rs,i,rs',i'}^{p,k,k',\mathbf{H}\text{-}\mathbf{H}_n} \right| \leq t_j \tag{23b}$$

$$\mathbf{q}^{\mathbf{H}-\mathbf{H}_n} \geq \mathbf{q}^{-} \tag{23c}$$

$$\mathbf{q}^{\mathbf{H}-\mathbf{H}_n} \leq \mathbf{q}^{+} \tag{23d}$$

$$\Sigma^{\mathbf{p},\mathbf{H}-\mathbf{H}_n} \succeq 0 \tag{23e}$$

$$\sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \sigma_{rs,rs,i,i'}^{q,\mathbf{H}-\mathbf{H}_n} = \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \sum_{k' \in \mathbf{K}_{rs'}} \sum_{k \in \mathbf{K}_{rs}} \left(\sum_{j \in \mathbf{D}} q_{rs,j} \right)^2 \sigma_{rs,i,rs',i'}^{p,k,k',\mathbf{H}-\mathbf{H}_n} = 0, \quad \forall rs \in \mathbf{R} \tag{23f}$$

Step 1.1.2 Use $\mathbf{H}_n$ to calculate the estimation error.

$$\sqrt{\sum_{a \in \tilde{\mathbf{A}}} \sum_{a' \in \tilde{\mathbf{A}}} \sum_{i \in \mathbf{D}} \sum_{i' \in \mathbf{D}} \left(\tilde{\sigma}_{a,i,a',i',o}^{v,\mathbf{H}_n} - \sigma_{a,i,a',i'}^{v}(\Sigma^{\mathbf{p},\mathbf{H}-\mathbf{H}_n}) \right)^2} = e_{\mathbf{H}_n}. \tag{24}$$

Step 1.2 Calculate the mean estimation error of the lasso parameter t_j.

$$e\left(t_j\right) = \frac{1}{10} \sum_{i=1}^{10} e_{\mathbf{H}_n} \tag{25}$$

Step 2: Determine t.

$$t = \arg \min_{j=1,2,\cdots,m} e\left(t_j\right) \tag{26}$$

At each iteration of the cross-validation procedure, 90% observed data is used for calibration while 10% observed data is used for validation. It can be seen from the above cross-validation procedure that the lasso parameter is determined to obtain a "best" shrinkage of the estimates with minimum estimation errors on the basis of the whole observed data. It should be noted that the determined value of lasso parameter t is strictly less than value of s (the L1-norm of covariance matrix by conventional LS method). $t < s$ will cause shrinkage of the estimated parameters of the conventional LS method towards zero, and some parameters may be exactly equal to zero (Tibshirani, 1996). In such a way, the covariance, which is actually zero, is shrunk towards zero or even driven to be zero.

Remark 5: The lasso parameter t is determined on the basis of the observed data, the values of t may be different for different set of the observed data even if for the same network.

Remark 6: On the one hand, the larger the value of m, the more accurate the lasso parameter t. On the other hand, the larger the value of m, the more the computational time. Thus, how to select the value of m remains interesting investigations in further study.

3. Solution algorithm

If the lasso parameter t is known, the mean and covariance of multi-class OD demands can be obtained by solving the constrained optimization problem (21). It should be noted that the problem (21) is a non-linear constrained optimization problem. The main difficulty for solving it comes from the non-linear constraint (21e). In the literature of optimization, the non-linear constrained optimization problem is usually transformed into an unconstrained one.

Penalty method is one of the most widely used approaches for transforming the constrained optimization problem into the unconstrained one. In this paper, the exterior penalty method is employed due to its simplicity and efficiency. In connection to the constraints (21b)-(21f), the following penalty terms are proposed to be added in the objective function:

$$\sum_{rs\in\mathbf{R}}\sum_{rs'\in\mathbf{R},rs\neq rs'}\sum_{k\in\mathbf{K}_{rs}}\sum_{k\in\mathbf{K}_{rs'}}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left|\sigma^{p,k,k'}_{rs,i,rs',i'}\right|\le t \Rightarrow \mu_1^{(u)}\left(\sum_{rs\in\mathbf{R}}\sum_{rs'\in\mathbf{R},rs\neq rs'}\sum_{k\in\mathbf{K}_{rs}}\sum_{k\in\mathbf{K}_{rs'}}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left|\sigma^{p,k,k'}_{rs,i,rs',i'}\right|-t\right)^2 \tag{27}$$

$$\mathbf{q}\ge\mathbf{q}^- \Rightarrow \mu_2^{(u)}\left\|\min\left(\mathbf{q}-\mathbf{q}^-,0\right)\right\|^2 \tag{28}$$

$$\mathbf{q}^+\ge\mathbf{q} \Rightarrow \mu_3^{(u)}\left\|\min\left(\mathbf{q}^+-\mathbf{q},0\right)\right\|^2 \tag{29}$$

$$\Sigma^{\mathbf{p}}\succeq 0 \Rightarrow \mu_4^{(u)}\left(\min\left(\lambda_{\min},0\right)\right)^2 \tag{30}$$

$$\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sigma^{q}_{rs,rs,i,i'}=\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}\left(\sum_{j\in\mathbf{D}}q_{rs,j}\right)^2\sigma^{p,k,k'}_{rs,i,rs',i'}=0 \Rightarrow \mu_5^{(u)}\left(\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}\left(\sum_{j\in\mathbf{D}}q_{rs,j}\right)^2\sigma^{p,k,k'}_{rs,i,rs',i'}\right)^2 \tag{31}$$

where $\mu_1^{(u)}$, $\mu_2^{(u)}$, $\mu_3^{(u)}$, $\mu_4^{(u)}$ and $\mu_5^{(u)}$ are the five positive penalty coefficients at iteration u ; $\lambda_{\min}$ is the minimal eigenvalue of $\Sigma^{\mathbf{p}}$. Equation (30) means that the minimal eigenvalue of $\Sigma^{\mathbf{p}}$ should be non-negative. It implies that all the eigenvalues of $\Sigma^{\mathbf{p}}$ are non-negative. Otherwise, a penalty is given in the objective function. Then, the penalized objective function for optimization problem (21) can be expressed as following.

$$\min_{\mathbf{q},\Sigma^{\mathbf{p}}}\left\{z=z_1(\mathbf{q},\Sigma^{\mathbf{p}})+z_2(\mathbf{q},\Sigma^{\mathbf{p}},\mu_1^{(u)},\mu_2^{(u)},\mu_3^{(u)},\mu_4^{(u)},\mu_5^{(u)})\right\} \tag{32}$$

where

$$z_1(\mathbf{q},\Sigma^{\mathbf{p}})$$

$$=\sum_{a\in\tilde{\mathbf{A}}}\sum_{i\in\mathbf{D}}\left(\tilde{v}^{\mathbf{H}}_{a,i,o}-v_{a,i}\left(\mathbf{q}\right)\right)^2+\sum_{a\in\tilde{\mathbf{A}}}\sum_{a'\in\tilde{\mathbf{A}}}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left(\tilde{\sigma}^{v,\mathbf{H}}_{a,i,a',i',o}-\sigma^{v}_{a,i,a',i'}(\mathbf{q},\Sigma^{\mathbf{p}})\right)^2 \tag{33}$$

$$z_2(\mathbf{q},\Sigma^{\mathbf{p}},\mu_1^{(u)},\mu_2^{(u)},\mu_3^{(u)},\mu_4^{(u)},\mu_5^{(u)})$$

$$=\mu_1^{(u)}\left(\sum_{rs\in\mathbf{R}}\sum_{rs'\in\mathbf{R},rs\neq rs'}\sum_{k\in\mathbf{K}_{rs}}\sum_{k\in\mathbf{K}_{rs'}}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\left|\sigma^{p,k,k'}_{rs,i,rs',i'}\right|-t\right)^2+\mu_2^{(u)}\left\|\min\left(\mathbf{q}-\mathbf{q}^-,0\right)\right\|^2+\mu_3^{(u)}\left\|\min\left(\mathbf{q}^+-\mathbf{q},0\right)\right\|^2$$

$$+\mu_4^{(u)}\left(\min\left(\lambda_{\min},0\right)\right)^2+\mu_5^{(u)}\left(\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sum_{k'\in\mathbf{K}_{rs'}}\sum_{k\in\mathbf{K}_{rs}}\left(\sum_{j\in\mathbf{D}}q_{rs,j}\right)^2\sigma_{rs,i,rs',i'}^{p,k,k'}\right)^2 \qquad (34)$$

Based on the exterior penalty method, the flowchart of the proposed solution algorithm is presented in Figure 2. It should be noted that the gradient of the objective function (32) is difficult to obtain due to the penalty term corresponding to the non-linear inequality constraint (Equation (30)). Therefore, some derivative-free optimization methods should be employed for solving the minimization problem in the upper level, such as the simplex search method (Lagarias et al., 1998) and generalized pattern search methods (Torczon, 1997; Audet and Dennis, 2003). Details on the derivative-free optimization methods can be referred to Conn et al. (2009). In this paper, the simplex search method (Lagarias et al., 1998) is used to solve the unconstrained optimization problem in Step 2, which is available in the Matlab optimization toolbox by the subroutine "fminsearch".

The convergence of the employed exterior penalty solution algorithm can be guaranteed by the following proposition.

Proposition 1: Suppose that each $\{\mathbf{q}^{(u+1)},\Sigma^{\mathbf{p},(u+1)}\}$ is the exact global minimizer of unconstrained minimization problem (32), then, every limit point $\{\mathbf{q}^{*},\Sigma^{\mathbf{p},*}\}$ of the sequence $\{\mathbf{q}^{(u+1)},\Sigma^{\mathbf{p},(u+1)}\}$ is a global solution of the constrained optimization problem (21) (Nocedal and Wright, 2006).

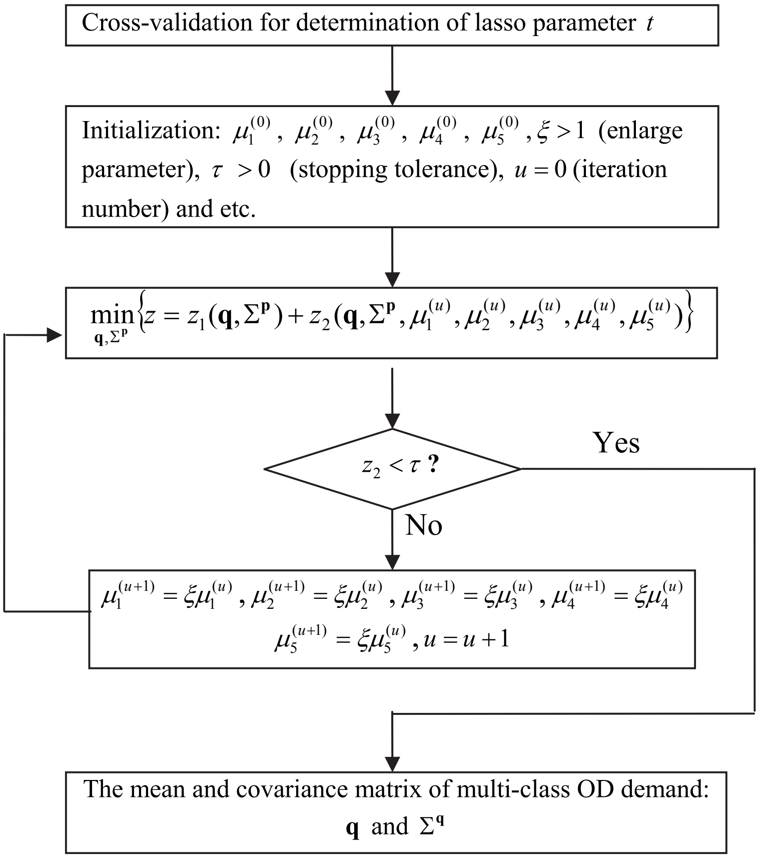

Cross-validation for determination of lasso parameter t
Initialization: $\mu_1^{(0)}$, $\mu_2^{(0)}$, $\mu_3^{(0)}$, $\mu_4^{(0)}$, $\mu_5^{(0)}$, $\xi>1$ (enlarge parameter), $\tau>0$ (stopping tolerance), $u=0$ (iteration number) and etc.
$\min_{\mathbf{q},\Sigma^{\mathbf{p}}}\left\{z=z_1(\mathbf{q},\Sigma^{\mathbf{p}})+z_2(\mathbf{q},\Sigma^{\mathbf{p}},\mu_1^{(u)},\mu_2^{(u)},\mu_3^{(u)},\mu_4^{(u)},\mu_5^{(u)})\right\}$
$z_2<\tau$?
Yes
No
$\mu_1^{(u+1)}=\xi\mu_1^{(u)}$, $\mu_2^{(u+1)}=\xi\mu_2^{(u)}$, $\mu_3^{(u+1)}=\xi\mu_3^{(u)}$, $\mu_4^{(u+1)}=\xi\mu_4^{(u)}$
$\mu_5^{(u+1)}=\xi\mu_5^{(u)}$, $u=u+1$
The mean and covariance matrix of multi-class OD demand:
$\mathbf{q}$ and $\Sigma^{\mathbf{q}}$

Fig. 2. The flowchart of the proposed solution algorithm

4. Numerical examples

4.1. Preliminary

To demonstrate the properties of the proposed model, a small tractable network is chosen. A simple network with six nodes, seven links, two OD pairs (1→3 and 2→4) and four paths, as shown in Figure 3, is used to illustrate the applications of the proposed stochastic multi-class OD demand estimation model. It is assumed that there are three classes of vehicles, i.e. private car (pc), taxi (tx) and goods vehicle (gv). That is to say, $\mathbf{D}=\{pc,tx,gv\}$. The corresponding pcu for these three vehicles are set as: 1 pc = 1 pcu, 1 tx = 1 pcu and 1 gv = 2.0 pcu. The mean path choice proportions are all set to be $1/6$, i.e. $p_{rs,i}^{k}=1/6$, $\forall k\in\mathbf{K}_{rs}, rs\in\mathbf{R}$, $i\in\mathbf{D}$. Suppose that the classified traffic counts during the same hourly period of weekdays have been collected over the whole year (about 300 days). Thus, the sample size is set to be h=300. The actual mean and covariance of the multi-class OD demands are given in Tables 2 and 3 respectively, while the observed link flows by vehicle class and their covariances are shown in Tables 4 and 5, respectively.

Table 2. The actual mean multi-class OD demands (pcu/hour)

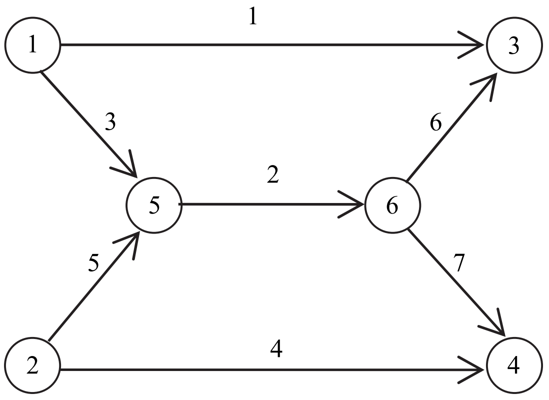
OD pair	Private car ($\hat{q}_{rs,pc}$)	Taxi ($\hat{q}_{rs,tx}$)	Goods vehicle ($\hat{q}_{rs,gv}$)
13	1200	1200	1200
24	1200	1200	1200

Table 3. The actual multi-class OD demand covariance matrix (pcu/hour)^2

$\text{cov}[Q_{rs,i},Q_{rs',i'}]$		OD Pair 1→3			OD Pair 2→4		
		Private car	Taxi	Goods vehicle	Private car	Taxi	Goods vehicle
OD Pair 1→3	Private car	216000	-216000	0	0	0	0
	Taxi	-216000	216000	0	0	0	0
	Goods vehicle	0	0	0	0	0	0
OD Pair 2→4	Private car	0	0	0	216000	-216000	0
	Taxi	0	0	0	-216000	216000	0
	Goods vehicle	0	0	0	0	0	0

Table 4. The observed classifed link flows (pcu/hour)

Link no	Private car ($\tilde{v}_{a,pc}$)	Taxi ($\tilde{v}_{a,tx}$)	Goods vehicle ($\tilde{v}_{a,gv}$)
1	600	600	600
2	600	600	600
4	1200	1200	1200



Path defined by sequence of links

Path 1: 1
Path 2: 3-2-6
Path 3: 5-2-7
Path 4: 4

Fig. 3. A small test network

Table 5. The observed classified link flow covariance matrix (pcu/hour)^2

$\text{cov}[\widetilde{V}_{a,i},\widetilde{V}_{a',i'}]$		Link 1			Link 2			Link 4		
		Private car	Taxi	Goods vehicle	Private car	Taxi	Goods vehicle	Private car	Taxi	Goods vehicle
Link 1	Private car	180000	-144000	0	-72000	36000	0	0	0	0
	Taxi	-144000	216000	0	36000	-108000	0	0	0	0
	Goods vehicle	0	0	72000	0	0	-72000	0	0	0
Link 2	Private car	-72000	36000	0	360000	-288000	0	-72000	36000	0
	Taxi	36000	-108000	0	-288000	432000	0	36000	-108000	0
	Goods vehicle	0	0	-72000	0	0	144000	0	0	-72000
Link 4	Private car	0	0	0	-72000	36000	0	180000	-144000	0
	Taxi	0	0	0	36000	-108000	0	-144000	216000	0
	Goods vehicle	0	0	0	0	0	-72000	0	0	72000

4.2. Stochastic interactions of mode and path choices

One of the outstanding features of the proposed model is capability of figuring out the interactions between vehicle classes by a statistical method. Specifically, the mode-path choice proportion $P^k_{rs,i}$ is used to capture the stochastic interactions of mode and path choices. The expected value of $P^k_{rs,i}$ is assumed to be known in this example. However, the covariance matrix of all $\text{cov}[P^k_{rs,i},P^{k'}_{rs',i'}]$ in the whole network is treated as decision variable. The resultant mode-path choice proportion covariance matrix for this example is shown in Table 6. The covariance information in Table 6 can reflect the stochastic interactions between vehicle classes as well as the path choice proportions. For example, the covariance between traffic flows of paths 1 and 2 using private car is negative, i.e. -0.0054, which means the more private cars use path 1 accompanies with less private cars use path 2. Such path choice behavior is reasonable. The reason is that paths 1 and 2 are two parallel paths of the same OD pair. The more amount of traffic flow of the same vehicle class on one path will result in the less amount on the other one. The stochastic interactions between vehicle classes can also be demonstrated by covariance information in Table 6. For path 1, the covariance between private car and taxi traffic flows is also negative, i.e. -0.0107. Such negative covariance means that for the same path the higher usage of private car leads to the lower usage of taxi. Meanwhile, for path 1, the covariance of private car and goods vehicle traffic flows is zero. This demonstrates that the traffic flows of private car and goods vehicle is not linearly dependent. They may be independent and not have interactions. In reality, the mode-path choice proportion covariance information provides an alternative simple way to capture the interactions between path choices and mode choices from a statistical viewpoint.

Table 6. The resultant mode-path choice proportion covariance matrix

$\text{cov}[P^k_{rs,i},P^{k'}_{rs',i'}]$			OD Pair 1→3						OD Pair 2→4					
			Private car		Taxi		Goods vehicle		Private car		Taxi		Goods vehicle	
			Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 3	Path 4	Path 3	Path 4	Path 3	Path 4
OD Pair 1→3	Private car	Path 1	0.0143	**-0.0054**	**-0.0107**	0.0029	**0**	0	0	0	0	0	0	0
		Path 2	-0.0054	0.0143	0.0029	-0.0107	0	0	0	0	0	0	0	0
	Taxi	Path 1	-0.0107	0.0029	0.0156	-0.0089	0	0	0	0	0	0	0	0
		Path 2	0.0029	-0.0107	-0.0089	0.0156	0	0	0	0	0	0	0	0
	Goods vehicle	Path 1	0	0	0	0	0.0056	-0.0056	0	0	0	0	0	0
		Path 2	0	0	0	0	-0.0056	0.0056	0	0	0	0	0	0
OD Pair 2→4	Private car	Path 3	0	0	0	0	0	0	0.0143	-0.0054	-0.0107	0.0029	0	0
		Path 4	0	0	0	0	0	0	-0.0054	0.0143	0.0029	-0.0107	0	0
	Taxi	Path 3	0	0	0	0	0	0	-0.0107	0.0029	0.0156	-0.0089	0	0
		Path 4	0	0	0	0	0	0	0.0029	-0.0107	-0.0089	0.0156	0	0
	Goods vehicle	Path 3	0	0	0	0	0	0	0	0	0	0	0.0056	-0.0056
		Path 4	0	0	0	0	0	0	0	0	0	0	-0.0056	0.0056

According to Equation (11), the corresponding multi-class path flows can be calculated easily, which is shown in Table 7. It should be noted that the multi-class OD demand covariance matrix can be calculated using the results in Table 7 according to Equation (13). For example, the summation of the four italic covariance numbers (-138341+37415+37415-138341) in Table 7 can obtain the OD demand covariance between private car and taxi (-201852) for OD pair 1→3. Such covariance information of stochastic multi-class OD demands can be used to evaluate the network performance with respect to different vehicle classes for network with uncertainty. Due to the length limitation of this paper, similar illustration can be found in section 4.1.3 in Shao et al. (2014).

Table 7. The resultant multi-class path flow covariance matrix (pcu/hour)^2

$\mathrm{cov}[F_{rs,i}^{k}, F_{rs',i'}^{k'}]$			OD Pair 1→3						OD Pair 2→4					
			Private car		Taxi		Goods vehicle		Private car		Taxi		Goods vehicle	
			Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 3	Path 4	Path 3	Path 4	Path 3	Path 4
OD Pair 1→3	Private car	Path 1	185178	-69929	***-138341***	***37415***	0	0	0	0	0	0	0	0
		Path 2	-69929	185178	***37415***	***-138341***	0	0	0	0	0	0	0	0
	Taxi	Path 1	-138341	37415	201722	-115345	0	0	0	0	0	0	0	0
		Path 2	37415	-138341	-115345	201722	0	0	0	0	0	0	0	0
	Goods vehicle	Path 1	0	0	0	0	72000	-72000	0	0	0	0	0	0
		Path 2	0	0	0	0	-72000	72000	0	0	0	0	0	0
OD Pair 2→4	Private car	Path 3	0	0	0	0	0	0	185178	-69929	-138341	37415	0	0
		Path 4	0	0	0	0	0	0	-69929	185178	37415	-138341	0	0
	Taxi	Path 3	0	0	0	0	0	0	-138341	37415	201722	-115345	0	0
		Path 4	0	0	0	0	0	0	37415	-138341	-115345	201722	0	0
	Goods vehicle	Path 3	0	0	0	0	0	0	0	0	0	0	72000	-72000
		Path 4	0	0	0	0	0	0	0	0	0	0	-72000	72000

Table 8. The resultant multi-class OD demand covariance matrix (pcu/hour)^2

$\mathrm{cov}[Q_{rs,i}, Q_{rs',i'}]$		OD Pair 1→3			OD Pair 2→4		
		Private car	Taxi	Goods vehicle	Private car	Taxi	Goods vehicle
OD Pair 1→3	Private car	230499	***-201852***	0	0	0	0
	Taxi	-201852	230499	0	0	0	0
	Goods vehicle	0	0	0	0	0	0
OD Pair 2→4	Private car	0	0	0	230499	-201852	0
	Taxi	0	0	0	-201852	230499	0
	Goods vehicle	0	0	0	0	0	0

4.3. Lasso vs LS methods

To test the performance of lasso method in comparison to the conventional LS methods, three scenarios are used and shown in Table 9. In Scenario A, the observed traffic data contains the “full information” of the multi-class OD demands. It is easy to verify that the traffic flow information of links 1, 2 and 4 covers all the needed information for estimation of multi-class OD demand. That is to say, the observed traffic data from links 1, 2 and 4 is sufficient to estimate the multi-class OD demands. Thus, Scenario A is called the “full information” scenario in this example. In Scenarios B and C, the observed traffic data only contains the “partial information” of the multi-class OD demands. In Scenario B, the observed traffic data only includes the data of private car and taxi but not the goods vehicles. In Scenario C, the observed traffic data only includes the traffic flow information for OD pair 13. In this example, root mean squared error (RMSE) is used to show the estimation errors for the three scenarios. The corresponding formulae of RMSEs for mean and covariance matrix of multi-class OD demands are given as below.

$$\mathrm{RMSE}_{\mathbf{q}} = \sqrt{\frac{1}{|\mathbf{R}||\mathbf{D}|}\sum_{i\in\mathbf{D}}\sum_{rs\in\mathbf{R}}\left(q_{rs,i}-\hat{q}_{rs,i}\right)^{2}} \tag{35}$$

$$\mathrm{RMSE}_{\Sigma\mathbf{q}} = \sqrt{\frac{1}{\left(|\mathbf{R}||\mathbf{D}|\right)^2}\sum_{i\in\mathbf{D}}\sum_{i'\in\mathbf{D}}\sum_{rs\in\mathbf{R}}\sum_{rs'\in\mathbf{R}}\left(\sigma_{rs,i,rs',i'} - \hat{\sigma}_{rs,i,rs',i'}\right)^2} \tag{36}$$

It can be seen from Table 9 that under full information condition (Scenario A) both lasso and LS methods have the same estimation results. As the RMSEs are small ($\mathrm{RMSE}_{\mathbf{q}}$ =0 (pcu/hour) and $\mathrm{RMSE}_{\Sigma\mathbf{q}}$ =6752 (pcu/hour)^2), the estimated mean and covariance matrix are closed to the actual ones. Thus, if the observed traffic data is sufficient, there is no difference between lasso and conventional LS method. However, if the traffic count data is insufficient (i.e. overfitting occurs), the lasso method could outperform the LS method, which can be evidenced by the results in Scenarios B and C. For instance, if the traffic data of good vehicles is unavailable, the lasso method could obtain a better estimation of multi-class OD demand covariance matrix than the conventional LS method. For lasso method in Scenario C, the estimated results are better than that of LS method. For example, in lasso method $\mathrm{RMSE}_{\Sigma\mathbf{q}}$ =28461 (pcu/hour)^2 which is smaller than in LS method ($\mathrm{RMSE}_{\Sigma\mathbf{q}}$ =45588 (pcu/hour)^2). In Scenario B, the performance of lasso method is still better than that of LS method. It is shown in this example, that the lasso method is an extension of the LS method. The estimation error of lasso method is smaller than that of LS method if the observed data is insufficient. That is to say the proposed lasso method is more suitable for handling the overfitting issue than the LS method. And the lasso method could outperform the conventional LS method on overcoming the identifiability difficulty for estimating the mean and covariance of multi-class OD demands.

Table 9. RMSEs of estimated mean and covariance matrix for multi-class OD demands

Scenario	Observed links	Descriptions	LS method		Lasso Method	
			(pcu/hour)	(pcu/hour)^2	(pcu/hour)	(pcu/hour)^2
A	1, 2, 4	Full information	**0** ($\mathrm{RMSE}_{\mathbf{q}}$)	**6752**($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)	**0**($\mathrm{RMSE}_{\mathbf{q}}$)	**6752**($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)
B	1, 2, 4 (only two vehicle classes)	Partial information	326($\mathrm{RMSE}_{\mathbf{q}}$)	9033($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)	251($\mathrm{RMSE}_{\mathbf{q}}$)	7389($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)
C	1,3 (only cover OD pair 1-3)	Partial information	936($\mathrm{RMSE}_{\mathbf{q}}$)	**45588** ($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)	758($\mathrm{RMSE}_{\mathbf{q}}$)	**28461** ($\mathrm{RMSE}_{\Sigma\mathbf{q}}$)

5. Conclusions and further studies

This paper proposed a new model for estimating the mean and covariance of stochastic multi-class OD demands based on the classified traffic counts for the same hourly period throughout the year. Different from the conventional OD demand estimation models, the proposed model utilized the statistical properties of the observed hourly traffic count data by vehicle class (or type) over the whole year. Also, the stochastic mode and path choices were explicitly considered in the proposed model using a random mode-path choice proportions. To overcome the identifiability difficulty in this paper, the lasso (least absolute shrinkage and selection operator) method was incorporated into the proposed model. A cross-validation procedure was proposed to determine the lasso parameter. An equivalent non-linear constrained optimization model was proposed and formulated as the corresponding OD demand estimation problem shown in this paper. A heuristic solution algorithm based on exterior penalty function method was adapted to solve the proposed model.

It was found in the numerical examples that (i) the estimated covariance of mode-path choice proportions can reflect the interactions of path and mode choices between different vehicle classes; (ii) the proposed model based on the lasso method could outperform the conventional least squares (LS) model on overcoming the identifiability difficulty for model application.

On the basis of the model proposed in this paper, some further extensions can be envisaged as follows.

- This paper only estimates the vehicle class covariance of stochastic multi-class traffic demands during the same hourly period over the year. Further investigation should be carried out on how to extend the proposed model to simultaneously estimate the spatial and temporal covariance of the multi-class traffic demands between different hourly periods in dynamic or time-dependent network models.
- Although the lasso method could overcome the identifiability difficulty to some extent in the proposed OD demand estimation model, multiple solutions may still exist. Also, the estimation errors still existed according the

results of the numerical example shown in this paper. How to address this issue in practice still reveals further investigations.

- As the mean of mode-path choice proportion is assumed to be known in this paper. How to relax this assumption by taking into account the traveler's path choice behavior in network with uncertainty would be worthwhile for further study. To this end, the bi-level modeling approach would be employed in the proposed model (Shao et al., 2013; Shao et al., 2014).
- The proposed modeling approach could also be modified to be applied in a multi-modal transportation network so as to consider the covariance between the other transport modes, such as metro and bus.
- The proposed modeling approach is tested on a hypothesis transportation network. The results may be very different for other test networks. Thus, it also leaves open the question of the scalability of the proposed method to realistic large scale networks. For such purpose, some real case studies need to be carried out using the proposed model in further studies.

Acknowledgements

The work described in this paper was jointly supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. 5196/10E and 5187/12E), Research Committee of The Hong Kong Polytechnic University (Project No. G-YJ93), National Natural Science Foundation of China (Project Nos. 71271205, 71171013 and 71371175), National Basic Research Program (973 Program) of China (Project No 2012CB725401), Qing Lan Project and the Fundamental Research Funds for the Central Universities (Project No. 2012QNA44).

References

Ashok, K., Ben-Akiva, M.E., 2000. Alternative approaches for real-time estimation and prediction of time-dependent Origin-Destination flows. Transportation Science, 34 (1), 21-36.

Ashok, K., Ben-Akiva, M.E., 2002. Estimation and prediction of time-dependent Origin-Destination flows with a stochastic mapping to path flows and link flows. Transportation Science, 36(2), 184-198.

Audet, C., Dennis, J.E., 2003. Analysis of Generalized Pattern Searches. SIAM Journal on Optimization, 13(3), 889-903.

Bell, M., 1991. The estimation of origin-destination matrices by constrained generalized least squares. Transportation Research Part B, 25(1), 13-22.

Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. Transportation Research Part B, 18(4-5), 289-299.

Castillo, E., Menendez, J.M., Jiménez, P., 2008a. Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. Transportation Research Part B, 42(5), 455-481.

Castillo, E., Menéndez, J.M., Sánchez-Cambronero, S., 2008b. Traffic estimation and optimal counting location without path enumeration using Bayesian networks. Computer Aided Civil and Infrastructure Engineering, 23(3), 189-207.

Chen, A., Ji, Z., Recker, W., 2002. Travel time reliability with risk sensitive travelers. Transportation Research Record, 1783, 27-33.

Chen, A., Zhou, Z., 2010. The α-reliable mean-excess traffic equilibrium model with stochastic travel times. Transportation Research Part B, 44(4), 493-513.

Chen, B.Y., Lam, W.H.K., Sumalee, A., Shao, H., 2011. An efficient solution algorithm for solving multi-class reliability-based traffic assignment problem. Mathematical and Computer Modelling, 54(5-6), 1428-1439.

Clark, S., Watling, D., 2005. Modeling network travel time reliability under stochastic demand. Transportation Research Part B, 39(2), 119-140.

Conn, A.R., Scheinberg, K., Vicente, L.N., 2009. Introduction to Derivative-Free Optimization. Philadelphia: Society for Industrial and Applied Mathematics.

Duthie, J.C., Unnikrishnan, A., Waller, S.T., 2011. Influence of demand uncertainty and correlations on traffic predictions and decisions. Computer-Aided Civil And Infrastructure Engineering, 26(1), 16-29.

Frank, C., 1978. A Study of Alternative Approaches to Combined Trip Distribution-Assignment Modeling. PhD thesis, Department of Regional Science, University of Pennsylvania, Philadelphia, PA.

Haas, C.N., 1999. On modeling correlated random variables in risk assessment. Risk Analysis, 19(6), 1205-1214.

Hazelton, M.L., 2003. Some comments on origin-destination matrix estimation. Transportation Research Part A, 37(10), 811-822.

Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM Journal of Optimization, 9(1), 112-147.

Lam, W.H.K., Huang, H.J., 1992a. A combined trip distribution and assignment model for multiple user classes. Transportation Research Part B, 26(4), 275-287.

Lam, W.H.K., Huang, H.J., 1992b. Calibration of the combined trip distribution and assignment model for multiple user classes. Transportation Research Part B, 26(4), 289-305.

Lam, W.H.K., Shao, H., Sumalee, A., 2008. Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply. Transportation Research Part B, 42(10), 890-910.

Li, B.B., 2009. Markov models for Bayesian analysis about transit route origin-destination matrices. Transportation Research Part B, 43(3), 301-310.

Maher, M., 1983. Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach. Transportation Research Part B, 20(6), 435-447.

Nakayama, S., Takayama, J., 2003. A traffic network equilibrium model for uncertain demands. Proceedings of the 82nd Transportation Research Board Annual Meeting, CD-ROM.

Nocedal, J., Wright, S.J., 2006. Numerical Optimization (2nd edition). New York: Springer.

Shao, H., Lam, W.H.K., Sumalee, A., Chen, A., 2013. Journey time estimator for assessment of road network performance under demand uncertainty. Transportation Research Part C, 35, 244-262.

Shao, H., Lam, W.H.K., Sumalee, A., Chen, A., Hazelton, M.L., 2014. Estimation of mean and covariance of peak hour origin–destination demands from day-to-day traffic counts. Transportation Research Part B, 68, 52-75.

Shao, H., Lam, W.H.K., Tam, M.L., 2006. A reliability-based stochastic traffic assignment model for network with multiple user classes under uncertainty in demand. Networks and Spatial Economics, 6(3-4), 173-204.

Spiess, H., 1987. A maximum likelihood model for estimating origin-destination matrices. Transportation Research Part B, 21(5), 395-412.

Sumalee, A.,Uchida, K., Lam, W.H.K., 2011. Stochastic multi-modal transport network under demand uncertainties and adverse weather condition. Transportation Research Part C, 19(2), 338-350.

Sun, S.L., Zhang, C.S., Yu, G.Q., 2006. A Bayesian network approach to traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems, 7(1), 124-132.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1), 267-288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society Series B, 73(Part 3), 273-282.

Tam, M.L., Lam, W.H.K. 2008. Using automatic vehicle identification data for travel time estimation in Hong Kong. Transportmetrica, 4(3), 179-194.

Torczon, V., 1997. On the convergence of pattern search algorithms. SIAM Journal on Optimization, 7(1), 1-25.

Transportation Research Board, 2000. Highway Capacity Manual. National Research Council, Washington, D.C., U.S.A.

Van Zuylen, H.J., Willumsen, L.G., 1980. The most likely trip matrix estimated from traffic counts. Transportation Research Part B, 14(3), 281-293.

Waller, S.T., Schofer, J.L., Ziliaskopoulos, A.K., 2001. Evaluation with traffic assignment under demand uncertainty. Transportation Research Record, 1771, 69-74.

Watling, D.P., 1994. Maximum-likelihood-estimation of an origin destination matrix from a partial registration plate survey. Transportation Research Part B, 28(4), 289-314.

Wong, S.C., Tong, C.O., Wong, K.I., Lam, W.H.K., Lo, H.K., Yang, H, Lo, H.P., 2005. Estimation of multiclass origin-destination matrices from traffic counts. Journal of Urban Planning and Development-ASCE, 131(1), 19-29.

Yau, K.K.W., Lo H.P., Fung S.H.H., 2006. Multiple-vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention, 38(6), 1157-1161.

Zhao, Y., Kockelman, K.M., 2002. The propagation of uncertainty through travel demand models: An exploratory analysis. Annals of Regional Science, 36(1), 145-163.