

1
2 **Mobile Sensors as Traffic Probes: Addressing Transportation Modeling and**
3 **Privacy Protection in an Integrated Framework**
4
5
6
7
8

9 **Xuegang (Jeff) Ban***

10 Department of Civil and Environmental Engineering
11 Rensselaer Polytechnic Institute (RPI)
12 110 Eighth Street, Room JEC 4034
13 Troy, NY 12180-3590
14 Phone: (518) 276-8043 Fax: (518) 276-4833
15 Email: banx@rpi.edu
16
17
18
19

20 **Marco Gruteser**

21 WINLAB / Electrical & Computer Engineering
22 Rutgers University
23 Tech Centre of New Jersey
24 671 Route 1 South
25 North Brunswick, NJ 08902
26 Phone: (732) 993-4561
27 Email: gruteser@winlab.rutgers.edu
28
29
30
31

32 **Submitted to the 7th International Conference on Traffic & Transportation Studies**

33
34 **January 10, 2010**
35
36
37
38
39

40 ** Corresponding Author*
41

ABSTRACT

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Mobile traffic sensor data, once widely available, will significantly enhance current transportation modeling applications such as arterial performance measurement. Receiving and processing mobile sensor data however may involve severe privacy concerns if not properly designed. In contrast to the fact that the current research on transportation modeling and privacy protection is rather separated, we propose in this paper a framework on privacy-aware transportation modeling (PATM) and application-aware privacy protection (AAPP). The proposed framework focuses on the interactions between transportation modeling and privacy preserving, being aware of privacy when developing transportation models as well as application needs when designing privacy preserving mechanisms. Using two case studies, we show how PATM and AAPP may be applied in privacy-preserving mobile data collection while satisfying the application needs at the same time. The paper is concluded by discussing how to design a unified approach to guarantee privacy and data needs for various applications.

KEYWORDS: Transportation Modeling; Traffic Modeling; Privacy Protection; Location Privacy; Mobile Traffic Sensors; GPS Cellular Phones; Virtual Trip Lines; Arterial Performance Measurement; OD Estimation;

1 1. Introduction and Motivation

2 *Mobile traffic sensors* refer to those that move with the traffic flow. They are in contrast to fixed-
3 location sensors like loop detectors that dominant current traffic detection system. Broadly
4 speaking, mobile traffic sensors include any monitoring or data collection system with a
5 dedicated device equipped with vehicles. In this sense, they include for example probe vehicles
6 (such as those equipped with Electric Toll Collection (ETC) tags), cellular phones, portable GPS
7 devices (like GPS cellular phones), GPS navigation systems, Bluetooth Mac Address Matching
8 (BMAM, see Wasson et al. 2008), vehicles in Vehicle Infrastructure Integration (VII, now called
9 IntelliDrive (2009)), etc. Most mobile sensors need to communicate with either satellites (GPS)
10 or cellular towers (cell phones) or dedicated roadside infrastructures (ETC, VII, BMAM) to
11 derive the mobile component's position, speeds, and other relevant information.

12 In this paper, we focus on mobile sensors that can provide detailed tracking capability including
13 GPS and cellular techniques. The current penetration of these mobile sensors is low (about 17%
14 according to HarrisInteractive (2007)). As indicated by the International Telecommunication
15 Union (ITU, 2009), the penetration of cell phones is nearly 100% in developed countries and
16 nearly 50% in developing countries in 2007. Since most future cellular phones may be equipped
17 with GPS, the high penetration of mobile sensors is expected in the near future. In addition, the
18 implementation of IntelliDrive will make the high penetration of mobile sensors possible.

19 *Mobile sensor data collection in transportation applications*

20 Mobile sensors can potentially provide detailed traces (trajectories) of individual vehicles, which
21 contain rich information about traffic conditions especially when the data are widely available.
22 As mobile data are usually *samples* of real traffic flow, traditional modeling techniques based on
23 fixed-location sensor data such as flow, occupancy, etc., may not work directly. As a result novel
24 modeling techniques are needed to use mobile data more effectively (Zhou, 2004; Ban et al.
25 2009; Ban and Hao, 2010). In the past issues related to deploy mobile sensors (such as coverage
26 and penetration) and traffic modeling using mobile data have been investigated (Turner and
27 Holdener, 1995; Quiroga and Bullock, 1998; Cheu et al. 2002; Qiu et al. 207; Smith and
28 Fontaine, 2007). On the other hand, receiving and processing raw mobile trajectory data leads to
29 severe privacy concerns (Congressional Record, 2001; Karger and Frankel, 1995; Agre, 1995;
30 Warrior et al. 2003). The reality is that both agencies and private sector organizations who for
31 various reasons have collected a large volume of trajectory data are extremely reluctant to share
32 this information, and privacy protection is one of the main reasons. The first author personally
33 experienced the difficulty of obtaining ETC data from transportation management agencies: due
34 to privacy concerns, the data must be deleted within 24 hours. Trucking companies like FedEx
35 are continuously monitoring the movement of their delivery trucks. These data however are not
36 shared as they are considered valuable business property and might contain sensitive information
37 about their drivers. As a result, almost no mobile data are now publicly available.

38 To protect privacy, it is a common practice to place restrictions on sharing data via policies or
39 regulations. It is also often beneficial for organizations to avoid the collection of personally
40 identifiable information. Possession of such data leads to the following disadvantages. First,

1 managing such data incurs overhead, at least for organizations that need to comply with privacy
2 regulations. Examples include US government organizations covered under the Privacy Act of
3 1974 and businesses in the EU, which need to comply with the EU Data Directive 95/46/EC.
4 These regulations, for example, require organizations to provide notice to data subjects, to allow
5 access to records by data subjects, and to keep the data secure. Second, holding privacy sensitive
6 data presents public relations and liability risks if such data is compromised. Privacy breaches
7 occur frequently as public records indicate (Privacy Rights Clearinghouse, 2009), due to stolen
8 laptops, dishonest insiders, or external hackers, among others. Third, even without such breaches,
9 it is difficult to guarantee that the data will not be used for purposes other than the originally
10 intended one. Data may be subpoenaed for civil cases or new owners of an organization may
11 want to put the data to new uses. Travelers aware of such risks and concerned about their privacy
12 may avoid services that collect personal data, thus leading to lower participation rates.

13 Due to these privacy concerns, several recent applications involved with mobile data collection
14 have put more emphasis on privacy protection:

- 15 • An AITS project originally led by researchers from Rensselaer Polytechnic Institute
16 applied GPS devices to collect traffic and behavioral data from a group of recruited
17 drivers (List et al. 2005). To protect privacy, trajectory data were *not* collected directly;
18 instead the concept of *monuments* was developed, which are pre-defined locations to
19 collect mobile sensor data such as travel times between two monuments on arterial streets
20 (He et al. 2002; Demers et al. 2006).
- 21 • Nokia recently developed a mobile data collection system based on GPS cellular phones
22 (Herrera et al. 2009; Ban et al. 2009). Privacy protection is a major consideration of the
23 system and the concept of virtual trip lines (VTL) was developed to represent pre-defined
24 locations along roadways (Hoh et al. 2007). No trajectory information is collected by the
25 system; instead only speeds at VTL locations are collected (travel times between VTLs
26 may also be collected at sparse locations as discussed in Section 3).
- 27 • VII (IntelliDrive) contains a subsystem on archiving probe vehicle data. Privacy
28 protection is considered by encrypting the data transmitted between vehicles and roadside
29 devices, as well as between the devices and data archival servers. Overall, the privacy
30 issues of VII are addressed via nine privacy principles (Jacobson, 2007). The recent VII
31 Proof of Concept Test (RTIA, 2008) concludes that privacy is “still an important issue
32 and requires further evaluation and research.” Especially, privacy needs to be ensured via
33 the concept of “*privacy by design*” – to ensure privacy by designing the data collection
34 system so that privacy sensitive information is not collected or revealed. This is *in*
35 *contrast to* traditional mobile data collection practices for which data are collected first
36 and then processed to prevent the release of privacy sensitive information.

37 The above discussions and existing mobile sensor related studies revealed that (1) mobile data
38 are extremely valuable for traffic modeling and traveler information; (2) appropriate privacy
39 protection mechanisms need to be in place in order to share large volume of mobile sensor data
40 among transportation agencies, private sectors, academia, and the public; and (3) although there
41 have been sparse investigations and implementations, the privacy issue has not been well
42 addressed in the transportation community.

1 *Location privacy*

2 Although past surveys indicated that the public has been indifferent to their location privacy, as
3 the public is more aware of the adverse consequences of privacy leaks, they will be more
4 concerned about protecting their privacy before sharing data (Krumm, 2008). Duckham and
5 Kulik (2006) defined four general methods to ensure location privacy including *regulatory*
6 *strategies, privacy policies, anonymity, and obfuscation*. The first two are related to policies,
7 while the last two are on the computational (technical) aspects. The policy approaches, although
8 effective in certain cases, tend to be *conservative* as they usually prohibit the release of
9 information which may not be necessary. Also, since privacy issues are multiple-facet and
10 complicated, they have already created many controversies in designing proper policies and
11 regulations for protecting privacy (Solove, 2008). In this paper, we focus on technical
12 approaches for privacy protection (i.e. anonymity and obfuscation) rather than policy-related
13 mechanisms. This is called *location privacy* – which define quantitatively what privacy violation
14 is and how to collect/process mobile data (Krumm, 2008). Anonymity aims to maintain the
15 privacy of location data by replacing the associated name with an untraceable ID; Obfuscation
16 means to degrade the quality of location measurements by aggregation, adding noise, etc. to
17 reduce the possibility of privacy violation. Anonymity by simply removing identifiers from
18 location data has been found not effective to guarantee privacy (Gruteser and Hoh, 2005). Hoh et
19 al. (2007) further showed, using a dataset of week-long GPS traces from 239 drivers, that they
20 were able to find home locations of 85% of a subset of 65 drivers. Therefore novel privacy
21 schemes are usually needed to protect location privacy.

22 *A framework to consider privacy and modeling simultaneously*

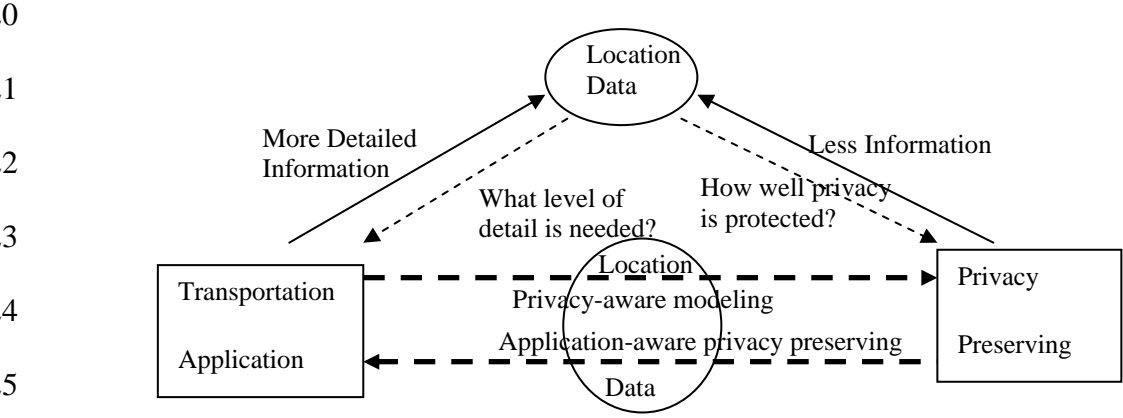
23 In summary, without proper considerations of privacy, travelers will be reluctant to share their
24 location data; the public and private sectors will also be concerned due to liability reasons. Novel
25 data collection schemes are needed to promote data sharing among different entities including
26 public and private sectors, researchers, and the public. As a result, for privacy researchers, how
27 to design privacy-preserving mobile data collection schemes is critical. As the mobile data
28 quality may be compromised after applying privacy-protection techniques (e.g. data noise is
29 added), this also calls for transportation researchers to develop new modeling approaches that
30 can extract meaningful transportation knowledge based on privacy-preserving mobile data.

31 In this paper, we propose a modeling framework that emphasizes *privacy-aware* transportation
32 modeling (PATM) techniques and *application-aware* privacy protection (AAPP) mechanism.
33 The framework can hopefully address both transportation application data needs and privacy
34 protection. We present two specific examples on how the framework can be applied to
35 transportation applications. The first example is for arterial performance measurement using
36 VTL data. We show how crucial performance measures such as delays, queue lengths etc. for
37 signalized intersections can be estimated using VTL data. The second example is regarding how
38 to improve freeway origin-destination (OD) demand estimation by fusing both detector data and
39 privacy-preserving mobile sensor data. We conclude by further discussing how to apply the
40 proposed framework to other types of transportation applications.

1 **2. Privacy-aware Transportation Modeling and Application-aware Privacy Protection**

2 Transportation modeling and privacy protection have been studied separately so far. On the one
 3 hand, with the primary goal being extracting information as much as possible, transportation
 4 researchers are aggressive in acquiring information: they tend to assume available data to the
 5 finest detail and consider privacy protection as a restricting factor for them to obtain “detailed”
 6 data. Developing privacy protection techniques, on the other hand, ensures policy makers, public
 7 agencies, private sectors, researchers, and the driving public that individual privacy is protected.
 8 This promotes sharing of mobile data: the public feels that their sensitive information will not be
 9 released and agencies/private sectors/researchers are more comfortable to share data as they have
 10 the mechanism to ensure privacy. However privacy experts have focused on protecting
 11 individual’s privacy, while mostly using very basic models of application data requirements,
 12 which may lead to the unnecessary prohibition of the release of some insensitive data.

13 As a result, there needs a fundamental change to the current thinking and modeling practices of
 14 transportation applications and privacy preserving. That is, we need a new paradigm that
 15 facilitates the close collaboration between privacy experts and transportation researchers to
 16 develop a holistic framework to (1) design effective privacy protection mechanisms, and (2)
 17 meanwhile to develop novel modeling approaches to extract information/knowledge using
 18 privacy-preserving mobile data. We denote this paradigm *privacy-aware* transportation modeling
 19 (PATM) and *application-aware* privacy preserving (AAPP), as depicted in Figure 1.



26 **Figure 1 A Holistic View for Privacy Protection and Transportation Modeling**

27 Figure 1 depicts the traditional way of transportation modeling and privacy preserving (on the
 28 top) and the newly proposed holistic modeling framework (on the bottom). From the traditional
 29 perspective, with one pushing for more information on the modeling side and another prohibiting
 30 the release of information from the privacy preserving side, consensus can hardly be reached on
 31 which level of detail the data should be released. To address these challenges, our proposed
 32 approach focuses on the interactions between transportation modeling and privacy preserving,
 33 being aware of privacy when developing transportation models as well as application needs
 34 when designing privacy preserving mechanisms. As a result, both privacy protection and
 35 application needs can be satisfied to the maximum extent possible.

1 The challenge of the newly developed PATM and AAPP framework is that as privacy protection
2 mechanisms will likely compromise the original mobile data, privacy experts need to know what
3 is the best way to do so and transportation researchers need to develop novel transportation
4 modeling techniques for knowledge/information extraction based on the privacy preserving
5 mobile data. While knowledge extraction and privacy preserving do have seemingly conflicting
6 objectives, the authors believe that one can satisfy both in most cases using innovative designs
7 for privacy-protection and modeling. This is due to the fact that privacy does not necessarily
8 need to be compromised to satisfy application data requirements: An application-aware design of
9 privacy algorithms can retain features important for the application, while still achieving privacy
10 by removing features that are less important. For example, cloaking techniques to depersonalize
11 location data can achieve the same level of privacy by retaining high spatial resolution and
12 reducing temporal resolution or by providing high temporal resolution and reducing spatial
13 resolution (Gruteser and Grunwald, 2003). Similarly, as shown in previous research (Ban et al.
14 2009; Ban and Hao, 2010), in order to extract transportation information like crucial traffic
15 measures (such as real time queue length or delay at a signalized intersection), it is NOT
16 necessary to know the detailed trajectory of every vehicle. Rather, some aggregated mobile data
17 (such as individual travel times between an upstream and a downstream location of the
18 intersection) can suffice, which imposes little privacy concerns (Hoh et al. 2008).

19 As will be presented in the next two sections, in the current practice, the proposed scheme can be
20 generally initiated from the transportation modeling side. However one has to keep in mind that
21 both sides need to be considered. The challenge is to reach a point that both objectives are met in
22 the best possible way without compromising the other party severely. For this purpose, an
23 iterative process is necessary.

24 **3. Virtual Trip Lines for Arterial Performance Measurement**

25 The first case study is based on the mobile sensor data collection system developed by Nokia that
26 collects virtual trip lines (VTL) data. VTLs are pre-defined (and *virtual*) locations on roadways.
27 When crossing a VTL, a vehicle equipped with mobile sensors will report its speed. As shown in
28 Hoh et al. (2008), using VTL to regulate location and speed reports reduces privacy concerns,
29 compared to probe vehicle systems which periodically transmit location reports. Privacy is
30 increased because VTLs can be placed into less privacy sensitive areas and areas where traffic
31 information is most important (such as major highways or arterial intersections). When location
32 updates are anonymized and trip lines are placed with sufficient spacing, it also becomes difficult
33 to track the path of an individual vehicle for an extended duration. Thus this approach avoids the
34 collection of potentially sensitive trip information and in particular trip endpoint information.

35 Arterial modeling has for long suffered from insufficient data coverage. Mobile devices such as
36 GPS cellular phones provide a great potential for low-cost, wide-area data collection of arterials.
37 For arterial performance measurement, the key is the performance of signalized intersections as
38 they contribute to the majority of the delay on arterial streets. With both data collection and
39 privacy protection in mind, PATM and AAPP can be naturally applied to extend the VTL
40 approach. The process can be generally initiated from the application side. First, transportation
41 researchers can examine the original VTL data as described in Hoh et al. (2008) which only

1 contains speed information and a trip line identifier, and the placement of VTLs does not allow
2 tracking of a vehicle across multiple VTLs. They then conclude that it is very difficult to obtain
3 intersection performance measures based on the original VTL data. Transportation researchers
4 can further examine the data needs for estimation arterial performances and find that ravel time
5 or delay information of individual vehicles passing an intersection can be used to estimate
6 intersection delay patterns, arrival volumes, and queue lengths. They subsequently introduce the
7 requirement to obtain travel times across an intersection. The privacy experts then need to check
8 whether collecting intersection travel times could severely impact privacy. In general, privacy
9 increases if an adversary cannot determine the complete path of a traveler. Since cars have to
10 follow roadways, on roadway segments confusion is only possible if the spacing of trip lines is
11 sufficiently large so that the order of cars changes in between with high probability. At
12 intersections, however, privacy can increase significantly if it is not possible to determine
13 whether a car took a turn or traveled straight ahead. While introducing travel time monitoring
14 across an intersection reveals turn information for cars, a high degree of privacy can still be
15 preserved if a small subset of intersections is chosen for monitoring so that any given vehicle
16 trip still passes through other unmonitored intersections with high probability. As a result,
17 collecting travel times for major intersections only does preserve high level of privacy violation.
18 At this point, a consensus is reached between transportation researchers and privacy experts that
19 travel times of an intersection can be collected that can be used for arterial intersection
20 performance measurement purpose without severely violate privacy.

21 The VTL case study in this section serves to illustrate how PATM & AAPP can be designed to
22 provide both adequate privacy protection and data quality for applications. The process is in
23 general iterative before a proper balance can be reached.

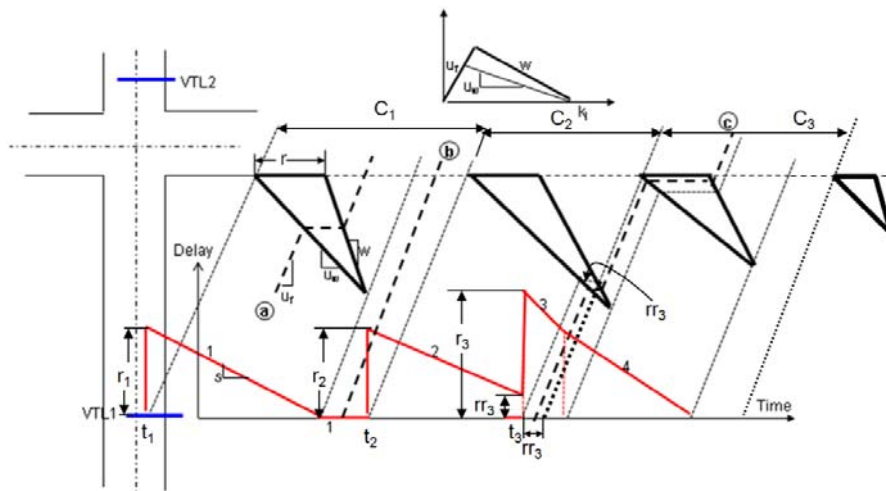
24 Intersection modeling based on VTL travel times are discussed in detail in Ban et al. (2009) and
25 Ban and Hao (2010). For the sake of completeness, we briefly present the major findings here,
26 focusing on how to estimate intersection arrival volume which was not discussed previously in
27 detail. As intersection delays are usually a function of incoming flows and signal timing settings,
28 it is possible to *reconstruct* flow and signal timing information from measured travel times or
29 delays. This however requires a *reverse* thinking process compared with traditional arterial
30 modeling methods based on loop detectors: one needs to reconstruct traffic states (volume, queue
31 length, etc.) from travel times while traditional methods all use traffic volume as an input (to
32 estimate delays). The key is to identify *discontinuities* or *nonsmoothness* of intersection delays
33 from sampled travel times, which indicate a traffic state change or signal status change.

34 In Ban et al. (2009), we showed that if the penetration of mobile sensors is sufficiently high
35 (e.g. $\geq 20\%$), the time-dependent intersection delay pattern can be estimated via the sampled
36 travel times. Figure 2(a) below depicts how this can be done. The figure shows a typical
37 signalized intersection with two VTLs installed upstream (VTL1) and downstream (VTL2)
38 respectively. Under the assumption that a queue never passes VTL1, we can use the bold solid
39 triangles in the figure to represent how queue forms and dissipates based on shockwave theory
40 (Lighthill and Whitham, 1955). The horizontal part of the triangles represents the duration of red
41 time. If delays due to vehicle decelerations and accelerations are ignored and the arrival rate is
42 uniform within one cycle, delays can be fully determined by the triangles. In the figure, dashed

1 lines represent trajectories of vehicles, while dotted lines are boundaries when the *discontinuities*
 2 of delays happen. We aim to characterize vehicle delays as a function of the time when it passes
 3 VTL1. Here we assume that data have been collected and thus one can perform “post-
 4 processing” to re-construct a *mapping* from the time that a vehicle passed VTL1 to its
 5 experienced delay at the intersection. As shown by the trajectories of vehicles (dashed lines), if a
 6 vehicle approaches the intersection in red time or if the queue length is not zero (e.g. trajectory *a*
 7 in the figure), the vehicle will join the end of the queue first and thus be delayed. The delay
 8 encountered by the vehicle is the horizontal part of trajectory *a*. Otherwise, if a vehicle arrives
 9 during green time and there is no queue (e.g. trajectory *b*), the vehicle will pass the intersection
 10 with no delay. The (red) delay curve at the bottom of Figure 2(a) will spike up at the time that
 11 allows a vehicle to travel to the intersection in free flow just before the start of the red time.
 12 More importantly, by analyzing the geometry of the triangles, one can observe that if a vehicle
 13 passes by VTL1 at a time that would make it get to the intersection just after the start of the red
 14 time, delay for this vehicle will be the maximum for the specific cycle. After that, delays will be
 15 reduced linearly until no delay is reached. This is represented by the line segments marked as “1”
 16 of the delay curve at the bottom of Figure 2(a). The slope of the delay reduction part, denoted as
 17 delay reduction rate *s*, can be calculated analytically as (Ban et al. 2009):

$$18 \quad s = \frac{u_f(w - u_w)}{w(u_f + u_w)} = \frac{v}{k_j} \left(\frac{1}{u_f} + \frac{1}{w} \right) - 1 \quad (1)$$

19 Here *w* is the shock wave speed, *u_f* is the free flow speed, *u_w* is the wave speed when a vehicle
 20 joins the queue, *k_j* is the jam density, and *v* is traffic flow which is assumed to be constant within
 21 a cycle. The three parameters *u_f*, *w*, *k_j* are specific to actual arterial locations, which also
 22 determine the *fundamental diagram* of the location. Notice that since $w \geq u_w$ always hold (refer to
 23 the fundamental diagram at the top of Figure 2(a)), *s* is positive, meaning that delay always
 24 reduces from its maximum (when traffic light turns red) to some minimum value (when light
 25 turns green and no queue exists) for normal situations.



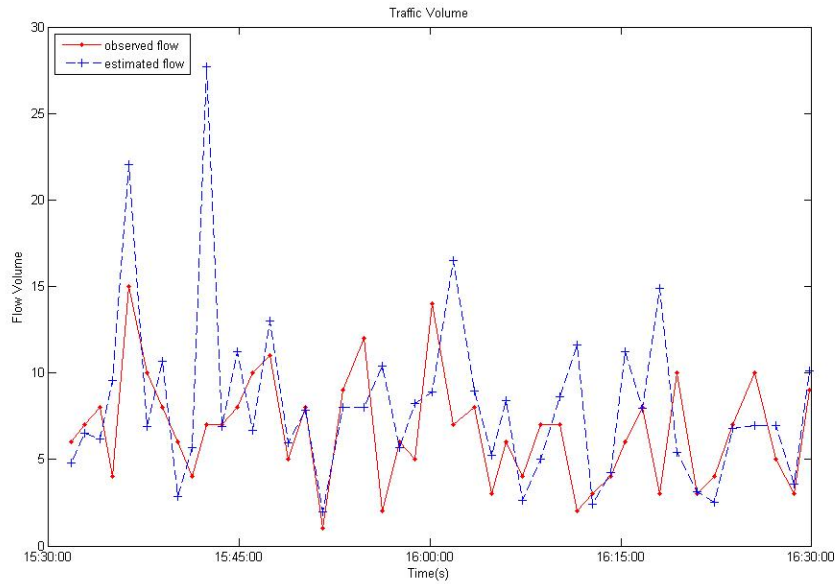
26

27

Figure 2(a) Intersection Delay Pattern

1 In Ban et al. (2009a), we further developed a two-step estimation method to construct the delay
 2 pattern (red lines at the bottom of Figure 2(a)) using sampled travel times. The delay pattern
 3 turns out to be critical as it can be used to further estimate traffic volume and queue length.
 4 Equation (1) clearly shows that if the delay pattern can be estimated, the delay reduction rate s is
 5 known. Then the only unknown in (1) is the traffic volume v , which can be calculated as follows:

$$6 \quad v = \frac{(s+1)k_f}{\frac{1}{u_f} + \frac{1}{w}} \quad (2)$$



7

8

Figure 2(b) Performance of Volume Estimation

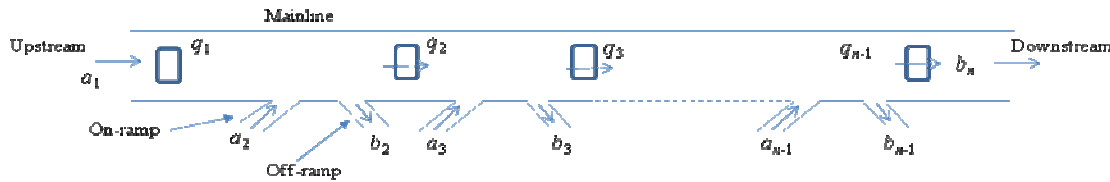
9 It was found in Ban et al. (2009) that under relatively high penetration (>20%), VTL data can be
 10 used to estimate the signalized intersection delay patterns relatively accurately. Under even
 11 higher penetration (>60%), the signal timing parameters (like cycle length, duration of red and
 12 green times, etc.) can also be derived. Here we present the estimation results of intersection
 13 arrival volumes using VTL data based on equation (2). Figure 2(b) depicts the estimated volume
 14 (dashed line with '+' signs) and observed volume (solid line with '.' signs) for a signalized
 15 intersection in micro-simulation (Ban et al. 2009). Although significant discrepancies between
 16 the two lines can be observed, we still see that the estimated volume can follow the trend of the
 17 observed volume fairly closely. If we aggregated volumes over 5-minute, 15-minute, 30-minute,
 18 and the entire 1-hour period, the estimation accuracy can be much improved. The percentage root
 19 mean square errors (PRMSE) are 29.5%, 12.8%, 7.4%, and 8.9% respectively. The equation for
 20 computing PRMSE is listed below:

$$21 \quad PRMSE = \frac{|\hat{v} - \bar{v}|}{\bar{v}} \times 100. \quad (3)$$

1 Here \hat{v} and \bar{v} are the estimated and observed volumes respectively for an appropriate period of
 2 time. As shown in Ban and Hao (2010), the time-dependent queue length of signalized
 3 intersections can also be estimated using VTL data. This shows the potential of using privacy
 4 preserving mobile data to estimate real time performances of signalized arterial intersections.

5 4. Mobile Data for Freeway Origin-Destination Demand Estimation

6 Most existing OD demand estimation methods are based on fixed locations sensors like loop
 7 detectors, which have been extensively studied for both the static case (Cascetta and Nyugen,
 8 1988; Bell, 1991; Yang et al. 1992) and the dynamic case (Cascetta et al. 1993; Chang and Tao,
 9 1996; Zhou and Mahmassani, 2007; Nie and Zhang, 2008). Here we illustrate how privacy-
 10 preserving mobile sensor data can be used to enhance existing OD estimation results. A well-
 11 recognized problem of using traffic counts from loop detectors to estimate OD demands is that
 12 the resulting model is usually under-determined: i.e. the number of unknown OD variables is far
 13 more than the number of locations (detector stations) for collecting traffic counts.
 14 Correspondingly, there are usually multiple OD demand patterns that can produce exactly the
 15 same set of observed traffic counts. We illustrate this using the example of estimating freeway
 16 (static) OD demands as shown in Figure 3. The figure depicts a freeway segment with n ramp
 17 locations (the first location denotes the upstream freeway mainline and the last location is the
 18 downstream freeway mainline). Denote a_i the inflow and b_i the exit flow for $1 \leq i \leq n$, and q_i is
 19 the mainline volume between location i and $i+1$ for $1 \leq i \leq n-1$. Assume $x_{i,j}$ is the demand from
 20 location i to location j for $1 \leq i \leq n, 1 \leq j \leq n$. As this is a freeway segment (i.e. one way), we have
 21 $x_{ij} = 0$ for all $i \geq j$. The OD estimation problem is to determine $x_{i,j}$ given a_i , b_i , and q_i .



22
23 **Figure 3 Freeway OD Estimation**

24 We should point out here that this static version of the freeway OD estimation problem is rather
 25 preliminary as more sophisticated dynamic models have been proposed in the literature (Chang
 26 and Wu, 1994). Therefore the problem is selected mainly to illustrate the concept in this paper. A
 27 simple optimization model for the freeway OD estimation problem can be given as below:

$$\min_{x_{ij}} z = \sum_{i=1}^{n-1} \left| \sum_{l=1}^i \sum_{j=i+1}^n x_{lj} - q_i \right| \quad (4-1)$$

$$28 \quad a_i = \sum_{j=i+1}^n x_{ij}, i = 1, 2, \dots, n-1 \quad (4-2)$$

$$b_j = \sum_{i=1}^{j-1} x_{ij}, j = 2, \dots, n \quad (4-3)$$

$$x_{ij} \geq 0, i = 1, 2, \dots, n; j = i+1, \dots, n \quad (4-4)$$

1 Here the objective function is the absolute deviation from the observed mainline traffic counts (q_i)
2 and the estimated traffic counts from OD demands $\sum_{l=1}^i \sum_{j=i+1}^n x_{lj}$. Equation (4-2) is the constraint that
3 the inflow at location i must exit at downstream locations ($i+1$ to n); equation (4-3) is the
4 constraint that the exit flow at location j must come from upstream locations (1 to $j-1$). The
5 model is clearly a linear programming problem by noticing that the absolute value in (4-1) can be
6 easily converted to linear terms by introducing extra variables. The model has $n(n-1)/2$
7 unknowns for x_{ij} and $3n-5$ observed values for a_i , b_i , and q_i . If n is large, the number of
8 unknowns is larger than the number of observed values – in which case the model is under-
9 determined. For example, here we assume $n=4$ and the observed values are given in Table 1. It
10 turns out that there are multiple optimal solutions that can exactly match the observed values (i.e.
11 the optimal objective value in (4-1) is zero) and Table 1 shows two of them.

12

Table 1 OD Estimation Results

i	ai	bi	qi	Res 1				Res2			
				1	2	3	4	1	2	3	4
1	3000		3000	500	1000	1500		500	1500	1000	
2	100	500	3500		500	500				1000	
3	500	1500	2500			500				500	
4		2500									

13

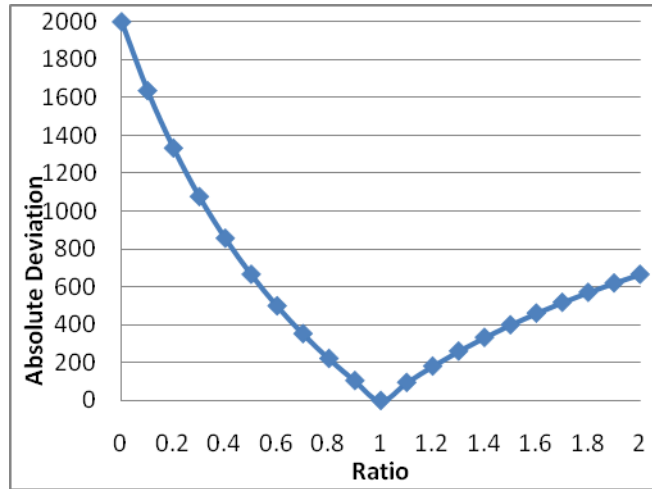
14 The above example shows that traditional models based on traffic counts may not produce the
15 “true” OD demand matrix. This can cause problems. Assume the true OD matrix is “Res1” but
16 “Res2” is generated by the estimation model. In this case, due to the error of predicting OD
17 demand, certain control strategies like ramp metering or pricing may fail due to excessive or
18 fewer demands at specific locations of the freeway.

19 If mobile data are available, they can provide a sample of the ratio between $x_{2,3}$ and $x_{2,4}$. Assume
20 the ratio is $\bar{\beta} = \bar{x}_{2,3} / \bar{x}_{2,4}$ where $\bar{x}_{2,3}$ and $\bar{x}_{2,4}$ are observed demands. The objective function of (4-1)
21 may be revised as:

$$22 \min_{x_{ij}} z = \sum_{i=1}^{n-1} \left| \sum_{l=1}^i \sum_{j=i+1}^n x_{lj} - q_i \right| + w \left| x_{2,3} - \bar{\beta} x_{2,4} \right| \quad (5)$$

23 Clearly the revised objective is a weighted summation of the total deviation at mainline traffic
24 count locations (the first term) and the deviation between $x_{2,3}$ and $x_{2,4}$ and their observed values.
25 Assume the observed demands are representative, i.e. $\bar{\beta} = 1$ for this particular example (both $x_{2,3}$
26 and $x_{2,4}$ are 500 for “Res1” as shown in Table 1). The revised model generates exactly “Res1”
27 indicating that the model performs better than model (3). In reality, $\bar{\beta}$ may not be completely
28 representative. Figure 4 shows how the results vary based on different $\bar{\beta}$'s. It can be seen from
29 the figure that even the estimates of $\bar{\beta}$ have significant errors, the model results are still
30 acceptable. For example, for $0.6 \leq \bar{\beta} \leq 1.5$ (i.e. $\bar{\beta}$ is estimated from 40% lower to 50% higher

1 than the true value), the absolute error of the estimated demand matrix from (4) and the true OD
 2 (“Res1” in Table 1) is less than 450, 10% from the total OD demands.



3
 4 **Figure 4 Performances of OD Estimation Results vs. Estimate of Ratio**

5 The example above shows that by combining aggregated mobile sensor data (i.e. the ratio of
 6 demands for different OD pairs) with traditional traffic counts, OD estimation results can be
 7 significantly improved. To obtain the aggregated ratio information, one does not need detailed
 8 vehicle trajectory data. This provides more flexibility for designing privacy protection
 9 mechanisms. One option is to apply random perturbation to the OD pair reported by each vehicle
 10 and then apply privacy preserving data mining techniques to reconstruct the original distribution
 11 of OD values (Agrawal and Srikant, 2000). If portions of the trajectory should be preserved also,
 12 a random path swapping (RPS) technique can be applied to the mobile data collection process.
 13 RPS randomly decides to switch (or not switch) the non-overlapping sections of the paths of two
 14 randomly selected individual traces between the same OD pair. This can be depicted Figure 5. If
 15 the OD zones are large, such a scheme can provide a basic level of privacy because (1) it
 16 provides “misleading” privacy information which can confuse a privacy adversary, and (2) even
 17 for the same individual, the adjusted traces over multiple days will unlikely to form any pattern
 18 (or if any pattern is formed, it should be very different from the true pattern), which may not be
 19 used by the adversary to identify the individual. RPS however does not modify a large portion of
 20 the route and does not change the total traffic demand for any OD pair so that $\bar{\beta}$ can be easily
 21 retrieved from the collected mobile data. These schemes therefore represent another example of
 22 collecting privacy-preserving mobile data to fulfill the needs of certain transportation
 23 applications (i.e. freeway OD estimation).

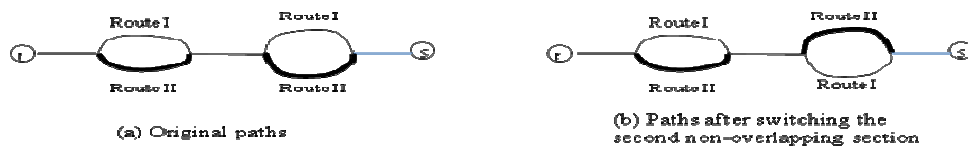


Figure 5: Random Route Switching Algorithm for Privacy Preserving

1 **5. Conclusions and Future Study**

2 In this paper, we proposed a privacy-aware transportation modeling (PATM) and application-
3 aware privacy protection (AAPP) framework for privacy-preserving transportation modeling
4 using mobile sensor data. The proposed framework focuses on the interactions between
5 transportation modeling and privacy preserving – being aware of privacy when developing
6 transportation models as well as application needs when designing privacy preserving
7 mechanisms. The framework was illustrated using two case studies: arterial performance
8 measurement using VTL travel times and freeway OD estimation using aggregated OD ratio
9 information. It was shown that the PATM & AAPP framework can be generally initiated by
10 investigating the application data needs and then adjusted by privacy preserving principles. The
11 process is usually iterative in order to reveal the minimum data requirements and to design the
12 most effective way to provide such data without severely comprising privacy. Therefore in most
13 cases, innovations are needed to design non-traditional data collection schemes, as well as to
14 develop novel transportation modeling techniques based on the new data formats.

15 The widely available mobile traffic sensors present great potentials, while at the same time great
16 challenges, for privacy protection and modeling using mobile data. The proposed PATM &
17 AAPP framework aims to address this issues; the results reported in this paper are just the first
18 step in the sense that the framework was only illustrated using two case studies. However, it
19 suffices to show that privacy protection and transportation modeling are two equally important
20 aspects to use mobile sensor data, which need to be addressed simultaneously. In the future, the
21 framework will be extended and tested on more applications. More importantly, theoretical
22 investigations are needed to provide a comprehensive framework on how to systematically
23 protect privacy while preserving critical information as much as possible. For this, novel privacy
24 protection mechanisms and mobile-data based traffic modeling methodologies are needed. In this
25 process, close collaborations between privacy protection experts and transportation modeling
26 researchers is the key.

27 **References**

- 28 1. Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. *SIGMOD Rec.* 29, 2,
29 439-450.
- 30 2. Agre, P.E. (1995). Transport informatics and the new landscape of privacy issues.
31 *Computer Professionals for Social Responsibility (CPSR) Newsletter*, 13(3).
- 32 3. Balke, K., Charara, H., & Parker, R. (2005). Development of a Traffic Signal
33 Performance Measurement System (TSPMS). Texas Transportation Institute, Report 0-
34 4422-2.
- 35 4. Ban, X., Herring, R., Hao, P., and Bayen, A. (2009) Delay pattern estimation for
36 signalized intersections using sampled travel times. To Appear in *Transportation*
37 *Research Record*.
- 38 5. Ban, X., and Hao, P. (2010) Real-time queue length estimation for signalized
39 intersections using sampled travel times. To be presented at the 89th Annual meeting of
40 Transportation Research Board.

- 1 6. Bell, M., 1991. The estimation of origin–destination matrices by constrained generalized
2 least squares. *Transportation Research Part B* 25, 13–22.
- 3 7. California Center for Innovative Transportation (CCIT), 2006. Final Report of Task
4 Order 3: Corridor Management Plan Demonstration. California Center for Innovative
5 Transportation, University of California, Berkeley.
- 6 8. Cascetta E. and Nguyen S. (1988) A unified framework for estimating or updating
7 origin/destination matrices from traffic counts. *Transportation Research Part B* 22, 437-
8 455.
- 9 9. Cascetta, E., Inaudi, D., Marquis, G., 1993. Dynamic estimators of origin–destination
10 matrices using traffic counts. *Transportation Science* 27 (4), 363–373.
- 11 10. Chang, G.-L., Wu, J., 1994. Recursive estimation of time-varying origin–destination
12 flows from traffic counts in freeway corridors. *Transportation Research Part B* 28 (2),
13 141–160.
- 14 11. Chang, G.-L., Tao, X., 1996. Estimation of dynamic O–D distribution for urban network.
15 In: *Proceedings of the 13th International Symposium on Transportation and Traffic*
16 *Theory*, pp. 1–20.
- 17 12. Cheu, R.L., Xie, C., Lee, D., 2002. Probe vehicle population and sample size for arterial
18 speed estimation. *Computer-Aided Civil and Infrastructure Engineering* 17 (1), 53–60.
- 19 13. Comert, G., and Cetin, M. (2008), Queue Length Estimation from Probe Vehicle
20 Location and the Impacts of Sample Size, To appear in the *European Journal of*
21 *Operational Research*.
- 22 14. Congressional Record (2001). Location privacy protection act of 2001. Available on
23 Internet: <http://www.techlawjournal.com/cong107/privacy/location/s1164is.asp>.
- 24 15. Darroch, J.N. (1964). On the traffic-light queue. *The Annals of Mathematical Statistics*,
25 380–388.
- 26 16. Demers, A., List, G., Wallace, W.A., Lee, E.E., Wojtowicz, J. (2006) Probes as a path
27 seeker: a new paradigm. *Transportation Research Record* 1944, 107-114.
- 28 17. Duckham, M. and Kulik, L. (2006) Location privacy and location-aware computing, in
29 *Dynamic & Mobile GIS: Investigating Change in Space and Time*, J. Drummond, et al.,
30 Editors. CRC Press: Boca Raton, FL USA., 34-51.
- 31 18. Gruteser, M., and Hoh, B. (2005). On the anonymity of periodic location samples. In
32 *Proceedings of the Second International Conference on Security in Pervasive Computing*.
- 33 19. Gruteser, M. and Grunwald, D. (2003). Anonymous usage of location-based services
34 through spatial and temporal cloaking. In *Proceedings of First ACM/USENIX*
35 *International Conference on Mobile Systems, Applications, and Services (MobiSys)*, San
36 Francisco, CA.
- 37 20. Haight, F.A., 1959. Overflow at a traffic flow. *Biometrika* 46 (3-4), 420–424.
- 38 21. HarrisInteractive (2007) National Study Shows GPS Adoption Rates Relatively Low, but
39 Offers Recommendations to Accelerate Market. Available on Internet:
40 Penetration<http://www.harrisinteractive.com/NEWS/allnewsbydate.asp?NewsID=1241>.
- 41 22. He, R., Liu, H.X., Kornhauser, A.L., and Ran, B. (2002). Study travel time variability
42 from probe vehicle data. *Proceedings of the seventh International Conference on:*
43 *Applications of Advanced Technology in Transportation*, Cambridge, MA, United States,
44 16-23.

- 1 23. Herrera, J.C., Work, D.B., Herring, R., Ban, X., and Bayen, A. (2009) Evaluation of
2 traffic data obtained via GPS-enabled mobile phones: the Mobile Century field
3 experiment. Accepted by Transportation Research Part C.
- 4 24. Hoh, B., et al., Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking,
5 in 14th ACM Conference on Computer and Communication Security (ACM CCS 2007).
6 2007: Alexandria, VA USA.
- 7 25. Hoh, B., Gruteser, M., Herring, R., Ban, J., Work, D., Herrera, J.C., and Bayen, A.. (2008)
8 Virtual trip lines for distributed privacy-preserving traffic monitoring. In The Sixth
9 Annual International conference on Mobile Systems, Applications and Services
10 (MobiSys 2008), Breckenridge, U.S.A., June 2008.
- 11 26. IntelliDrive (2009). Available on Internet: see <http://www.intelidriveusa.org/index.php>.
- 12 27. International Telecommunication Union (ITU, 2009). Internet Link (Accessed on Dec. 02,
13 2009): <http://www.itu.int/ITU-D/ict/statistics/ict/graphs/mobile.jpg>.
- 14 28. Jacobson, L. (2007). Vehicle Infrastructure Integration Privacy Policies Framework (ver
15 1.0.2). Reported of The Institutional Issues Subcommittee of the National VII Coalition.
- 16 29. Karger, P.A., and Frankel, Y. (1995). Security and privacy threats to ITS. In Proceedings
17 of the Second World Congresson Intelligent Transport Systems, volume 5, Yokohama,
18 Japan.
- 19 30. Krumm, J. (2008) A survey of computational location privacy. Personal and Ubiquitous
20 Computing.
- 21 31. Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves I Flood movement in long
22 rivers. II A theory of traffic flow on long crowded roads. Proceedings of Royal Society
23 (London) A229, 281–345.
- 24 32. List, G., Wallace, W.A., Demers, A., Salaszyk, P., Lee, E.E., Wojtowicz, J. (2006) Field
25 experience with a wireless GPS-based ATIS system. In Proceedings of the 12th ITS
26 World Congress (CD-ROM), San Francisco.
- 27 33. Liu, X., and Ma, W. (2008) A real-time performance measurement system for arterial
28 traffic signals. Presented at the 87th Annual Meeting of Transportation Research Board.
- 29 34. Liu, X., Wu, X., Ma, W., and Hu, H. (2009) Real time queue length estimation for
30 congested signalized intersections. Transportation Research, Part C, in press.
- 31 35. McNeil, D.R., 1968. A solution to the fixed cycle traffic light problem for compound
32 Poisson arrivals. Journal of Applied Probability 5, 624–635.
- 33 36. Newell, G. F. 1960. Queues for a fixed-cycle traffic light. Ann. Math. Statist. 31 589–597.
- 34 37. Newell, G.F., 1965. Approximation methods for queues with application to the fixed-
35 cycle traffic light. SIAM Review 7, 223–240.
- 36 38. Nie, Y., and Zhang, H.M. (2008). A variational inequality formulation for inferring
37 dynamic origin–destination travel demands Transportation Research Part B, 42, 635–662.
- 38 39. Privacy Rights Clearinghouse (2009). A Chronology of Data Breaches, Available on
39 Internet: <http://www.privacyrights.org/ar/ChronDataBreaches.htm> (Accessed:
40 12/18/2009).
- 41 40. Qiu, Z., Chen, P., Jing, J, and Ran, B (2007) Cellular probe technology applied in
42 advanced traveler information. *International Journal of Technology Management*, in
43 press.

- 1 41. Quiroga, C.A., Bullock, D. (1998). Travel time studies with global positioning and
2 geographic information systems: an integrated methodology. *Transportation Research*
3 *Part C* 6(1-2), 101-127.
- 4 42. RITA (2008). VII Proof of Concept Test Final Reports. Available on Internet:
5 <http://www.resourceguide.its.dot.gov/default.asp?SID=3&SSID=9>.
- 6 43. Skabardonis, A., Geroliminis, N., 2008. Real-time Monitoring and Control on Signalized
7 Arterials. *Journal of Intelligent Transportation Systems*. 12 (2), 64–74.
- 8 44. Smaglik, E. J., Sharma, A., Bullock, D.M., Sturdevant, J.R., and Duncan, G. (2007).
9 Event-Based data collection for generating actuated controller performance measures.
10 *Transportation Research Record* 2035, 97-106.
- 11 45. Smith, B.L., and Fontaine M.D., 2007. Private-sector provision of congestion data.
12 NCHRP report 70-1.
- 13 46. Solove, D. J. (2008) *Understanding Privacy*. Harvard University Press, Cambridge,
14 Massachusetts.
- 15 47. Turner, S., Holdener, D., 1995. Probe vehicle sample sizes for real-time information: The
16 Houston experience. In: *Proceedings of the Vehicle Navigation & Information Systems*
17 *Conference*, Seattle, WA, United States, pp. 3–10.
- 18 48. Warrior, J., McHenry, E., and McGee, K. (2003) They know where you are [location
19 detection], *Spectrum, IEEE* , 40(7), 20-25.
- 20 49. Wasson, J.S., J.R. Sturdevant, and D.M. Bullock. Real-Time Travel Time Estimates
21 Using Media Access Control Address Matching. *ITE Journal*, Vol. 78, No. 6, 2008, pp.
22 20-23.
- 23 50. Webster, F.V., 1958. Traffic signal settings. Road Research Laboratory Technical Paper
24 No. 39, HMSO, London.
- 25 51. Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin–destination
26 matrices from link traffic counts on congested networks. *Transportation Research B* 26,
27 417–434.
- 28 52. Zhou, X., 2004. Dynamic origin–destination demand estimation and prediction for off-
29 line and on-line dynamic traffic assignment operation. Ph.D. Thesis, University of
30 Maryland, College Park.
- 31 53. Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic
32 origin–destination demand estimation and prediction in a day-to-day learning framework.
33 *Transportation Research Part B* 41 (8), 823–840.