Uncertain Penetration Rate Issues in Mobile Intelligent Transportation Systems

Quang Tran Minh, Muhammad Ariff Baharudin, Eiji Kamioka

Graduate school of Engineering and Science, Shibaura Institute of Technology, Tokyo, Japan
{m709504, nb11507, kamioka}@shibaura-it.ac.jp

ABSTRACT

This paper thoroughly discusses the essential issues remaining in mobile phone technologies which impede the realization of mobile phone based traffic state estimation systems (M-TESs). Concretely, the inherent issues in mobile phone based applications, namely **low** and **uncertain** penetration rate issues, which affect the M-TES's effectiveness, are resolved. A unique GA-based velocity-density estimation mechanism is proposed to improve the traffic state estimation accuracy when the penetration rate is low but still be relevant. A novel ANN-based prediction model is proposed to cope with unacceptably low and uncertain penetration rate issues. Moreover, a reasonable selection method is proposed aiming at selecting an appropriate traffic state estimation model without the actual penetration rate information. Experimental evaluations reveal the effectiveness as well as the robustness of the proposed solutions.

Keywords: *mobile probes, low penetration rate, context-aware, genetic algorithm, neural network, ITS, M-ITS.*

1 INTRODUCTION

Transportation and road traffic are important parts of any economy. Meanwhile, traffic congestion still remains as a serious issue in almost every big city across the world. The Ministry of Land Infrastructure and Transport of Japan reported in 2006 that the economic loss caused by traffic jam is around $100 billion annually [1]. The Urban Mobility Report [2] reported in 2007 that traffic congestion causes 4.2 billion hours of extra travel time in the United States, accounting for 2.9 billion extra gallons of fuel [3]. Traffic state estimation is one of the most important fields in Intelligent Transportation Systems (ITS) research to alleviate traffic congestion. Existing systems utilize road-side fixed sensors such as loop detectors [4], RFID readers [5], cameras [6], etc., for real-time traffic data collection. However, these approaches face on the coverage limitation since it is impractical to install a huge number of sensors at every street.

In recent years, with the advances of mobile phone technologies, mobile devices have been utilized as traffic probes [7], [8]. Since mobile phones are available everywhere and mobile phone networks have already been deployed, the essential issues in traditional road-side fixed sensor approaches such as coverage limitation, real-time effect, investment cost, etc., can be overcome. As a result, the mobile phone based ITS (M-ITS) research is entering a new stage accelerating the realization of mobile phone based traffic state estimation systems (M-TES).

In spite of the aforementioned advantages, the M-TES faces on several issues ranging from comprehensively estimating traffic state using less informative traffic data reported by mobile devices [9], to effectively removing errors rooted from low and uncertain penetration rate [3], [10]. **Firstly**, in M-TES, traffic data is collected by mobile phones on which GPS (Global Positioning System) receiver is only the common sensor available. However, the GPS data such as position (longitude, latitude), direction, velocity, etc., of vehicles in a traffic flow is not sufficient traffic state information than the drivers expect. **Secondly**, the penetration rate, namely the fraction of the number of vehicles that report data to the estimation server out of the total number of vehicles traveling through the considered road segment, is commonly low, or more seriously, is unknown at the estimation time. To the best of our knowledge, no research thoroughly discusses these issues. This article aims at proposing notable approaches coping with the **low** and **uncertain** penetration rate issues. Contributions of this paper are summarized as follows:

- Proposing reasonable solutions for penetration rate related issues. As a result, a novel GA-based velocity-density estimation mechanism and a unique ANN-based prediction model are proposed to deal with difficulties rooted from ***low*** penetration rate.
- A practical selection method aiming at selecting appropriate estimation model among two estimation models mentioned above under the condition of ***uncertain*** penetration rate is proposed.

This paper is organized as follows: Section 2 reviews the related work. Section 3 describes the problem formulation. A novel GA-based velocity-density inference model is proposed in section 4. Section 5 proposes an appropriate ANN-based prediction model to predict traffic state when the penetration rate becomes unacceptably low or unknown. Section 6 proposes a practical selection method for selecting an appropriate traffic state estimation model. The effectiveness of the proposed approaches is thoroughly evaluated in section 7. Section 8 concludes this work and draws out the future research directions.

2 RELATED WORK

Existing traffic state estimation systems such as VICS [11], NAVITIME [12] in Japan, the ITS project at Kansas, USA [13] mainly rely on road-side fixed sensors for traffic data collection. These road-side fixed sensor systems are costly in terms of initial installation and maintenance, thus they confront the coverage limitation issue. Theoretically, Ad-hoc network technology [1] can help to improve the coverage but it is not matured enough to be applied in real-world applications. Mobile Millennium Project (MMP) [14] is closely related to this work which employs GPS-enable mobile phones as traffic probes for real-time data collection. An estimation server at the system center processes traffic data, estimates traffic state, and informs drivers of estimated traffic information. The drawbacks of this project are as follows:

1) The MMP estimated traffic state by employing the dynamical theory to analyze vehicle flows on road networks. Obviously, the dynamical theory may work effectively in an environment of short traffic flows. It may reveal, however, serious errors when applying to environments of complex network with long roads.

2) Low and uncertain penetration rate issues were not discussed in the MMP.

R. Herring, at el. [3] proposed a statistical learning model to estimate traffic state in terms of *travel time* and *congestion state* considering low penetration rate. As claimed, their model works effectively even if the penetration rate is as low as *5%*. However, several issues are remaining that need to be thoroughly discussed. **First**, this work did not mention the effect of the traffic flow density on traffic state. **Second**, the congestion state was defined as a "binary" indicator which accepts only two states, namely *"congested"* and *"not-congested"*. Obviously, this setting biases the estimation accuracy since even the "blind" guessing approach also has an opportunity to reach *50%* of accuracy. **Third**, this work employed the Paramics simulator [15] to generate synthetic data which gives information about every vehicle. To imitate a low penetration rate dataset, namely *5%* for example, a large portion of data *(95%)* was removed and only a subset of *5%* of data was kept. In fact, this process could not generate the appropriate low penetration rate dataset as it was defined in their work. Therefore, the relation between the estimation error and the penetration rate obtained in that research should be clarified.

Our previous work in [9] has proposed a notable traffic state quantification model by which less informative traffic data reported by mobile devices can be effectively processed to granularly quantify traffic state levels. This work focuses on solving the inherent issues of **low** and **uncertain** penetration rate, that is to say, one of the most essential issues in mobile phone based applications such as the M-TES.

3 PROBLEM FORMULATION

This section presents the problem formulation, namely estimating traffic state based on only the GPS data reported by mobile phones considering **low** and **uncertain** penetration rate issues.

3.1 Traffic State Estimation in M-TES

Traffic characteristics commonly vary from road segments to road segments, thus traffic state should be estimated based on a road segment basis. Considering a road network of N road segments, the set of all road segments is denoted as $V=\{i|i = 1..N\}$. For any road segment i, traffic data (GPS data) is available at any time t. However, this data is the event-based data which cannot be directly transformed into traffic state. Therefore, traffic state should be aggregated in predefined time intervals, namely in t-second windows, such as in each minute, for example. Concretely, traffic state is estimated at times $k = 0, t, 2t,...$, where t is the aggregation time mentioned above.

Obviously, *velocity* and *density* of a traffic flow directly reflect traffic condition of a road segment. Therefore, these factors should be estimated independently using the data reported by mobile phones before integrating for traffic state estimation.

Definition 1: The *average velocity* of the traffic flow in the road segment i during time k, denoted as $V_{Avg}^{k,i}$, is the average velocity of all vehicles travelling in the considered road segment and is calculated in equation *(1)*.

$$V_{Avg}^{k,i} = \frac{\sum_{j=1}^{q} V_{t_m,j}^{k,i}}{qr}, (k-1)t \le t_m < kt \qquad (1)$$

Here, $V_{t_m,j}^{k,i}$ is the velocity of any individual vehicle j $(j = 1...q)$ detected at time t_m $(m = 1, 2,..., r)$ during time interval k $([k-1]t \le t_m < kt)$, q is the total number of vehicles, and r is the total number of detection times during time interval k.

Definition 2: The *density* of the traffic flow in the road segment i during time k, denoted as $D^{k,i}$, is defined in equation *(2)*. Where, $q^{k,i}$ is the total number of vehicles travelling through the road segment i during time k, and $C^{k,i}$ is the *maximum capacity* of the road segment i during time k.

$$D^{k,i} = \frac{q^{k,i}}{C^{k,i}} \qquad (2)$$

Traffic state of the considered road segment i during time k is estimated based on velocity ($V_{Avg}^{k,i}$) and density ($D^{k,i}$). Therefore, any error in estimating velocity or density will be propagated to the error of traffic state estimation. This section investigates the effect of low penetration rate on velocity and density estimations.

3.2 Effect of the Low Penetration Rate

In the M-TES, traffic state is estimated based on the data reported by mobile phones carried by vehicles. In practice, it is impossible to compel every mobile phone user to report data to the estimation server. Therefore, the penetration rate is commonly low declining the estimation effectiveness. The formal definition of the penetration rate and its effect on the M-TES's accuracy are as follows:

Definition 3: The *penetration rate* at the road segment *i* during time *k*, denoted as $\rho^{k,i}$, is defined in equation *(3)* and illustrated in Fig.1. Here, *p* is the number of vehicles that report data to the estimation server, and *q* is the total number of vehicles travelling in the road segment *i* during time *k*.

$$\rho^{k,i} = \frac{p}{q} \tag{3}$$

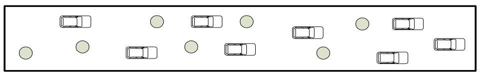


Figure 1. Vehicles that report data are denoted as the car-shape ones, the penetration rate here is $\rho^{k,i} = p/q = 8/15$

With a given penetration rate $\rho^{k,i}$, the average velocity estimation model described in equation *(1)* must be replaced with equation *(4)*. Here, $V^{k,i}_{t_m,j}$ is the velocity of an individual vehicle *j* *(j = 1..q)* detected at time t_m *(m = 1, 2,...r)* during time interval *k* *([k-1]t ≤ t_m < kt)*, *q* is the total number of vehicles traveling in the considered road segment, and *r* is the total data reporting times during time *k*.

$$V^{k,i,\rho^{k,i}}_{Avg} = \frac{\sum_{j=1}^{\rho^{k,i}q} V^{k,i}_{t_m,j}}{\rho^{k,i}qr}, (k-1)t \le t_m < kt \tag{4}$$

Under this condition of penetration rate, the average velocity estimation error, $E^{k,i}_V$, can be expressed in equation *(5)*, where $V^{k,i}_{Avg}$ is the "actual" average velocity estimated when every vehicle reports data to the estimation server (equation *(1)*), and $V^{k,i,\rho^{k,i}}_{Avg}$ is the average velocity estimated under the given penetration rate $\rho^{k,i}$ (equation *(4)*).

$$E^{k,i}_V = \left|1 - \frac{V^{k,i,\rho^{k,i}}_{Avg}}{V^{k,i}_{Avg}}\right| = \left|1 - \frac{1}{\rho^{k,i}} \cdot \frac{\sum_{j=1}^{\rho^{k,i}q} V^{k,i}_{t_m,j}}{\sum_{j=1}^{q} V^{k,i}_{t_m,j}}\right| \tag{5}$$

$$E^{k,i}_D = (1-\rho^{k,i}) * 100\% \tag{6}$$

Similar to average velocity, density estimation is also affected by the penetration rate. According to the density definition described in equation *(2)*, the density estimation error, denoted as $E^{k,i}_D$ in equation *(6)*, is directly affected by $\rho^{k,i}$. Obviously, if $\rho^{k,i}$ is *20%*, $E^{k,i}_D$ is as large as *80%*, an unacceptable error for any estimation model. Meanwhile, the penetration rate of *20%* or lower is usual in practice.

Figure 2 shows the effect of penetration rate on velocity and density estimation accuracy. Here, *D_err*, representing $E^{k,i}_D$, was calculated directly from equation *(6)*, and *V_err*, representing $E^{k,i}_V$, was obtained using simulation data. This figure reveals that the density estimation is more seriously, compared to that of the velocity estimation, affected by low penetration rate.

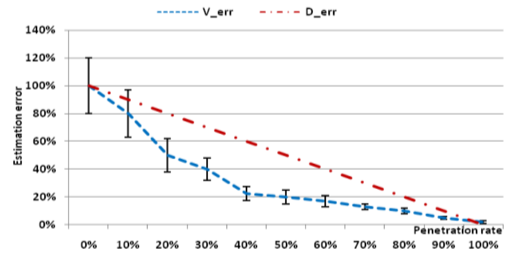


Figure 2. Effect of penetration rate on velocity and density estimations

4 GA-BASED VELOCITY-DENSITY INFERENCE CIRCUIT

In order to alleviate errors rooted from low penetration rate mentioned in the previous section, a novel genetic algorithm (GA) [17] based velocity-density inference circuit is proposed. This model is depicted in Fig.3, where both the *velocity* and *density* calculated directly from the *sensed* data collected by mobile phones, namely $V^{k,i}_{sensed}$ and $D^{k,i}_{sensed}$, are served as the primary inputs. The outputs of the circuit are the final estimated velocity and density, namely $V^{k,i}_{est}$ and $D^{k,i}_{est}$. The inferred velocity and density, namely $V^{k,i}_{infer}$, $D^{k,i}_{infer}$, obtained by applying the Greenshields model [18] are also taken into account. In addition, moving average values of estimated velocity and density at time *k*, namely $MV^{k,i}$ and $MD^{k,i}$, calculated using the corresponding values estimated in the previous phases are *fed* back to the estimation model. The GA component provides an optimal coefficient *g* motivating optimal estimations.

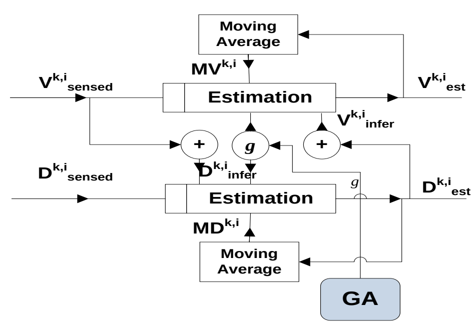


Figure 3. The optimal velocity-density inference model circuit

The philosophies behind this inference model are as follows: *1) Velocity* and *density* calculated directly and independently from *sensed* data help to void any error propagation. *2)* The Greenshields model [18] used to infer density from estimated velocity and vice versa, can help to avoid the over-error of density estimation when penetration rate is unacceptably low. *3)* Current traffic state has inherent relationships with previous traffic states at the same road segment. *4)* All of the estimation approaches (direct estimation using *sensed* data, inference using the Greenshields model, inference using the previous estimated data) may uphold their advantages while diminishing their inherent disadvantages if being appropriately integrated.

The overall velocity-density inference model is formally presented in equations *(7)* and *(8)*, where α, β, γ are the impact coefficients of the corresponding parameters. It should be noted that α, β, γ are encapsulated in a simplified parameter g, namely $g = \{\alpha, \beta, \gamma\}$, presented in Fig.3.

$$V_{est}^{k,i} = \alpha V_{sensed}^{k,i} + \beta MV^{k,i} + \gamma V_{infer}^{k,i} \quad (7)$$

$$D_{est}^{k,i} = \gamma D_{sensed}^{k,i} + \beta MD^{k,i} + \alpha D_{infer}^{k,i} \quad (8)$$

The moving averages and the immediate inferred velocity and density at time k, namely $MV^{k,i}$, $MD^{k,i}$, $V^{k,i}_{infer}$ and $D^{k,i}_{infer}$, respectively, are computed in equations *(9)*, *(10)*, *(11)*, and *(12)*. Here, D^{i}_{max} and V^{i}_{max} are the maximum density and the limited velocity of the road segment i; and ξ is the sliding window for moving average calculations which can be set by domain experts or by using simulation data.

$$MV^{k,i} = \frac{\sum_{j=k-\xi}^{k-1} V_{est}^{j,i}}{\xi} \quad (9)$$

$$MD^{k,i} = \frac{\sum_{j=k-\xi}^{k-1} D_{est}^{j,i}}{\xi} \quad (10)$$

$$D_{infer}^{k,i} = D_{\max}^{i}\left(1 - \frac{V_{sensed}^{k,i}}{V_{\max}^{i}}\right) \quad (11)$$

$$V_{infer}^{k,i} = V_{\max}^{i}\left(1 - \frac{D_{est}^{k,i}}{D_{\max}^{i}}\right) \quad (12)$$

It should be noted that $D^{k,i}_{infer}$ is inferred directly from $V^{k,i}_{sensed}$ (equation *(11)*), while $V^{k,i}_{infer}$ is inferred from the final estimated density, $D^{k,i}_{est}$ (not from $D^{k,i}_{sensed}$ while this factor is available). This inference policy helps to improve the overall estimation effectiveness. The reasons behind this design are as follows: The accuracy of $V^{k,i}_{sensed}$ is commonly better than that of $D^{k,i}_{sensed}$, as shown in Fig. 3, thus $D^{k,i}_{infer}$ is chosen to be inferred first. When $D^{k,i}_{infer}$ is available, $D^{k,i}_{est}$ can be obtained in equation *(8)*. At this time, both $D^{k,i}_{sensed}$ and $D^{k,i}_{est}$ are available for inferring $V^{k,i}_{infer}$. Obviously, the accuracy of $D^{k,i}_{est}$ must be better than that of the $D^{k,i}_{sensed}$, hence $D^{k,i}_{est}$ is employed to infer $V^{k,i}_{infer}$ in equation *(12)*. The estimation flow is as follow, where D_{est} and V_{est} are target variables:

$$V_{sensed} \rightarrow D_{infer} \rightarrow \boldsymbol{D_{est}} \rightarrow V_{infer} \rightarrow \boldsymbol{V_{est}}$$

The most important key contributing to the estimation model described above is the optimization of coefficients α, β, γ. In this work, an appropriate GA-based mechanism is proposed to optimize the coefficients α, β, γ leading to optimize the whole estimation model. Since α, β, γ are real-value numbers ranging in [0, 1], the proposed GA mechanism must have the ability of working with chromosomes (solutions) modeled by real-value numbers [19]. More concretely, the schema of a chromosome in the GA mechanism is coded as $g = \{\alpha, \beta, \gamma\}$ (the chromosome of *3* genes). The proposed GA mechanism is depicted as the pseudo code in Fig.4.

```
Generating the initial population, P, randomly;
count = 0; Fitness = 0; n = expected_iteration;
While (Fitness < threshold) && (count < n){
     Selecting ();
     Mating ();
     Crossover();
     Mutation();
     // P is updated
     for each g in P{
          Fitness = max(Fitness, g.fitness());
     }
     count++;
}
Output the best chromosome g whose fitness is highest
```

Figure 4. The pseudo code of the Genetic Algorithm

The most important point in applying the GA technique to a particular issue is to propose an appropriate *fitness function* for the selection procedure. Here, candidates are selected based on their *competitiveness* evaluated using the *fitness function* proposed in equation *(13)*.

$$f(g_i) = \frac{e(g_i)}{\bar{e}(g_j \mid g_j \in population)} \quad (13)$$

Here, $e(g_i)$ is the evaluation of candidate g_i ($g_i = \{\alpha_i, \beta_i, \gamma_i\}$), which is the estimation error caused by choosing g_i as the set of coefficient*;* and $\bar{e}(g_j)$ is the average evaluation of all individuals g_j in the current population. The evaluation $e(g_i)$ is defined in equation *(14)*, where V_{est_gi} is the estimated velocity (equation *(7)*) with the set of coefficient g_i, and V_{act} is the *"actual"* velocity.

$$e(g_i) = \frac{|V_{est\_gi} - V_{act}|}{V_{act}} \quad (14)$$

After obtaining the fitness of all chromosomes in a population, the GA selects only the ones whose *fitness* values $f(g_i)$ are relevant to the crossover process. The selection criterion is the minimal of the estimation error, $e(g_i)$, leading to the minimal of the *fitness* value, $f(g_i)$. That is to say, the candidates with smaller *fitness* values have the higher probability to be selected to the crossover process.

5 TRAFFIC STATE PREDICTION UNDER UNACCEPTABLY LOW PENETRATION RATE

In general, the *GA-based velocity-density inference* model proposed in section 4 improves the M-TES accuracy significantly. However, when the penetration rate becomes unacceptably low, namely just several percent or even zero, the proposed GA model cannot work properly. To address this issue, a prediction model should be proposed to predict traffic state of unacceptably low penetration rate road segments.

Obviously, traffic state of a road segment is affected by traffic state of the nearby road segments. In addition, the current traffic state of each road segment has a close relation with its previous states. If these *spatial-temporal* relations (rules) are known in advance, traffic state of the considered road segment can be predicted. These rules can be learned from historical traffic state data by any machine learning technique. In this work, an appropriate neural network (ANN) with multilayer perceptron (MLP) [20] is proposed to predict the average velocity and density of considered road segments when their penetration rates become **unacceptably low** or **unknown**.

5.1 The Overview of the ANN

An artificial neural network (ANN) is a computational model that is inspired by the structural and/or functional aspects of biological neural networks [20]. The basic model of an artificial neuron is founded upon the functionality of the biological neuron. Each neuron has a computational function (the cell body) to process and threshold (i.e. mapping the output data into a range of relevant/meaningful values such as the range of [0, 1], for example) the coming signals (i.e. the input data). The output data of a neuron is sent to other neurons for further process. This data is accompanied with a weight representing the strength (representing the biological synapse) of the connection between any two neurons. In the computational concept, this weight represents the impact factor of the mentioned data in the next computation taken place at the later neuron. Connection of neurons results in a neural network by which the output data at the "final" neuron(s) can represent some kinds of knowledge. Commonly, the ANN is suitable for classification and prediction issues. An artificial neuron is predicted in Fig.5.

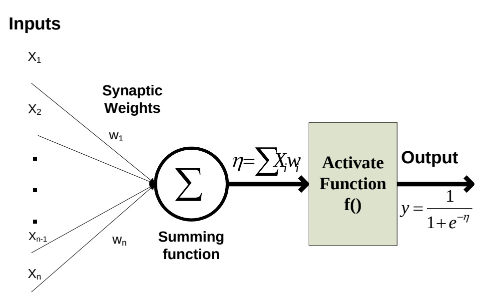


Figure 5. Modeling an artificial neuron

Figure 5 shows that a node (the neuron's cell body) is fed by several data via different entries (dendrites). Each data is combined with a particular weight representing the biological synapse. A negative weight reflects an inhibitory connection, while positive values designate excitatory connections. All inputs are summed altogether and modified by the weights. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1. The neuron output is calculated as $y = f(\eta)$, where η is the interval activity of the neuron identified by the summing function of $X_i w_i$ (the multiplication of input X_i and corresponding synaptic weight w_i). The particular activate function can be decided by the designer [20].

5.2 The Proposed ANN-based Prediction Model

In this work, the *spatial-temporal* relations between traffic states of road segments which are closed together are taken into account for training the ANN model. Figure 6 illustrates the proposed ANN model. Let *i* is the considered road segment whose penetration rate is unknown and its traffic state is needed to be predict. Obviously, the velocity and density of the road segment *i* at time interval *k* has some relations with those of any road segment *j* ($j \in V$) in the same region at the same time interval *k*. Therefore, the velocities and densities at time *k*, of any related road segment *j*, denoted as $V^{k,j}$, $D^{k,j}$, are fed to the ANN model for training the aforementioned *spatial* relation between road segment *i* and its related road segments.

In addition, traffic state is a complicate phenomenon that does not change frequently in terms of time. Concretely, given a road segment *i*, its traffic state at time *k* may have some relations (*temporal* relations) with its previous traffic states, namely at any time *t*, where $t < k$. Therefore, the proposed ANN model is also fed with the previous velocity and density of the considered road segment *i*, namely $V^{t,i}$, $D^{t,i}$ (where $t < k$), respectively.

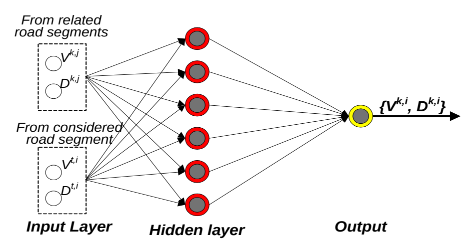


Figure 6. The ANN-based prediction model dealing with unacceptably low penetration rate

The proposed ANN-based prediction model can be formulated in equation *(15)*. The velocity and density of the considered road segment *i* at time *k*, denoted as $V^{k,i}_{pre}$, $D^{k,i}_{pre}$, are predicted by the so-called *predict()* function, where current traffic state of related road segments ($V^{k,j}$, $D^{k,j}$) and the previous traffic state of the considered road segment ($V^{t,i}$, $D^{t,i}$) mentioned above are served as the input parameters.

$$\{V^{k,i}_{pre}, D^{k,i}_{pre}\} = predict(V^{k,j}, D^{k,j}, V^{t,i}, D^{t,i}) : t < k, j \neq i, j \in V \quad (15)$$

6 SELECTING AN APPROPRIATE ESTIMATION MODEL

The ANN-based prediction approach proposed in the previous section can cope with the *unacceptably low* and *uncertain* penetration rate issues. However, since this model does not take the real-time traffic data into account but relies on only historical data, its effectiveness can not be improved when the penetration rate is relevant (large enough). In this case, the GA-based mechanism proposed in section 4 becomes prominent. Therefore, choosing the **"right"** traffic estimation model among two candidates above under the condition of **uncertain** penetration rate is an essential but challenging issue. This section proposes a notable method for selecting appropriate estimation model.

Let denote V_{act} the "actual" velocity which is represented by the solid line in Fig.7. It should be noted that V_{act} is unknown at the estimation time. Let denote V_{ann}, V_{cir} the velocities estimated using the ANN-based prediction (denoted as Ann) and the GA-based velocity-density inference (denoted as Cir) models, respectively. The task is to approximate the estimated velocity, namely V_{est}, based on V_{ann} and V_{cir} so that the difference between V_{est} and V_{act} is minimal.

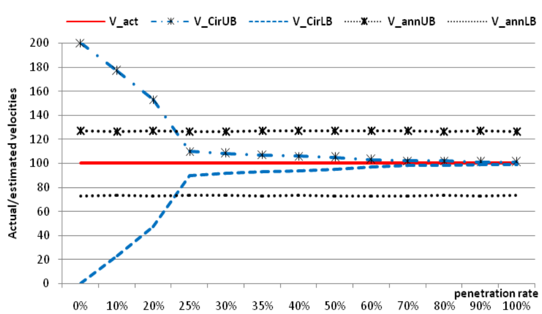


Figure 7. The relationship between "actual" and estimated velocities considering different penetration rates

As mentioned in the beginning of this section, the estimation error in the *Ann* method is stable regardless of penetration rate, the V_{ann} lies on one of the 2 straight dotted lines representing $V_{\_annUB}$ and $V_{\_annLB}$, respectively, where $V_{\_annLB} < V_{act} < V_{\_annUB}$. These two lines are paralleling with the solid line representing V_{act}. Different with *Ann* method, the *Cir* model is affected by the penetration rate, that is to say, the higher penetration rate is, the lower the error is. This relationship is represented by the error estimation curve depicted in Fig.3, discussed in section 3.2. Therefore, the V_{cir} lines on one of the 2 curves representing $V_{\_cirUB}$ and $V_{\_cirLB}$, where $V_{\_cirLB} < V_{act} < V_{\_cirUB}$. The relationship between V_{act}, $V_{\_annUB}$, $V_{\_annLB}$, $V_{\_annUB}$, and $V_{\_annLB}$ is illustrated in Fig.7.

As shown, the gap between V_{ann} (namely $V_{\_annUB}$ or $V_{\_annLB}$) and V_{act} is kept as a constant regardless of the penetration rate. In contrast, the gap between V_{ann} and V_{cir} drastically increases when the penetration rate becomes lower than the critical one (the place where V_{ann} and V_{cir} cross each other). This gap is narrow and stable when the penetration rate is relevant (larger than the critical one). Therefore, this relationship can be utilized to estimate V_{est}. Here, the estimation model for estimating V_{est} can be appropriately selected as described in equation *(16)*, where $\bar{V}$ is defined in equation *(17)*.

$$V_{est} = \begin{cases} V_{ann}, if \dfrac{|V_{cir} - V_{ann}|}{\bar{V}} \geq \theta\% \\ V_{cir}, other\_case \end{cases} \quad (16)$$

$$\bar{V} = \frac{|V_{ann} + V_{cir}|}{2} \quad (17)$$

Equation *(17)* represents that if the gap between V_{ann} and V_{cir} is larger than a threshold value, namely θ%, then V_{est} should be V_{ann} since the penetration rate is low. In other cases (relevant penetration rates), V_{est} should be approximated as V_{cir}. The threshold θ can be determined by experimental data as shown in the evaluation section.

7 EVALUATION

This section evaluates the effectiveness of the proposed solutions to **low** and **uncertain** penetration rate issues. Concretely, the effectiveness of the GA-based velocity-density inference mechanism (denoted as *Cir* mechanism), of the ANN-based prediction approach (denoted as the *Ann* model), and of the selection method for selecting an appropriate traffic state estimation model is evaluated.

7.1 The Experiment Environment

In this work, the TSF simulator [16] was utilized to generate synthetic data for evaluations. Different road segments were selected randomly as shown in Fig.8. For each selected road segment, different penetration rates, namely 20%, 30%, 40%,…, and different levels of density were configured at different simulations. In each simulation, two types of data were created concurrently as follows:

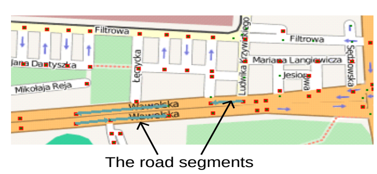


Figure 8. Road segmentation in the TSF

a) The GPS data reported by individual vehicles were recorded. Each record contains the *Time stamp* (in seconds), the *road segment Id*, the *position* (longitude, latitude), the *current velocity*, and the *vehicle Id* of the vehicle that reports data. The frequency of the data report timing was set to every *3s* (similar to the common GPS signal frequency).

b) The summarized traffic state information of the selected road segments was also recorded. Each record contains the information of the *Time interval Id* (in minutes), the *road segment Id*, the *average velocity*, and the *density*. The time interval for recoding the summarized traffic state data was set to every minute. This data was used to evaluate the accuracy of the proposed estimation methods applying the GPS data described in *a)*.

7.2 Effectiveness of the GA-based Mechanism and the ANN-based Prediction Approach

Figure 9 represents the effectiveness of the proposed GA velocity-density inference circuit (*Cir*) compared to the conventional estimation model in terms of estimating average velocity and density. The velocity and density estimation errors of the conventional method where the *sensed* data is applied directly are denoted as *Normal_V* and *Normal_D* while those corresponding errors of the *Cir* model are denoted as *GA_Circuit_V* and *GA_Circuit_D*, respectively. As shown, the proposed *Cir* model is prominent. Especially, when the penetration rate is relevant, namely larger than 25%, the estimation error in both velocity and density estimation of the proposed *Cir* model is tiny as lower than 5%.

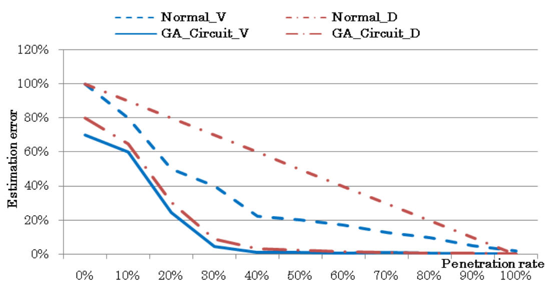
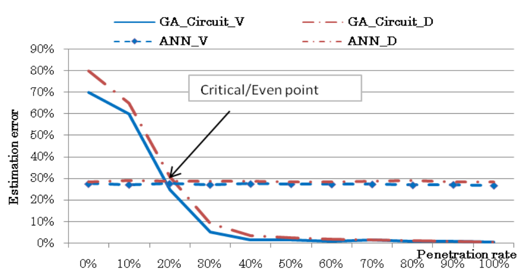


Figure 9. Effectiveness of the GA-based velocity-density inference



cricuit (Cir)

Figure 10. Effectiveness of the ANN-based prediction model (Ann) v.s. that of the GA-based velocity-denisty inference circuit (Cir)

Figure 10 shows the effectiveness of the ANN-based prediction method (*Ann*) compared to that of the *Cir* model regarding different penetration rates. Here, the errors of both the velocity and density, denoted as *ANN_V* and *ANN_D*, respectively, of the *Ann* method are around *27%* regardless of penetration rate. The accuracy of the *Ann* model is completely satisfies this research purpose when the penetration rate is unacceptably low, namely lower than the *critical* one (i.e. around *25%*). However, when the penetration rate is relevant, namely larger than the *critical* one, the *Cir* approach (denoted as *GA_Circuit_V* and *GA_Circuit_D* for the velocity and density estimation errors, respectively) is dominant.

7.3 Effectiveness of the Traffic State Estimation Selection Method

This section evaluates the proposed method on selecting the appropriate traffic state estimation model under the condition of **uncertain** penetration rate.

Figure 11 shows the effectiveness of the appropriate selection method proposed in section 5 regarding different decision thresholds (i.e. the values of θ - please refer to equation *(16)*). The figure reveals that the lower threshold supports for the selection method in the cases of low penetration rates, and vice versa. For example, with the threshold of 30% (θ=30%), the selection accuracy is almost higher than 80% if the penetration rate is low (≤20%). In contrast, the accuracy decreases drastically (≤40%) when the penetration rate is relevant (≥25%). If a completely high threshold, namely θ=50%, is selected, the selection accuracy is high when the penetration rate is relevant while the accuracy is drastically decreased when the penetration rate is low. Figure 11, also shows that the model works well (with high accuracy) in both the high or low penetration rate when the threshold is set to 40%. In this case, the accuracy is quite high (around 80% or 90%) in the cases of relevant penetration rates, while accuracy in the cases of low penetration rates is still high enough (around 60%).

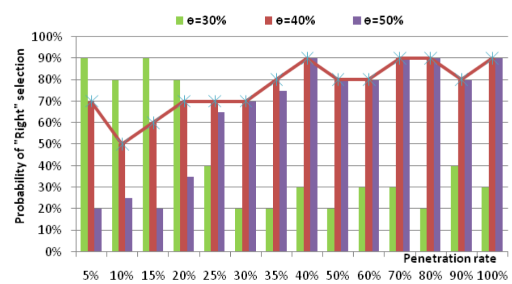


Figure 11. Effectiveness of the proposed method for selecting the "Right" traffic state estimation model

Table 1 summarizes the model selection effectiveness with regards to different thresholds. This table show the probability of **"right"** selection in high penetration rate (>25%), low penetration rate (≤25%) and the summary of the selection accuracy in all the cases. As shown, for a low threshold (θ=30%), the selection accuracy is high (76%) in cases of low penetration rates, but the accuracy declines drastically (26.67%) in cases of high penetration rate. The overall accuracy in this case is 44.29%. In contrast, for a high penetration rate (θ=50%), the relation between penetration rate and the selection accuracy is in the converse way. The model is optimal with the threshold of 40% where it works well for both the low and high accuracy resulting in a high selection accuracy of 76.43%. Therefore, if the selection method is applied in real world applications, the selection threshold should be 40%.

Penetration rate	**Selection thresholds and accuracy (%)**		
	θ =30%	θ =40%	θ =50%
All the case	44.29%	76.43%	65%
<=25%	76.00%	64.00%	33.00%
>25%	26.67%	83.33%	82.78%

Table 1. The average accuracy of the selection method

8 CONCLUSION AND FUTURE WORK

Novel solutions to essential issues of **low** and **uncertain** penetration rate in mobile phone based applications, especially in the M-TES, were proposed. More concretely, a unique GA-based velocity-density inference circuit (*Cir*) and a notable ANN-based prediction model (*Ann*) were proposed. The effectiveness of the proposed solutions was thoroughly evaluated using large amount of simulation data. The evaluation results show that the proposed approaches are effective and robust. The GA-based mechanism improves the estimation accuracy significantly even in cases of low penetration rates improving the scalability of the M-TES. The ANN-based prediction model makes the M-TES be viable even if the penetration rate becomes unacceptably low or even unknown.

In practice, the actual penetration rate cannot be accurately measured at the estimation time, hence it is challenging to select the **"right"** estimation model among the GA-based and the ANN-based prediction approaches. This research proposed practical selection methods for selecting appropriate traffic state estimation model. The proposed selection model reveals its robustness and feasibility since it help to select the **"right"** traffic state estimation model automatically even the "actual" penetration rate is unknown.

This proposed selection method, however, still requires information about the estimated values, namely V_{ann} and V_{cir}, which may decline the computational performance. Therefore, finding an appropriate approach by which appropriate estimation method can be selected correctly without any requirement of prior estimations is an interesting research direction. In addition, the effectiveness as well as the robustness of the proposed solutions should be confirmed by more real-field experiments before being applied to real world applications.

REFERENCES

[1] [1] N. Shibata, T. Terauchi, T. Kitani, K.Yasumoto, M. Ito, T. Higashishino, *"A Method for Sharing Traffic jam Inforamtion Using Inter-Vehicle Communication,"* mobiquitous, 3rd Annual International Conference on Mobile and Ubiquitous Systems – Workshops, pp. 1-7, 2006.

[2] TTI, Texas Transportation Institute: Urban Mobility Information: 2007 Annual Urban Mobility Report. http://mobility.tamu.edu/ums/.

[3] R. Herring, A. Hofeitner, S. Amin, T. A. Nasr, A. A. Khalek, P. Abbeel, A. Bayen *"Using Mobile Phones to Forecast Arterial Traffic through Statistical Learning,"* The 89th Annual Meeting of the Transportation Research Board, Aug. 2009.

[4] S. Tang and F.-Y. Wang, *"A PCI-based evaluation method for level of services for traffic operational systems,"* IEEE Trans. Intell. Transp. Syst.,vol. 7, no. 4, pp. 494–499, Dec. 2006.

[5] J. Wright and J. Dahlgren, *"Using vehicles equipped with toll tags as probes for providing travel times."* California PATH Working Paper UCB-ITSPWP-2001-13, Institute of Transportation Studies, University of California, Berkeley, CA, 2001.

[6] Y. Cho and J. Rice, *"Estimating velocity fields on a freeway from lowresolution videos,"* IEEE Trans. Intell. Transp. Syst., vol. 7, no. 4, pp. 463–469, Dec. 2006.

[7] J. Yoon, B. Noble, and M. Liu *"Surface Street Traffic Estimation"* MobiSys ' 07, San Juan, Puerto Rico, USA, June 11-14, 2007.

[8] K. Sohn and K. Hwang, *"Space-based passing time estimation on a freeway using cell phones as traffic probes,"* IEEE Trans. Intell. Transp. Syst., vol. 9, no. 3, pp. 559–568, Sep. 2008.

[9] T.M. Quang, E. Kamioka, *"Traffic State Estimation with Mobile Phones Based on The "3R" Philosophy,"* IEICE Transactions on Communications, Vol.E94-B, No.12, pp. 3447-3458, Dec. 2011.

[10] T. M. Quang and E. Kamioka: "*Adaptive Approaches in Mobile Phone Based Traffic State Estimation with Low Penetration Rate,"* Journal of Information Processing (IPSJ) Vol.20, No.1, pp. 297-307, Jan. 2012.

[11] Ministry of Land Infrastructure and Transport: *"The System Outline of VICS,"* http://www.vics.or.jp.

[12] Route search and location based consumer services in Japan http://www.navitime.co.jp.

[13] Kansas ITS: http://www.ksdot.org/burtransplan/burovr/inttrans.asp.

[14] http://traffic.berkeley.edu/theproject.html.

[15] Paramics Microsimulation Software for the simulation and visualisation of road traffic. http://www.paramics.com.

[16] P. Gora. *"Traffic Simulation Framework — a Cellular Automaton-Based Tool for Simulating and Investigating Real City Traffic,"* Recent Advances in Intelligent Information Systems ISBN 978-83-60434-59-8, pp. 641–653, 2009.

[17] Z. Michaelewicz, *"Genetic algorithms + data structures = evolution programs,"* (3rd ed.), Springer-Verlag, Berlin, 1996.

[18] B. D. Greenshields "*A Study in Highway Capacity,"* Highway Research Board Proceedings, pp. 448-477, 1935.

[19] A.H. Wright, *"Genetic algorithms for real parameter optimization,"* In: J.E. Rawlins, Editor, Foundations of genetic algorithms, Morgan Kaufmann, pp. 205–218, 1991.

[20] D. E. Rumelhart, B. Widrow, M. A. Lehr, "*The basic ideas in neural networks*," Communications of the ACM, v.37 n.3, pp.87-92, Mar. 1994.