Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 2546 - 2560, 2005

MISSING DATA TREATMENT AND DATA FUSION TOWARD TRAVEL TIME ESTIMATION FOR ATIS

Yuh-Horng WEN
Research Assistant Professor
ITS Research Center
Department of Electrical and Control Engineering
National Chiao Tung University
Hsinchu 30010, Taiwan, R.O.C.
Fax: +886-3-5729749
E-mail: yhwen@faculty.nctu.edu.tw

Tsu-Tian LEE
Professor
Department of Electrical and Control Engineering
National Chiao Tung University
President of National Taipei University of Technology
Taipei 10608, Taiwan, R.O.C.
Fax: :+886-2-27518845
E-mail: ttlee@cn.nctu.edu.tw

Hsun-Jung CHO
Professor
Department of Transportation Technology and Management
National Chiao Tung University
Hsinchu 30010, Taiwan, R.O.C.
Fax: +886-3-5720844
E-mail: hjcho@cc.nctu.edu.tw

Abstract: This study develops a travel time estimation process by integrating a missing data treatment and data-fusion-based approaches. In missing data treatment, this study develops a grey time-series model and a grey-theory-based pseudo-nearest-neighbor method to recover, respectively, temporal and spatial missing values in traffic detector data sets. Both spatial and temporal patterns of traffic data are also considered in travel time data fusion. In travel time data fusion, this study presents a speed-based link travel time extrapolation model for analytical travel time estimation and further develops a recurrent neural network (RNN) integrated with grey models for real-time travel time estimation. In the case study, field data from the national freeway no.1 in Taiwan is used as a case study for testing the proposed models. Study results showed that the grey-theory-based missing data treatment models were accurate for recovering missing values. The grey-based RNN models were capable of accurately predicting travel times. Consequently, the results of this study indicated that the proposed missing data treatment and data fusion approaches can ensure the accuracy of travel time estimation with incomplete data sets, and are therefore suited to implementation for ATIS.

Key Words: missing data treatment, data fusion, travel time estimation, grey and neural network models, ATIS

1. INTRODUCTION

Research into travel time estimation continues to attract the attention of Intelligent Transportation Systems (ITS) academicians and engineers, since accurate travel time estimation is essential for successful implementation of an Advanced Traveler Information Systems (ATIS). Many ITS studies and transportation agencies use the traffic data from

dual-loop detectors, which are readily available in many locales of freeways and urban roadways. A dual-loop detector consists of two consecutive single inductance loops spaced a few feet apart. In the dual-loop system, when the first single inductance loop detects a vehicle, a timer is started. The timer runs until the same vehicle is detected at the second single inductance loop. The distance between the two single-loop detectors is divided by the elapsed time and converted to miles per hour to calculate vehicle speed. The speed, length, and volume data are aggregated into 20-second intervals for outputs. Dual-loop detector systems are capable of archiving with traffic count (i.e., the number of vehicles that pass over the detector in that period of time), velocity and occupancy (i.e., the fraction of time that vehicles are detected).

A number of methods for estimating travel time have previously been proposed and developed, such as, Dailey (1993), Cremer and Schutt (1990), Nam and Drew (1996). Most of these models used the flow and occupancy measured by the loop detectors to calculate the speed at that point and extrapolated to get the link travel time. However, the authors found that link travel time differed significantly from the quotient of local velocity and the link distance. Since the link travel time for vehicles reflects traffic conditions averaged over a fixed distance over a variable amount of time, while the detector data only reflect traffic conditions averaged over a fixed time period at a single point in space. Other approaches have included local speed-based estimation process to yield accurate link travel time estimates from point data (Coifman, 2002; Oh et al., 2002; Lint and Zijpp, 2003). Coifman (2002) exploited traffic flow theory (triangular fundamental speed-flow relationship) to extrapolate local speeds to an extended link. The author developed an analytical model to estimate vehicle trajectories and then estimate link travel times. Lint and Zijpp (2003) proposed a piecewise linear local speed based trajectory method, and the speed estimation equations can then be derived the trajectories between two detectors.

Rapid forecasting of travel time with reasonable accuracy becomes an emerging issue towards travel time estimation. More and more studies focused on real-time data fusion approaches for travel time estimation. Data fusion is concerned with the problem of combining data and information coming from different sources into a single database when variables are absent or missing in some data sets (Saporta, 2002). ADVANCE (Advanced Driver and Vehicle Advisory Navigation Concept), one of the ITS projects in Chicago, have proposed a series of studies using neural networks and fuzzy neural networks for travel time prediction (Nelson and Palacharla, 1993; Palacharla and Nelson, 1999). Recently, Dharia and Adeli (2003) used counter propagation neural network for travel time forecasting. Lint et al. (2002, 2003) developed the state-space neural networks for travel time prediction, wherein the accuracy, robustness, and uncertainty were also discussed. However, raw real-time traffic data from loop detectors are not ready to be processed by numerical and data fusion tools for travel time estimation because they may contain a lot of missing data items. Real-time traffic data from roadside loop detectors are inevitably corrupted by unexpected missing values or appear to be giving nonsensical or erroneous data due to detector faults or transmission distortion. In most traffic data sets, missing values occurred at some time-spots of the traffic data time-series, due to temporary power or communication failures in the traffic surveillance system. Sometimes missing values inevitably occurred in a whole series of detector data from some failed detectors. Missing data treatment is an important preparation step for data-fusion tasks in travel time estimation, since inappropriate treatment of missing data may cause large errors and impinge upon the quality of estimated travel time information. Missing data should be pre-processed (recovered) so as to allow the whole or partial data set to be processed by a data fusion tool. Specifically, in addition to accurate travel time

estimation, an understanding of the quantitative effects of missing data on travel-time estimating performance is therefore also needed.

Missing data problems have been studied in many areas, such as speed recognition, image processing, system identification, and to a limited extent, traffic forecasting (Redfern et al., 1993; Chen et al., 2001). Few studies on travel time estimation concern the missing data recovery issues. Many studies just ignored or used naïve imputation models, such as exponential smoothing and linear interpolation, to recover missing values for travel time estimation (Lint et al., 2002, 2003). Lint et al. (2002) stated that more sophisticated approaches should be further studied and applying the travel time estimation framework in real time. Past studies on data imputation used certain means of approximation to fill in missing values, and the resultant completed data are analyzed by standard statistical analytical methods. Many imputation-based procedures were developed for recovering missing data, such as semi-parametric methods (Lawless and Kalbfleisch, 1999), Bayesian methods (Fitzgerald, 1999), the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), multiple imputation (Schafer, 1997), and Monte Carlo techniques (Gelfand and Smith, 1990). Sande (1996) used hot deck imputation, where recorded units in the sample are substituted by a value obtained from the nearest data record. Pawlak (1993) used regression imputation, where the missing variables for a unit are estimated by values derived from the known variables according to a given function. Huang and Zhu (2002) presented a pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets, where missing values are substituted by corresponding values of the pseudo-nearest-neighbors with the largest pseudo-similarity measure.

This study follows the procedure of Lint et al. (2003) to impute missing data and predict travel times using data fusion approaches, whereas models integrated with grey models and neural networks are discussed in the present paper. This study proposes a process to treat traffic detector data and travel time data fusion. In contrast to previous studies, this study attempts to integrate grey theory and modeling into data imputation and neural network-based data fusion models. Grey theory is an effective mathematical method, which is a multidisciplinary and generic theory dealing with systems characterized by poor information and/or for which information is lacking (Deng, 1988, 1989). Fields covered by grey theory include system analysis, data processing, modeling, prediction, decision making and control (Deng, 1989). In the missing data treatment stage, this study develops a grey-theory-based approach to impute missing values from temporal and spatial errors. This study uses a grey time series model to impute missing values occurring at some time-spots in time-series detector data. In addition, a grey-relational-based nearest neighbor method is proposed to recover spatial missing data. The model computes the relative geometric relationship between two vectors with partially missing elements, and finds the pseudo-nearest-neighbor that has maximum value of grey relational grade among all auxiliary vectors, thereby the missing values being substituted by corresponding values of the pseudo-nearest-neighbors. Furthermore, both off-line and real-time data fusion models are discussed for travel time estimation. Both spatial and temporal patterns of traffic data are considered in the data fusion models. This study uses the speed-based link travel time extrapolation approach (Lint and Zijpp, 2003) to off-line estimate travel times and uses recurrent neural networks (Elman, 1990) with grey-models for real-time travel time prediction. The grey accumulated generating operation (AGO) is an important feature of grey models, which focuses largely on reducing the randomness of data. The grey-based recurrent neural networks integrate grey modeling into recurrent neural networks that are capable of dealing with both randomness and spatial-temporal properties in traffic data implicitly. Moreover, field data from the national

freeway no.1 in Taiwan is used as an example for testing the proposed model. The paper is organized as follows. Section 2 presents the grey-theory-based missing data treatment models, and Section 3 presents the off-line and real-time travel time data fusion models. A case study is discussed in Section 4, and Section 5 contains conclusion remarks.

2. MISSING DATA TREATMENT

Missing values may occur at some time-spots of the traffic data time-series and/or occur in a whole series of detector data from some failed detectors. This study uses a grey time series model to impute temporal missing values occurring at some time-spots in time-series detector data. In addition, a grey-relational-based nearest neighbor method is proposed to recover spatial missing data.

Lint et al. (2003) uses an exponentially moving average (MA) model to recover those temporal missing or corrupt values. However, when high percentage data corruption occurred in the time series, it is difficult to recover missing values by forecast values of time series model, e.g., ARIMA. In response to these situations, we attempt to apply the grey time series models, GM(1,1). The grey time series models encompass a group of differential equations adapted for parameter variance. A GM series is defined as a time series in which the number of data points of the series only need to be more than or equal to four (Deng, 1989). Furthermore, assumptions regarding statistical distribution of data are unnecessary when applying the grey theory.

Let $\mathbf{x}_d^{(0)} = [x_d^{(0)}(1), x_d^{(0)}(2),..., x_d^{(0)}(n)]$ be an original time-series traffic data record from detector d with n time-spots. The AGO formation of $\mathbf{x}_d^{(0)}$ is $\mathbf{x}_d^{(1)} = [x_d^{(1)}(1), x_d^{(1)}(2),..., x_d^{(1)}(n)]$, where $x_d^{(1)}(k) = \sum_{i=1}^{k} x_d^{(0)}(i)$, k= 1, 2, …, n. It is easy to recover $\mathbf{x}_d^{(0)}$ from $\mathbf{x}_d^{(1)}$ as $x_d^{(0)}(k) = x_d^{(1)}(k) - x_d^{(1)}(k-1)$, the inverse operation of grey AGO is called IAGO. The GM(1,1) model can be formulated as $\frac{\mathrm{d}x_d^{(1)}}{\mathrm{d}k} + ax_d^{(1)} = u$. The forecasting functions of $x_d^{(0)}(k)$ can then be obtained as follows:

$$\hat{x}_d^{(0)}(k) = \left(x_d^{(0)}(1) - \frac{\hat{u}}{\hat{a}} \right)(1 - e^{\hat{a}})e^{-\hat{a}(k-1)} \qquad k = 2, 3, ... \tag{1}$$

where the parameters, a and u, can be determined by applying the least-squares method. GM(1,1) differs from conventional statistical models and time series models (e.g., ARIMA) in not demanding a large amount of data with a good statistical distribution. In other words, GM(1,1) is useful for modeling when there is only a small amount of data available, with poor statistical distribution.

Herein, missing values $\mathrm{NA}_d(t+1)$ from detector d at time point t+1, then can be replaced by a forecast $\hat{x}_d^{(0)}(t+1)$, using (1), wherein inputting a time series with presence values from time i to t, i.e., $x_d^{(0)}(i)$, $x_d^{(0)}(i+1)$, ..., $x_d^{(0)}(t)$, such as $\hat{x}_d^{(0)}(t+1) = \left(x_d^{(0)}(i) - \frac{\hat{u}}{\hat{a}} \right) \cdot (1 - e^{\hat{a}})e^{-\hat{a}(k)}$ k =i, i+1, ..., t.

For dealing with the spatial missing values, this study follows the concept of pseudo-nearest-neighbor approach (Huang and Zhu, 2002) and integrates with grey relational analysis to substitute missing traffic data. In model of Huang and Zhu (2002), they developed a weighted correlation measure between two vectors with partially missing element values, named as pseudo-similarity; and then found the pseudo-nearest-neighbor that has the largest pseudo-similarity value. However, by assumptions of Huang and Zhu (2002), each data attribute is governed by univariate Gaussian randomly distribution in case that the data set is Gaussian randomly distributed. Most studies on missing data treatment also focus on the methods for handling the randomly distributed data (Little and Rubin, 1987). In the approach of Huang and Zhu (2002), data sets are first required to ensure their distribution types to be Gaussian distributions. However, we do not sure whether traffic data sets follow Gaussian distributions. Grey relational analysis (Deng, 1988) provides an alternative approach to identify the correlations between two vectors with partially missing elements. The degree of correlation between two vectors, the grey relational grade, can be represented by the relative geometric relationship between them in an imaging grey space without making *a prior* assumption about data distribution types. Similar grey relational-based nearest neighbor approach has been developed in Huang and Lee (2004) for missing attribute value prediction.

Let $\mathbf{x} = [x(1), x(2),..., x(n)]$ be a traffic data set. We use $\mathbf{x}_k = [x(1), x(2),..., x(k), \mathrm{NA}(k+1), ..., \mathrm{NA}(n)]$, $k<n$, to represent a data vector with (n-k) missing values, and let $\mathbf{x}_k$ be a data vector with missing values should be recovered. Consider m compared incomplete data vectors, $\mathbf{x}_l$ ($l = 1,...,m$), of data set $\{\mathbf{x}\}$. Then we can have the $\mathbf{x}_k$ and $\mathbf{x}_l$ be expressed as $\mathbf{x}_k = [x_k(1), x_k(2),..., x_k(d), S(d+1),..., S(k), \mathrm{NA}(k+1), \ldots, \mathrm{NA}(n)]$ and $\mathbf{x}_l = [x_l(1), x_l(2), ..., x_l(d), S(d+1),..., S(k), \mathrm{NA}(k+1), \mathrm{NA}(n)]$, where $d \le \min(k, l)$, and $S(i)$ represents the value that is missing in one of the vectors $\mathbf{x}_k$ and $\mathbf{x}_l$ but not in both. The grey relational coefficient, $\xi_{kl}(i)$, between data vector, $\mathbf{x}_k$ and $\mathbf{x}_l$ at a certain time point i=t, can be expressed by the following equation, which represents the relative geometric relationship between two vectors at time point i=t,

$$\xi_{kl}(t) = \frac{\min\limits_{l}\min\limits_{t}\left|x_k(t) - x_l(t)\right| + \rho \max\limits_{l}\max\limits_{t}\left|x_k(t) - x_l(t)\right|}{\left|x_k(t) - x_l(t)\right| + \rho \max\limits_{l}\max\limits_{t}\left|x_k(t) - x_l(t)\right|} \tag{2}$$

where $\left|x_k(t) - x_l(t)\right|$ represents the absolute difference between the two vectors at point i=t, $\min\limits_{l}\min\limits_{t}\left|x_k(t) - x_l(t)\right|$, $\max\limits_{l}\max\limits_{t}\left|x_k(t) - x_l(t)\right|$ are, respectively, the minimum and maximum distance for all points in all compared incomplete data vectors, $\mathbf{x}_l$ ($l = 1,...,m$). ρ ($0 \le \rho \le 1$) is a distinguishing coefficient used to adjust the range of the comparison environment, and to control level of difference of the relational coefficients. In cases where data variation is large, ρ usually ranges from 0.1 to 0.5 for reducing the influence of extremely values. Note that the complete data vector can be taken as the basis of the presented values of data vectors. For all $i \le d$, the grey relational grade, γ_{kl}, can be expressed as

$$\gamma_{kl} = \frac{1}{d}\sum_{i=1}^{d}\xi_{kl}(i) = \frac{1}{d}\sum_{i=1}^{d}\frac{\min_{l}\min_{i}|x_k(i)-x_l(i)| + \rho\max_{l}\max_{i}|x_k(i)-x_l(i)|}{|x_k(i)-x_l(i)| + \rho\max_{l}\max_{i}|x_k(i)-x_l(i)|} \tag{3}$$

The aim of grey relational grade is to recognize the geometric relationship between two data vectors in relational space. In this study, we use the grey relational grade to measure the pseudo-similarity between vectors $\mathbf{x}_k$ and $\mathbf{x}_l$. Then, a larger grey relational grade γ_{kl} is given to the vectors having less missing elements. The grey-relational-based pseudo-nearest neighbor method finds a data vector $\mathbf{x}_l$ such that: $\mathbf{x}_l$ has the presence of value $x_l(k+1)$; $\mathbf{x}_l$ has the maximum grey relational grade, based on the present data values, the present value $x_l(k+1)$ of $\mathbf{x}_l$ is used to replace the $\mathrm{NA}(k+1)$ in vector $\mathbf{x}_k$. The algorithm of the grey-relational-based pseudo- nearest-neighbors substitution method for data imputation is presented as follows:

Pre-condition: data vector $\mathbf{x}_k$ with missing values, i.e.,
$\mathbf{x}_k = [x(1), x(2),..., x(k), \mathrm{NA}(k+1),..., \mathrm{NA}(n)]$.

Post-condition: the missing values in $\mathbf{x}_k$ being substituted by corresponding values of the grey-relational-based pseudo- nearest-neighbors, i.e., $\mathbf{x}_k = [x(1), x(2),..., x(n)]$.

Computation:

For each data vector $\mathbf{x}_k$

{ For each missing values $x_k(i)$ of $\mathbf{x}_k$

{ For each $\mathbf{x}_l \in \{\mathbf{x}\} - \mathbf{x}_k$

{ If the $x_k(i)$ value is non-missing

Compute γ_{kl} }

Find the $\mathbf{x}_l^*$ that has the maximum value of γ_{kl} among all $\mathbf{x}_l$ examined}

Replace the element $x_k(i)$ of $\mathbf{x}_k$ by the $x_l(i)$ value of $\mathbf{x}_l^*$ }

3. DATA FUSION FOR TRAVEL TIME ESTIMATION

Travel time can be simply defined as the time spent in traveling from one point to another. Specifically, travel time estimation requires more comprehensive considerations about temporal and spatial dimensions. In this study, both off-line and real-time models are discussed. This study follows the speed-based link travel time extrapolation approach (Lint and Zijpp, 2003) to off-line estimate travel times. Furthermore, this study uses recurrent neural networks (RNN) (Elman, 1990) with grey-models for real-time travel time prediction. Then, the grey-based RNN model is trained with the off-line estimated travel times using the speed-based link travel time extrapolation approach.

3.1 Off-line Travel Time Estimation Using Speed-based Extrapolation

Herein, the modeling of off-line link travel time estimation follows the formulation of Lint and Zijpp (2003), it is summarized as follows. Speed-based travel time extrapolation aims to reconstruct vehicles trajectories by generalizing loop-detector speeds over space, and hence

calculate travel time. Data from detectors can comprise a space-time grid of region {*k*, *d*}, *k*=1, 2, …, *m*, *d*=1,2,…,*n*, where *k* is detector locations and *d* is time duration for study. Consider region {*k*, *d*} as a rectangular area in space-time from $\{s_k, t_d\}$ to $\{s_{k+1}, t_{d+1}\}$. Consider the time-dependent speed $v_i(t)$ of vehicle *i* traversing a section between detectors *k* and *k*+1, $s_i(t)$ is the trajectory function of vehicle *i* driving at section *k* during duration *d*:

$$s_i(t) = s_i^0 + \left(\frac{V(k,d)(s_{k+1} - s_k)}{V(k+1,d) - V(k,d)} + s_i^0 - s_k \right) \left(e^{[\frac{V(k+1,d)-V(k,d)}{s_{k+1}-s_k}](t-t_i^0)} - 1 \right) \tag{4}$$

where $\{s_i^0, t_i^0\}$ is the initial point of vehicle *i*'s trajectory entering cell {*k*, *d*}. However, if $V(k,d) = V(k+1,d)$, then $s_i(t) = s_i^0 + V(k,d)(t - t_i^0)$. From Eq. (4), the exit time of vehicle *i* can be calculated as

$$t_i^* = t_i^0 + \frac{(s_{k+1} - s_k)}{V(k+1,d) - V(k,d)} \ln \left(\frac{\frac{V(k,d)(s_{k+1} - s_k)}{V(k+1,d) - V(k,d)} + s_{k+1} - s_k}{\frac{V(k,d)(s_{k+1} - s_k)}{V(k+1,d) - V(k,d)} + s_i^0 - s_k} \right) \tag{5}$$

where $s_i^0 + \left(\frac{V(k,d)(s_{k+1} - s_k)}{V(k+1,d) - V(k,d)} + s_i^0 - s_k \right) \left(e^{[\frac{V(k+1,d)-V(k,d)}{s_{k+1}-s_k}](t-t_i^0)} - 1 \right) > s_{k+1}$, and $s_i^* = s_i(t_i^*) = s_{k+1}$. If the vehicle velocity is too low to let $s_i(t_{d+1}) < s_{k+1}$, we can use Eq. (4) to calculate the vehicle trajectory $s_i(t_{d+1})$. Furthermore, consider again the situation that $V(k,d) = V(k+1,d)$, then the exit time t_i^* can easily be reduced as $t_i^* = t_i^0 + (s_{k+1} - s_i^0)/V(k,d)$. The trajectory-based travel time estimation algorithm can be presented as follows:

Step 1: Start vehicle *i*, set section *k* (first section *k*=1).
Step 2: Vehicle *i* enters region {*k*, *d*} at $\{s_i^0, t_i^0\}$. Calculate exit time $\{t_i^*\}$ and location $\{s_i^*\}$ of vehicle *i* leaving cell {*k*, *d*}, using Eqs. (5) and (4), respectively. If $s_i^* = s_{k+1}$, then *k*=*k*+1; otherwise, $t_i^* = t_{d+1}$ and *d*=*d*+1.
Step 3: If $k \geq m$, then stop, the end of trajectory of vehicle *i* is reached; record its departure time and link travel time. Otherwise, return to Step 2.
Step 4: Let *i*=*i*+1, return to Step 1, and vehicle *i*+1 enters region {*k*, *d*} at $\{s_{i+1}^0, t_{i+1}^0 = t_i^0 + h\}$, where *h* is the headway between consecutive vehicles. This algorithm continues until $t_i^0 + h \geq n$.

From this algorithm, we add location of vehicles in time-space form to the individual trajectories when and where they leave their current region {*s*, *d*}. Then, this exit-time-space location determines in turn the entry-time-space-location of vehicles in the next region, and we can deduce the link-level vehicle trajectories and hence link travel time. Furthermore, from the results of Lint and Zijpp (2003), the speed-based trajectory method had very high accuracy for estimating travel time, it shown about 2% mean relative errors.

3.2 Real-time Travel Time Estimation Using Grey-based Recurrent Neural Networks

Many studies used neural network models to predict travel times, e.g., Dharia and Adeli (2003) used counter propagation neural network for travel time forecasting, and Lint et al. (2002, 2003) developed the state-space neural networks for travel time prediction. However, in travel time data fusion, the "dynamic" inherence of travel times refer both spatial and temporal patterns of traffic data should be considered. The recurrent neural network (RNN) model (Elman, 1990) is a dynamic network, in which it has an internal feedback. Elman (1990) emphasizes the capability of the RNN model to learn complex spatial-temporal patterns. RNNs are two-layer backpropagation networks, with the addition of a feedback connection from the output of the hidden layer to its input. This feedback path allows Elman RNNs to learn to recognize and generate temporal patterns, as well as spatial patterns. The RNN learns to interpret current inputs in the context of its previous internal states. However, uncertainty inherent to the distribution of traffic data complicates real-time travel time estimation. Stochastic (random) fluctuations in traffic data and travel times and unexpected abnormal fluctuations (i.e., if congestion and incidence occurring) significantly influence the future traffic information. Ma et al. (1994) noticed that the uncertainty and randomness in training data degrades the forecasting performance of neural network. It is expected that better forecasting performance can be achieved if the randomness in the training data can be reduced. This study proposes a novel grey-based recurrent neural network, which integrated grey models into the recurrent neural network, for dynamic travel time estimation.

Figure 1 depicts the structure of the grey-based RNN model, which consists of two trainable layers of neurons (hidden and output layers), and two layers which simply distribute their activations to the hidden layer (input and context layer), with feedback from the first-layer output to the first layer input. In Fig. 1, $f_{d'}$ and $s_{d'}$ represent, respectively, the volume and speed data from the upstream station; while $f_{d''}$ and $s_{d''}$ represent those from downstream station. $f_{d'}^{(1)}(t)$ and $s_{d'}^{(1)}(t)$ are, respectively, the AGO formations of $f_{d'}$ and $s_{d'}$ at time interval *t*; while $f_{d''}^{(1)}(t)$ and $s_{d''}^{(1)}(t)$ are, respectively, the AGOs of $f_{d''}$ and $s_{d''}$. Grey-AGO formations aim to include the measurements of volume and speed parameters from previous time intervals 0 to *t* at the same station. Then the current grey-AGOs (at time *t*) for both volume and speed parameters from upstream and downstream stations are used as inputs to the grey-based RNN model. It is reported that grey AGO is able to reduce the randomness in data (Hsieh, 2003). The output of grey-based RNN model represents the link travel time within the same section at some future time interval (*t*+*n*). The hidden neuron function for travel time estimation can be described as:

$$Z_j(t+n) = f'_H\left(\begin{array}{l}[\gamma_{1j} f_{d'}^{(1)}(t) + \gamma_{2j} s_{d'}^{(1)}(t) + \gamma_{3j} f_{d''}^{(1)}(t) + \gamma_{4j} s_{d''}^{(1)}(t)] + \\ \sum_j \delta_j z_j(t+n-1)\end{array}\right) \tag{6}$$

where $f'_H(\cdot)$ is the hidden neuron function for travel time estimation, γ_{1j}, γ_{2j}, γ_{3j}, γ_{4j} are weights connecting the input layer units (i.e., volume and speed parameters from upstream and downstream stations) to hidden unit *j* at time *t*, δ_j is the weight connecting context layer units to hidden unit *j*, $z_j(t+n)$ is a output of hidden node *j* at time *t*+*n*, and context layer input at time *t*+*n*-1 then is given as $z_j(t+n-1)$.

The output neuron of grey-based RNN can then be presented as

$$\hat{T}^{(1)}_{d'd''}(t+n) = f'_O(\beta_1 + \sum_j \beta_j z_j(t+n)) \tag{7}$$

where $\hat{T}^{(1)}_{d'd''}(t+n)$ is the output of grey-based RNN at time interval $t+n$, and $f'_O(\cdot)$ is the output neuron function for travel time estimation. Final, the travel time estimates at time interval $t+n$, $\hat{T}^{(0)}_{d'd''}(t+n)$, are obtained by using IAGO. The recurrent neural network model then undergoes a training process during which the weights associated with the interconnections are determined.

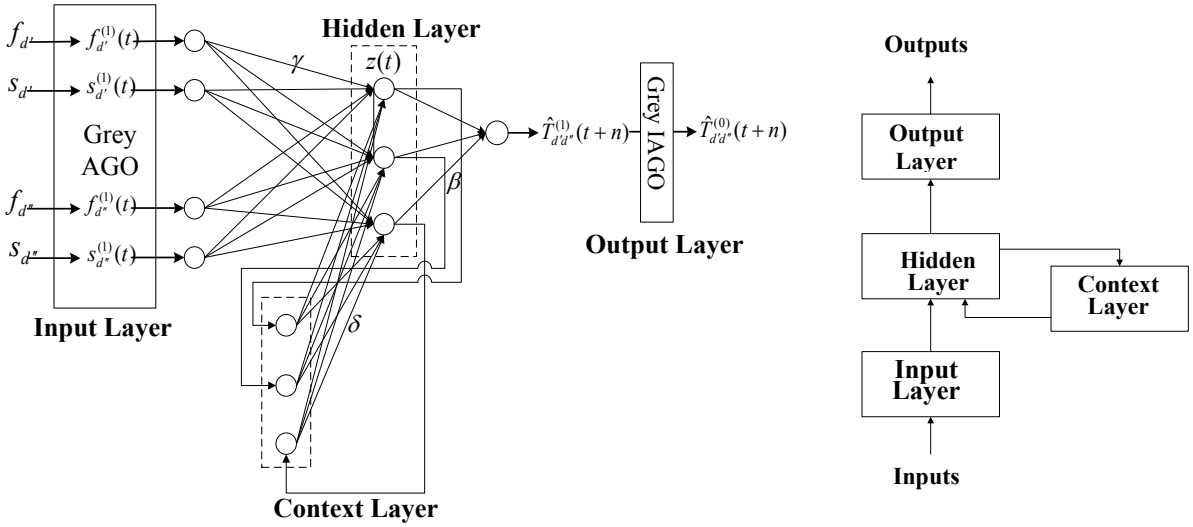

Figure 1. Architecture of Grey-based Recurrent Neural Network

For computation of δ, the derivative chain rule is formulated, then used to update the weight coefficients between context layer and the hidden layer of the training procedure of RNN. The learning algorithm in the RNN model is the same as that in the backpropagation networks, using the gradient descent rule, which adjusts the weights based on the derivatives of the error with respect to the weights. The entire input sequence is presented to the RNN network, and its outputs are calculated and compare with the target sequence to generate an error sequence. Mean squared error (MSE) is used as a performance function in training the RNN model. For each time step (epoch), the error is backpropagated to find gradients of errors for each weight and bias. The gradient is actually an approximation since the contributions of weights and biases to errors via the delayed recurrent connection are ignored. The gradient is then used to update the weights with the backprop training function.

Relevant neural network studies for travel time prediction used the results from simulation models (e.g., TSIS, CORSIM) as learning data (Nelson and Palacharla, 1993; Palacharla and Nelson, 1999; Dharia and Adeli, 2003). Some studies used probe report travel times from simulated vehicles. In contrast to those studies, the grey-based RNN model herein is trained with the offline estimated travel times (i.e., the trajectory-based travel time estimation algorithm, Eqs. (4)-(5)) for target sequences. The piecewise linear speed-based trajectory method has been proven that it produces unbiased estimates of the true travel times (Lint and Zijpp, 2003). The training process continues until the prediction errors (MSE) become very small and the network parameters converge to values that allow it to perform the desire mapping for each input-output correlation. Once the grey-based RNN is trained and its

performance is shown to be satisfactory, it can then be used on-line to provide link travel time estimation based on new real-time data.

4. CASE STUDY

This section presents a case study that demonstrates the results of traffic detector data treatment and travel time data fusion on freeway. The field data we analyzed were collected from a section of Taiwan freeway no. 1 with ten detectors located at about 500m intervals along a 6.03 km-long link. This study only used data from main-line dual-loop detectors between the on-ramp and off-ramp of the link. Our main goal was to study the morning peak-hour (8:00 AM to 9:00 AM) during one week of September in 2002. Substantial cleanup, identifying missing data and eliminating detector data that appeared to be giving nonsensical values of the data are necessary. After initial inspections of the data, it revealed many nonsensical values and missing elements of the data set. From raw data, one of the 12 detectors made no response and did not record any data. Another two detectors got demonstrably faulty values of vehicle counts, velocity, and occupancy that almost all values from this malfunctioning detector are revealed as default-value. Data from other three detectors temporally revealed many clearly flawed entries. Herein, those faulty values are removed from the data records. Ultimately, as a result of the data cleanup process, approximately 25% of the data were missing.

For demonstrating the performance of the grey-theory-based missing data treatment methods, we first use a group of complete data records from six healthy detectors for comparison. Then, we convert the original data records at three out of the six detectors into test datasets. The test datasets with increasing amount (10%-40%) of temporal missing values, in which missing values is randomly drawn from a uniform distribution in the time series. Detector data with partial or whole spatial missing values (10%-40%) are also randomly broken. Then, 50%-80% mixed temporal and spatial missing values are also tested. The accuracy of the proposed models was verified by comparing their results with the original (clean) data values. Figure 2 shows the mean relative errors for these three data records with respect to varying missing rate that ranges from 10% up to 80%. The horizontal axis denotes the percentage of data values absent in the data records, while the vertical axis denotes the mean relative error percentage compared with the original complete data sets. Furthermore, the overall model errors averaged about 20% with 50%-80% mixed missing data. From the experimental results of Huang and Zhu (2002), the errors were higher than 20% when missing data rate up to 40%. The results of our missing data recovery method were shown to be sufficiently accurate to treat with the missing values, by comparing the results of Huang and Zhu (2002).

Furthermore, the performances of off-line travel time estimation inputting data with and without performing missing data treatments were compared. For comparison, we input complete (clean) data records, and input test datasets containing 33% missing values with and without performing the proposed missing data treatment models, respectively, to the off-line travel time estimation model. The estimated travel times using data with missing data treatment highly approximated to the estimated travel times inputting complete data. However, the travel time estimation using incomplete data records without missing data treatments got high deviation. From this experimental result, it implied that the proposed missing data treatment models ensured the accuracy of off-line travel time estimation with incomplete data sets.

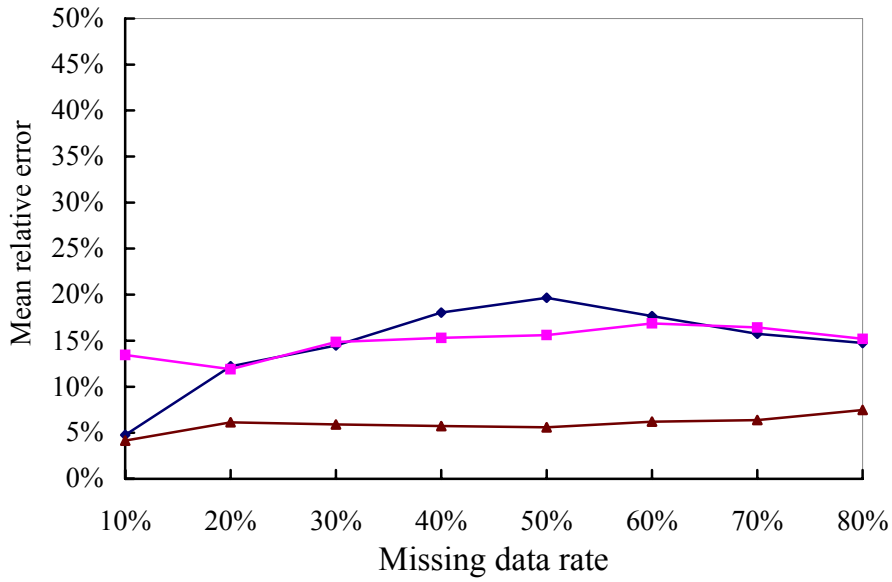

Figure 2. Mean Relative Errors versus the Percentage of Missing Values on Data Records

The grey-based RNN model was further applied to dynamic forecasting section travel times. The neural-network-toolbox of MATLAB was used to build the RNN networks, train them and run the RNN model to dynamic estimate the travel time. The function 'newelm' in neural-network-toolbox of MATLAB allows users to create an Elman recurrent network model. Herein, the grey-based RNN models are trained using the above off-line estimated travel times. Two RNN procedures were tested; the first involved incorporating missing data treatment methods while using the recovered data for RNN training and the second involved using partially missing data without missing data treatments in the RNN training procedures. The complete data records are used to off-line estimate the target travel time sequence. At each training time step, input vectors are presented to the grey-based RNN network, and generate MSEs; the errors are backpropagated to find gradients of errors for each weights and bias, then the gradients are used to update the weights with the learning functions. The grey-based RNN model with inputs with missing data treatment has higher training performance (smaller MSEs), while those two training results are very similar. After 1000 epochs for training, the mean squared error of grey-based RNN with inputs with missing data treatment is 0.00133, while the RNN without missing data treatment got 0.00154 MSE. Table 1 lists the performance results of the grey-based RNN for estimating travel times for various prediction horizons (i.e., 20 seconds, 1, 5, and 10 minutes). The results show that the grey-based RNN models are capable of estimating travel times up to 10 minutes into the future with a high degree of accuracy (about 96-97%).

Table 1. Forecasting Performance of Grey-based RNN for Travel Time Estimation

Prediction horizon	Prediction errors	
	MSE	Average relative error (%)
20 seconds	0.001563	3.180
1 minute	0.001527	3.206
5 minutes	0.001523	3.149
10 minutes	0.001587	3.131

Figure 3 shows the estimated section travel times using the grey-based RNN models trained by off-line estimated travel times inputting detector data with and without missing data treatment. The off-line estimated travel times are also shown in Figure 3. From Figure 3, both grey-based RNN results were close to the off-line travel time estimation, it appears that both RNN approaches are robust to data input failures, albeit that the RNN with data treatments may perform better. Since only about 25% data corruptions occurred in this case, the RNN travel time prediction results with and without missing data treatment were close. However, inputting missing data in the training procedure makes the internal states of grey-based RNN more difficult to interpret, and makes the model less useful for analytical analysis. Restated, this case study demonstrates how the grey-theory-based missing data treatment and grey-based RNN models may be applied to the rapid forecasting of link travel time; and the RNN models were capable of accurately predicting travel times inputting incomplete traffic datasets.

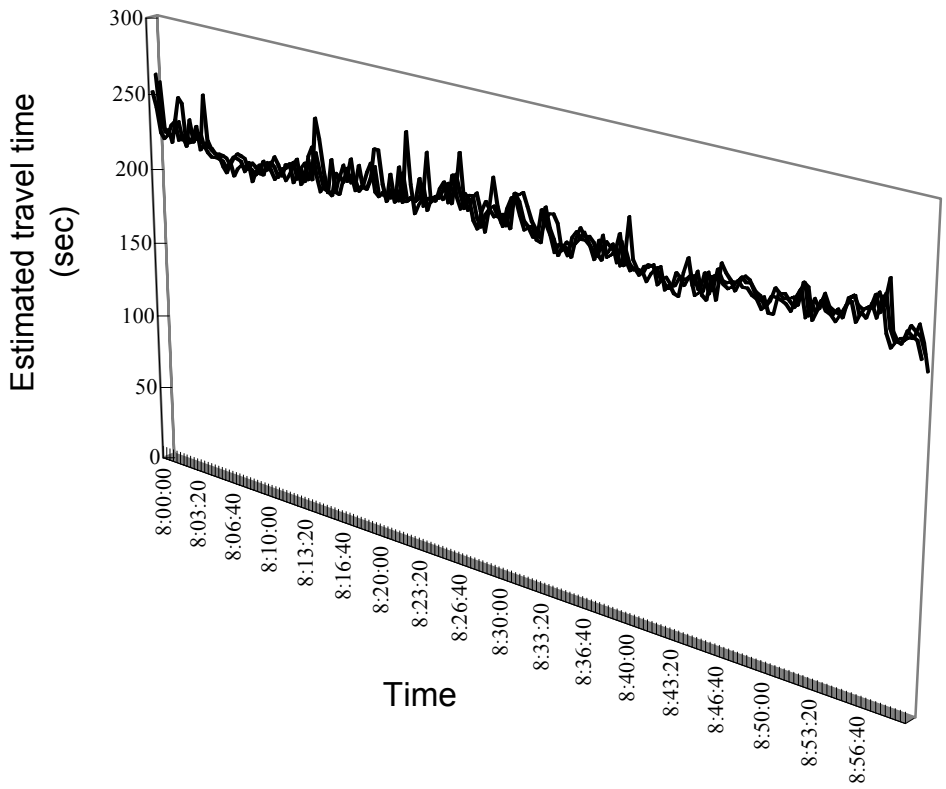

Figure 3. Estimated Link Travel Time using Grey-based RNN Models and Off-line Model

5. CONCLUSIONS

This study focuses on missing data treatment and data fusion for traffic detector data. In contrast to previous studies, this study attempts to integrate grey modeling into data imputation and neural network-based data fusion models. Grey-theory-based nearest neighbor method and grey time-series model are developed to recover temporal and spatial missing values on datasets. This study uses a grey time series model to impute missing values occurring at some time-spots in time-series detector data. Grey-relational-based nearest neighbor method is proposed to recover spatial missing data. We proposed a

modification to the pseudo-nearest-neighbor approach, in which we released the assumption of Gaussian data sets using the grey relational analysis. The grey relational grade is used to calculate the relative geometric relationship between two vectors, and the procedure is to find a pseudo-nearest-neighbor that has maximum value of grey relational grade among all auxiliary vectors, thereby the missing values being substituted by corresponding values of the pseudo-nearest-neighbors. Both spatial and temporal patterns of traffic data are considered in travel time data fusion. A speed-based link travel time extrapolation model for analytical travel time estimation and a grey-based recurrent neural network model for real-time travel time prediction are presented. The grey-based RNN model consists of an input layer comprising grey-AGOs of speed and volume parameters from the upstream and downstream stations. The feedback path allows the RNN models to learn to recognize and generate temporal patterns, as well as spatial patterns. The output of the grey-based RNN model is the travel time within the same section (link) at some future state.

Application of the proposed models to field data from the Taiwan national freeway no.1 was demonstrated in a case study. Study results showed that the data treatment models for faulty data recovery were sufficiently accurate to treat with temporal and spatial missing data. The results also indicated that the proposed missing data treatment ensured the accuracy of travel time estimation with incomplete datasets. Moreover, the grey-based RNN data fusion models were capable of accurately predicting travel times (about 96-97% accuracy). Grey-based RNN approaches were also shown to be robust to missing data inputs. Consequently, the proposed approaches can ensure the accuracy of travel time estimation with incomplete data sets, and are therefore suited to implementation for ATIS.

ACKNOWLEDGEMENTS

This study was supported in part by the Ministry of Education of Taiwan, ROC under Grants EX-91-E-FA06–4-4. Partial studies have been presented in 2005 IEEE International Conference on Networking, Sensing and Control (Wen et al., 2005).

REFERENCES

Chen, H., Grant-Muller, S., Mussone, L., Kontgomery, F. (2001) A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. **Neural Computing and Applications, Vol. 10**, 277-286.

Coifman, B. (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. **Transportation Research-A, Vol. 36,** 351-364.

Cremer, M. and Schutt, H. (1990) A comprehensive concept for simultaneous state observation, parameter estimation and incident detection. **12th International Symposium on Transportation and Traffic Theory**, 1990.

Dailey, D.J. (1993) Travel time estimation using cross-correlation techniques. ***Transportation Research-B*, Vol. 27**, 97-107.

Dempster, P., Laird, N.M., Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. **Journal Royal Statistical Soc Series B-Statistical Methodology, Vol. 39**, 1-38.

Deng, J. L. (1988) Properties of relational space for grey system. In J. L. Deng, (ed.), **Essential Topics on Grey System-Theory and Applications**, China Ocean, Beijing, China.

Deng, J. L. (1989) Introduction to grey system theory. **Journal of Grey System, Vol. 1, No. 1,** 1-24.

Dharia, A. and Adeli, H. (2003) Neural network model for rapid forecasting of freeway link travel time. **Engineering Applications of Artificial Intelligence, Vol. 16,** 607-613.

Elman, J. L. (1990) Finding structure in time. **Cognitive Science, Vol. 14**, 179-211.

Fitzgerald, W.J. (1999) The restoration of missing data using Bayesian numerical methods. **Proc ICASSP'99, Vol. I-VI,** 1999.

Gelfand, E., Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. **Journal of American Statistical Association, Vol. 85,** 398-409.

Huang, C.C. and Lee, H.M. (2004) A grey-based nearest neighbor approach for missing attribute value prediction. **Applied Intelligent, Vol. 20,** 239-252.

Huang, X. and Zhu, Q. (2002) A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. **Pattern Recognition Letters, Vol. 23,** 1613-1622.

Hsieh, C.H. (2002) Grey neural network and its application to short-term load forecasting problem. **IEICE Transaction on Information and Systems**, **Vol. E58-d**, 897-902.

Lawless, J. F., Kalbfleisch, J.D. (1999) Semi-parametric methods for response-selective and missing data. **Journal Royal Statistical Soc Series B-Statistical Methodology, Vol. 61, No. 2,** 413-438.

Lint, J.W.C., Hoogendorn, S.P., Zuylen, H.J. (2002) Freeway travel time prediction with state-space neural networks. **Transportation Research Records**, **No. 1811**, 30-39.

Lint, J.W.C., Hoogendorn, S.P., Zuylen, H.J. (2003) Toward a robust framework for freeway travel time prediction: experiments with simple imputation and state-space neural networks. **Proc. 82nd Transportation Research Board Annual Meeting,** Washington, D.C., 2003.

Lint, J.W.C. and Zijpp, N. J. (2003) An improved travel-time estimation algorithm using dual-loop detectors. **Proc. 82nd Transportation Research Board Annual Meeting,** Washington, D.C., 2003.

Little, R. and Rubin, R. (1987) **Statistical Analysis with Missing Data,** Wiley, New York.

Ma, H., El-Keib, A.A. and Ma, X. (1994) Training data sensitivity for artificial neural network-based power load forecasting. **Proc. 26th Southern Symposium on System Theory,** 650-652.

Nam, D.H. and Drew, D.R. (1996) Traffic dynamics: method for estimating freeway travel times in real time from flow measurements. **Journal of Transportation Engineering, Vol. 122,** 185-191.

Nelson, P. and Palacharla, P. (1993) A neural network model for data fusion in ADVANCE. **1993 Transtech Pacific Rim Conference,** Seattle, July 25, 1993.

Oh, J. S., Jayakrishnan, R., Recker, W. (2002) Section travel time estimation from point detector data. Rep. UCI-ITS-WP-02-11, Institute of Transportation Studies, U.C. Irvin.

Palacharla, P. and Nelson, P. (1999) Application of fuzzy logic and neural networks for dynamic travel time estimation. **International Transaction in Operational Research, Vol. 6,** 145-160.

Pawlak, M. (1993) Kernel classification rules from missing data. **IEEE Transactions on Information Theory, Vol. 39, No. 3,** 979-988.

Redfern, E. J., Watson, S.M., Tight, M.R., Clark, S.D. (1993) A comparative assessment of current and new techniques for detecting outliers and estimating missing values in transport related time series data. **Proc. PTRC European Transport, Highways and Planning 21st Summer Annual Meeting, Vol. 363**, 1993.

Sande, I. (1996) Hot deck imputation procedures. **Symposium on Incomplete Data Proceedings, Vol. III,** 1996.

Saporta, G. (2002) Data fusion and data grafting. **Computational Statistics and Data Analysis, Vol. 38,** 465-473.

Schafer, J. L. (1997) **Analysis of Incomplete Multivariate Data.** Chapman & Hall, London.

Wen, Y.H., Lee, T.T., Cho, H.J. (2005) Hybrid models toward traffic detector data treatment and data fusion. **2005 IEEE Intl. Conf. on Networking, Sensing and Control,** Tucson, March 19-22, 2005.