Development of Archiving and Data Fusion Strategies for Travel Time Data

James K. Richardson
Graduate Research Assistant

Brian L. Smith, Ph.D., P.E.
Associate Professor

Center for Transportation Studies
University of Virginia

Final Report
Mid-Atlantic Universities Transportation Center

December, 2008
A Research Project Report
For the Mid-Atlantic Universities Transportation Center (MAUTC)
A U.S. DOT University Transportation Center

Dr. Brian L. Smith
Department of Civil Engineering
Email: bls2z@virginia.edu

Center for Transportation Studies at the University of Virginia produces outstanding transportation professionals, innovative research results and provides important public service. The Center for Transportation Studies is committed to academic excellence, multi-disciplinary research and to developing state-of-the-art facilities. Through a partnership with the Virginia Department of Transportation’s (VDOT) Research Council (VTRC), CTS faculty hold joint appointments, VTRC research scientists teach specialized courses, and graduate student work is supported through a Graduate Research Assistantship Program. CTS receives substantial financial support from two federal University Transportation Center Grants: the Mid-Atlantic Universities Transportation Center (MAUTC), and through the National ITS Implementation Research Center (ITS Center). Other related research activities of the faculty include funding through FHWA, NSF, US Department of Transportation, VDOT, other governmental agencies and private companies.

Disclaimer: The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.
There is currently a robust field of research dedicated to estimating travel time on road networks. Nearly all of this work has focused on the task of deriving travel time from loop-detector data and various probe monitoring approaches. However, there has been little effort devoted to effectively and efficiently manage travel time data originating from multiple sources. Archived Database Management Systems (ADMS) emerged as a response to the need to manage and store the massive amounts of data being generated by “traditional” ITS technologies (in most cases, point detector data). Yet, while ADMS systems have been successful in storing vast amounts of point data such as speed, occupancy, and volume, there is currently no consensus on the best approach to storing and managing travel time data. This research explored the issues involved in managing travel time data and looks at information technology and statistical methods to address the key challenges in this area. The results presented in this report include a new data model to accommodate travel time data according to three distinct dimensions of the data: the source, the spatial extent, and the time of measurement. In addition, the report presents a proposed data fusion method based on weighting values derived from space, time, and source specific functions.
ABSTRACT

There is currently a robust field of research dedicated to estimating travel time on road networks. Nearly all of this work has focused on the task of deriving travel time from loop-detector data and various probe monitoring approaches. However, there has been little effort devoted to effectively and efficiently manage travel time data originating from multiple sources. Archived Database Management Systems (ADMS) emerged as a response to the need to manage and store the massive amounts of data being generated by “traditional” ITS technologies (in most cases, point detector data). Yet, while ADMS systems have been successful in storing vast amounts of point data such as speed, occupancy, and volume, there is currently no consensus on the best approach to storing and managing travel time data. This research explored the issues involved in managing travel time data and looks at information technology and statistical methods to address the key challenges in this area. The results presented in this report include a new data model to accommodate travel time data according to three distinct dimensions of the data: the source, the spatial extent, and the time of measurement. In addition, the report presents a proposed data fusion method based on weighting values derived from space, time, and source specific functions.
INTRODUCTION

There is currently a robust field of research and development dedicated to measuring travel time on road networks. Much of this work has focused on the task of deriving travel time estimates from loop-detector data. Lately, there has also been a great deal of interest in using other technologies such as cellular telephones, GPS devices, and roadside cameras to estimate mean link travel times. A related challenge that has not received nearly as much attention, is what to do with the data after it is generated? Transportation agencies and information service providers will find themselves in the position of having multiple sources of travel time data. In most cases, the multiple sources use different roadway link structures, provide data at different measurement time intervals, and provide different levels of quality. Thus, the very real challenge being faced is how to concurrently manage this data and use it effectively.

In recent years, Archived Database Management Systems (ADMS) have emerged as a response to the need to manage and store the massive amounts of data being generated by “traditional” intelligent transportation system (ITS) technologies. Yet, while ADMS systems have been successful in storing vast amounts of event data (i.e. incidents), and point detector data (i.e. speed, occupancy, and volume), there is currently no consensus on the best approach to storing and managing travel time data. This report presents the results of research to address the challenge of managing and archiving travel time data, assessing a number of different technologies and methodologies that hold high potential.

STATE OF THE PRACTICE

ADMS systems have been developed primarily to support archiving of point detector (i.e. inductive loop detectors, video detection systems, etc.) data and event data. Commonly used data structures reflect this fact. For example, the ADMS Virginia system is modeled along the classic “star schema” found in many data warehousing systems. A star schema consists of a set of “fact tables” and “dimension tables”. In the case of the ADMS Virginia system, the fact tables hold information such as traffic flows whereas the dimension tables hold information such as detector locations. The data model is illustrated below in Figure 1.
The current ADMS data model has performed well in storing data such as speed, occupancy, and volume generated by loop detectors. However, there is currently no accommodation in the database for storing travel times. One of the goals of this research is to develop a new data model for archiving travel time data. To set a foundation for this work, it is useful to examine some of the challenges inherent in archiving travel time.

TRAVEL TIME DISTRIBUTION

Travel time is measured over arbitrary space and time dimensions. Estimates of travel time are inferences of the space-mean speed of a particular road segment at a particular time. For example, the travel times as measured by two probe vehicles may be quite different for the same spatial extent at the same time. Thus, the travel time of vehicles over a particular link during a particular time interval may be considered as a random variable with an unknown underlying distribution, expected value, and variance. The Central Limit Theorem states that the distribution of the sample mean of a random variable tends towards the normal distribution as the number of samples grows large, and that the mean and variance of this distribution will be the same as the underlying distribution of the population. Thus, given a large enough sample of travel time estimates over a common space / time extent, we can compute an acceptably accurate space-mean speed for the extent at a particular time.
TRAVEL TIME AS MULTI-DIMENSIONAL DATA

In order to archive travel time data there are three critical dimensions of the data that must be considered:
- Source of Travel Time Estimate
- Spatial Extent
- Time

Source Dimension

As discussed in the introduction, travel time data will likely be provided by a number of different sources external to the archive. Although we presently can estimate travel times using loop detector data stored in a traditional archive, we are also looking to bring in different sources of travel time data.

A number of private firms currently provide travel time estimates using probe based estimates such as wireless location technology (WLT) and automatic vehicle identification (AVI). (see 1) From the perspective of this research, the method by which the travel time is measured is less important. However, it is important that we know the source of the estimate since we will be making inferences about travel time based on this information.

This raises a number of important questions about our “confidence” in the data provided by a particular travel time source. For example, if we have two travel time estimates for a common space / time extent from two different sources, how can we determine which estimate is closer to the true population mean? Different methods can produce different estimates of travel times. By assigning a value to our confidence in the estimate we can weight the estimates.

Spatial Dimension

Inherent in any travel time data is a spatial dimension. Travel time data is a measure of the time to move from a starting position along a link to an ending position. Let’s call this a travel time segment. One of the major challenges in archiving travel time data is finding a common spatial segmentation strategy.

In order to uniquely identify a spatial extent we need the following data:
- Starting position of the extent
- Ending position of the extent
- Route identifier

We can use this data to uniquely identify any travel time spatial extent. However, a problem occurs when we start dealing with data from many different sources. Each source can adopt its own segmentation strategy. This makes it increasingly difficult to integrate new data sources into the archive. In order to avoid this problem, the traffic information and mapping industry has developed standardized TMC codes to identify spatial extents.

TMC (Traffic Management Channel) is a specific application of the FM Radio Data System (RDS) used for broadcasting real-time traffic.(7) TMC codes were originally developed in Europe to identify important points on the road network. The system is being deployed in the United States by a consortium of vendors. Using the US-based TMC codes
spatial extents can be identified as the path between TMC-points. Several different travel time data vendors are, or will be, adopting TMC codes as the basis of their segmentation strategy.

**Time Dimension**

The final dimension of travel time data is the time dimension. For the purposes of archiving travel time we will consider the time dimension to refer to the extent-in-time (i.e. the time interval) during which vehicles entering a given spatial extent will experience the estimated space-mean speed (i.e. travel time estimate).

There are a number of interesting questions that must be considered. How large of a time extent can be used in estimating travel time? What is the relationship between the size of the time-extent and the accuracy of the travel time estimate? How can we compare estimates across time?

**A PROPOSED DATA MODEL FOR ARCHIVING TRAVEL TIME**

Given these three dimensions of travel time, a data model for archiving travel time is proposed. This model, illustrated in Figure 2, shows the relationships between the main entities.

![Diagram of Proposed Data Model for Archiving Travel Time Data](image-url)
### Table 1 – Entity Descriptions

<table>
<thead>
<tr>
<th>Entity</th>
<th>Primary Key</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT_SOURCES</td>
<td>TT_SOURCE_ID</td>
<td>Stores information about each of the sources of travel time data (i.e. travel time providers)</td>
</tr>
<tr>
<td>TT_SEGMENTS</td>
<td>TT_SEGMENT_ID</td>
<td>Each segment where travel time is being measured will have TT_SOURCE_ID as a foreign key so we know the source of the measurements for that segment. Additional attributes may include the start and end points of the segment either as coordinates or mileposts.</td>
</tr>
<tr>
<td>TRAVEL_TIME</td>
<td>TT_ID</td>
<td>A segment can have many travel time measurements over time. Each measurement would be stored here.</td>
</tr>
</tbody>
</table>

The following tables describe in more detail the proposed attributes in each of the tables in the travel time archives data model.

#### Table 2 - TT_SOURCES Attribute Descriptions

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Data Type</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT_SOURCE_ID</td>
<td>NUMBER</td>
<td>Primary Key</td>
</tr>
<tr>
<td>SOURCE_NAME</td>
<td>VARCHAR</td>
<td>Description travel time estimation source</td>
</tr>
<tr>
<td>SOURCE_CONFIDENCE</td>
<td>NUMBER</td>
<td>A measurement of confidence in the source’s estimates</td>
</tr>
</tbody>
</table>

#### Table 3 - TT_SEGMENTS Attribute Descriptions

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Data Type</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT_SEGMENT_ID</td>
<td>NUMBER</td>
<td>Primary Key</td>
</tr>
<tr>
<td>TT_SOURCE_ID</td>
<td>NUMBER</td>
<td>Foreign Key</td>
</tr>
<tr>
<td>SEGMENT_START</td>
<td>NUMBER</td>
<td>Starting point of segment</td>
</tr>
<tr>
<td>SEGMENT_END</td>
<td>NUMBER</td>
<td>Ending point of segment</td>
</tr>
<tr>
<td>SEGMENT_CONFIDENCE</td>
<td>NUMBER</td>
<td>General confidence level in this segment’s travel time estimates</td>
</tr>
</tbody>
</table>
Table 4 – TRAVEL.TIME Attribute Descriptions

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Data Type</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT_ID</td>
<td>NUMBER</td>
<td>Primary Key</td>
</tr>
<tr>
<td>TT_SEGMENT_ID</td>
<td>NUMBER</td>
<td>Foreign Key</td>
</tr>
<tr>
<td>DATE_TIME</td>
<td>DATE</td>
<td>The date and time of the estimate (i.e. observation)</td>
</tr>
<tr>
<td>TT_ESTIMATE</td>
<td>NUMBER</td>
<td>Travel time estimate</td>
</tr>
<tr>
<td>TT_CONFIDENCE</td>
<td>NUMBER</td>
<td>Confidence level in the travel time estimate for this segment and time</td>
</tr>
</tbody>
</table>

**QUERYING THE ARCHIVE**

To see how querying the proposed archive would work, it is useful to start with an example. We may want to generate the estimated travel time to travel on I-66 between Manassas and Fairfax in Northern Virginia on a particular day and time interval. In a travel time archive, we will have the ability to retrieve many different estimates from different sources at different times over different spatial extents. This leads to the question, what is the “best” travel time estimate? How do we know which estimates are better? Can we combine estimates somehow? Figure 2 presents an overview of the query process, with details provided following the figure.

![Figure 2 – Overview of Travel Time Query Process](image)

Let’s begin with the process of building the query. A typical query for a travel time estimate will be composed of two basic elements:
- Spatial Extent (i.e. Start of segment / End of segment)
- Time extent

Let’s return to our example: we want a travel time estimate for eastbound I-66 between Manassas and Fairfax at a specified time extent. We might have travel time estimates in the archive for that exact segment, or a segment that overlaps with that segment, or a segment that is “close”. We may choose to use one, some, all, or none of these segments.

Naturally, the estimates in the archive will be for many different times. Let’s assume we wish to know the travel time for the Manassas-Fairfax segment on Monday June 2 between 3:00 – 3:15 PM. Our archived travel time estimates can either match exactly that time, come close to that time (by some criteria), or be too far away (by some criteria).

The problem can be expressed in a more general form:

- **Exact**: Archived travel time estimate exactly matches the query parameter.
- **Intersects**: Archived travel time estimate intersects with the query parameter.
  - Degree of intersection: How much do the extents intersect?
- **Disjoint**: Archived travel time estimates do not match the query parameters
  - Proximate: The disjoint extents are proximate by some threshold.
  - Distant: The disjoint extents are not proximate.

**Figure 3 - Illustration of Parameter Matching**

**Determining which Estimates to Retrieve**

*Exact match*
Naturally, an exact match will be retrieved from the archive. If we have travel time estimates in the archive for the same exact spatial extent and time extent, those estimates will be used in satisfying the query.

*Intersects*
If a spatial or time extent in the archive intersects with the extent specified in the query we then consider the degree to which the extents intersect. In some cases we will accept estimates that intersect and in other cases we will reject them. This will play a role in how the data is weighted.
**Disjoint**

A disjoint spatial extent or time extent is one where the extents do not intersect (either in space or time). For example, if we have an estimate for a spatial extent that is disjoint from the requested extent but is proximate to the requested extent we may decide to use it. Likewise, if the time parameter is disjoint but proximate we can look at the distance of the estimate from the query parameter (in units of time) and decide whether or not to use it.

**DATA FUSION**

Once we have retrieved travel time estimates from the archive we begin the process of fusing the estimates together to produce a final travel time estimate for the specified query parameters. We will use three data fusion techniques.

Data fusion is defined as any set of techniques and tools used to combine measurement data from multiple inputs into a common representational format. (5) The goal of data fusion is to improve the quality of the information produced such that it is better than if each distinct measurement were used separately. Figure 4 illustrates this process.

![Figure 4 – Visualization of the Data Fusion Process](image)

Travel time can be expressed as a function of space, time, and source. In practice, travel time estimates are computed from the space-mean speed estimate $s$. The space-mean speed is the average speed to traverse some arbitrary link. Travel time and space-mean speed are related by the following equation (we assume speed, $s$, in mph, travel time, $TT$, in hours, and link length, $L$, in miles):

$$\frac{L}{s} = TT$$

(1)

Our data fusion function will be a function of two variables:
\[ s_i = \text{the space — mean speed for the } ith \text{ estimate} \]

We can compute the new space-mean speed estimate using a weighted mean. The weighted mean is defined by:

\[ \frac{\sum_{i=1}^{n} w_i s_i}{\sum_{i=1}^{n} w_i} \]  

(2)

The weights will be a function of the three dimensions of travel time: spatial extent, point in time, data source.

**Spatial Fusion**

In spatial fusion, we attempt to fuse travel time estimates for different spatial extents. We can let the weights be a function of the percentage overlap of the archived spatial extent with the queried extent (exact, intersects, disjoint). In some cases archived data will not be available for a portion of the queried extent. In these cases we can attempt to use a proximate extent as a substitute for the missing data or indicate that the data is not available. If a proximate extent is used, the weighting value for the estimate can be lowered to indicate a lower confidence in the estimate. An inverse distance factor can be used to lower the weight in the estimate (e.g. 1/distance). The weighting function for spatial fusion has the following form:

- **Exact**: weight = 1
- **Intersects**: weight = % of segment
- **Disjoint**
  - Proximate: % of segment * 1/distance

For example, travel time estimates for a particular spatial extent could be given the following weights from the spatial fusion function:

\[ w_1 = 0.5 \text{ (i.e. 50% intersection)} \]
\[ w_2 = 1 \text{ (i.e. exact spatial match)} \]

**Temporal Fusion**

A similar strategy for fusing estimates across time can be used. If we want to know the travel time for a particular spatial extent at a particular date and time we may need to consider using estimates from times that are “close” to the specified time parameter. The weight of the estimates can be determined using an inverse distance function.

The distance function is based on day of the week and time of day. Estimates from the same day of the week at the same time are considered to be “close”. Moving away (in either
direction) from the target time is adding to the distance and therefore lowering the weight of the estimate. If we are estimating travel time on a Wednesday at 3:00 PM then estimates from other Wednesdays at 3:00 PM would be weighted higher than estimates from Saturday at 3:10 PM. A maximum distance parameter can be defined to indicate which travel time estimates are too distant to be used in the fusion function.

Source Fusion

The third fusion function we will need is a function to fuse travel time estimates from multiple sources. In this case, the weighting function will be a function of the confidence in the source’s estimate. One way in which we can calibrate the confidence level in a travel time source is to periodically use GPS probe data and compare the estimates produced by the source against the travel times reported by the probe vehicle. The errors in the estimates may then be used as weights for that particular source.

Definition of Weighting Function

The weighting function is a product of the three weights described above: spatial, temporal, source.

\[ w_i = (\text{spatial weight}) \times (\text{temporal weight}) \times (\text{source weight}) \]  

(4)

The weighting function will have a maximum value of 1 and a minimum value of 0. This reflects the weight given to the estimate in the weighted mean. As an example, if we have an estimate for a spatial extent and time extent that exactly matches our query parameters we would have a weighting function that looks like:

\[ w_i = (1) \times (1) \times (\text{source weight}) \]  

(5)

where source weight is the calibrated weighting factor given to the particular source’s estimates.

Returning the Results of the Data Fusion Process

The data fusion process is an iterative process. We begin by retrieving the results of the query for travel time estimates. Then for each estimate, an appropriate weight based on the spatial, temporal, and source parameters is determined. Each estimate is then weighted appropriately. The result of the data fusion process is a weighted space-mean speed that estimates travel time by fusing the estimates of many sources over an arbitrary spatial extent and time extent. The weighting values used in the data fusion process can be returned to the
user as a measure of the quality of the estimates. The quality measurement can be expressed as a product of the respective weights of the estimates input into the data fusion process.

CONCLUSION

Management of travel time data will be of great importance to transportation professionals. Both planners and operations specialists will have an interest in using travel time archives. Strategies for managing travel time data were discussed in this report. These strategies include developing a new data model to accommodate travel time data according to three distinct dimensions of the data: the source, the spatial extent, and the time of measurement. Three data fusion functions were mentioned as a final area of interest in this project. Data fusion will likely be a valuable component of any archiving strategy. A weighted mean function is proposed to fuse travel time estimates. Weighting values are derived from space, time, and source specific functions. These values are applied to estimates and a new estimate is derived from the weighted mean function.

REFERENCES

1. Smith, B.L. *Investigation of Travel Time Data Service Requirements*. Publication UVACTS-5-0-108, University of Virginia Center for Transportation Studies, 2002
3. Ou, Qing, O, Van Lint, H. Hoogendoorn, S. Piece-wise Inverse Speed Correction using Individual Travel-times”, Presented at 87th Annual Meeting of the Transportation Research Board.