OTREC

FINAL REPORT

OREGON TRANSPORTATION RESEARCH AND EDUCATION CONSORTIUM

Improving Travel Information Products via Robust Estimation Techniques

OTREC-RR-09-04
March 2009

A National University Transportation Center sponsored by the U.S. Department of Transportation's Research and Innovative Technology Administration

IMPROVING TRAVEL INFORMATION PRODUCTS VIA ROBUST ESTIMATION TECHNIQUES

Final Report

OTREC-RR-09-04

by

David Maier
Kristin Tufte
Rafael J. Fernández-Moctezuma
Portland State University

for

Oregon Transportation Research
and Education Consortium (OTREC)
P.O. Box 751
Portland, OR 97207

March 2009

Technical Report Documentation Page

1. Report No. OTREC-RR-09-04	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Improving Travel Information Products via Robust Estimation Techniques	5. Report Date March 2009
	6. Performing Organization Code
7. Author(s) David Maier Kristin Tufte Rafael Fernández-Moctezuma	8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Computer Science Portland State University P.O. Box 751 Portland, OR 97207-0751	10. Work Unit No. (TRAIS)
	11. Contract or Grant No. 07-64
12. Sponsoring Agency Name and Address Oregon Transportation Research and Education Consortium (OTREC) P.O. Box 751 Portland, Oregon 97207	13. Type of Report and Period Covered
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Traffic-monitoring systems, such as those using loop detectors, are prone to coverage gaps, arising from sensor noise, processing errors and transmission problems. Such gaps adversely affect the accuracy of Advanced Traveler Information Systems. This project will explore models based on historical data that can provide estimates to fill such gaps. We build on an initial study by Mr. Rafael J. Fernandez-Moctezuma, using both a linear model and an artificial neural network (ANN) trained on historical data to estimate values for reporting gaps. These initial models were 80% and 89% accurate, respectively, in estimating the correct speed range, and misclassifications were always between adjacent speed ranges (in paricular, the free-flow range and congested range were never confused). Going forward, we will investigate other non-linear models, such as Gaussian Mixtures, that provide further statistical metrics, in contrast to the uninterpreted weights of ANNs.

This work will exploit the Portland Transportation Archive Listing (PORTAL) at the Intelligent Transportation Systems Laboratory at PSU. Dr. Tufte helps supervise development of PORTAL, and Mr. Fernandez used PORTAL data in his study. PORTAL holds more than two years of Portland-area freeway-loop-detector data at both detailed and aggregated levels, and is an ideal resource for the proposed work.

Initially we will be building and testing estimators in off-line mode. We will select a highway segment (comprising multiple detector stations) that is representative in terms of pattern of outages. We will build models for this segment, then examine their performance on estimates for synthetic gaps (so we can compare estimates to reported values). Later, using live loop-detector data (which PORTAL supports), we will work towards on-line estimation over the local freeway network, which requires computing estimates in a timely manner. Our end target is improvements in end-user travel information products, such as the Portland-Metro Speed Map on ODOT's Trip Check.

Our main evaluation metric will be the trade-off curve bewteen accuracy of prediciton and percentage of gaps that can be filled.

This research supports national surface-transportation research priorities, including the Systems Management Information area (ITS JPO). Within that area, it relates to (2) Data Management (techniques and guidance for processing and managing data associated with highway and transit monitoring) and (5) Data Dissemination (exchanging information about transportation services and providing that information to travelers). [Page 3-15, U.S. Department of Transportation Research, Development, and Technology Plan, 6th Edition]

17. Key Words	18. Distribution Statement No restrictions. Copies available from OTREC: www.otrec.us

19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 48	22. Price

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Oregon Transportation Research and Education Consortium (OTREC) for their funding of this project and El Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support of Rafael Fernández- Moctezuma.

DISCLAIMER

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the material and information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation University Transportation Centers Program and the Oregon Transportation Research and Education Consortium (OTREC) in the interest of information exchange. The U.S. Government, OTREC, and CONACYT assume no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the U.S. Government, OTREC or CONACYT. This report does not constitute a standard, specification, or regulation.

IMPROVING TRAVEL INFORMATION PRODUCTS VIA ROBUST ESTIMATION TECHNIQUES

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Transportation-related data such as sensed data from inductive loop detectors and other sensors is subject to noise and loss due to communication failures, hardware malfunctions, software glitches and many other causes. In addition, the volume of transportation-related data collected is increasing as data collection becomes cheaper and easier. For example, the Portland-area transportation data archive, PORTAL, contains over 700GB of data. This archive contains many gaps due to missing data or invalid data values. The focus of this project is to investigate data imputation in a real-time context.

Performing imputation in real time has limitations that do not occur with imputation for archived data. This project involved an initial evaluation of several alternative imputation methods using inductive-loop data from PORTAL, and an analysis of the distribution of data gap length in PORTAL data. In addition, linear and nonlinear regression techniques were tested as possible imputation methods, and we have investigated the appropriate configuration of these models for the PORTAL loop detector data. The project's primary conclusion is that a successful system for filling missing values will require a combination of imputation methods. Different methods will be required due to different gap lengths and patterns of data loss.

1.0 INTRODUCTION

Transportation-related data is being collected in increasing volumes for systems analysis, operations support, traveler information and many other uses. Sensed data, such as data from inductive loop detectors, is subject to noise and loss from a variety of sources. In archival sources, a missing or suspect datum is sometimes replaced by an *imputed* value, an estimate of what the value would have been if correctly observed. Such data replacement is usually performed with some knowledge of the underlying generating process, either based on theoretical considerations or empirical models. In the archival setting, imputation can be performed off-line in batch mode at particular time intervals. A computationally intensive process is acceptable if it can be amortized over many values.

Within a larger context, there lies an interest in processing live data streams to support (near) real-time information products, such as speed maps and travel-time estimates, and enhancing such streams through imputation of missing values in a way that introduces minimal delay. Thus, there are additional requirements on the imputation methodology beyond those found in the archival setting. One is that imputation is "temporally one-sided," in which case users are restricted to methods that only require inputs from the current time or the past. Second, the methods must be computationally efficient on an individual-value basis, as imputation will take place for one or a few values at a time. As with the off-line case, estimation accuracy is important. However, what constitutes an appropriate level of accuracy can only be determined in the context a particular end-use of the data stream and its requirements.

This project involved an initial evaluation of several alternative imputation methods in light of these requirements. We have used actual inductive-loop data from the Portland metro area that was taken from the Portland Transportation Archive Listing (PORTAL).

This report begins with a brief overview of imputation strategy as it relates to the project. It then describes representative imputation methods including heuristic techniques (such as rolling forward the previous known measurement and using historical averages) as well as statistical techniques (such as linear and nonlinear regression models). The techniques reviewed illustrate different rationalizations of the data imputation process, in particular, temporal and spatial frameworks. An effort was made to guide the reader from the simple to higher order methods, pointing out advantages and disadvantages of each.

We then describe the highway segment that was evaluated, present a discussion of its high-level traffic dynamics and provide exemplar data. An initial examination shows that there is indeed correlation in data between different sensor stations, which is a necessity for some of the spatially based imputation methods.

Some methods appear sensitive to the length of periods of consecutive missing values (*gaps*). Thus we analyzed the lengths of gaps in a sample of PORTAL data to see if long gaps were a common occurrence, under different definitions of what constituted "missing" data. While there was no consistent pattern of gap duration across different conditions, long gaps (> 1 hour in duration) sometimes made up more than half of the missing data. Given that long gaps are not an infrequent occurrence, we tested the sensitivity of the simple roll-forward heuristic against gap length. Indeed, estimation accuracy declined with increasing gap length. Thus, it appeared worthwhile to evaluate other imputation methods to see if they could provide improvements for larger gap durations, with the roll-forward error providing the target for improvement.

We tested both linear and nonlinear regression techniques as possible imputation methods, detail work on Gaussian Mixture Models (GMM). Unlike the simple heuristic case, there is a need to choose an appropriate configuration of the model, both in the inputs to include in the model as well as the appropriate number of components to use (which depends on the input choice). Multiple input choices (involving upstream stations, downstream stations or both) were examined and the report shares both costs of parameter exploration and training as well as the error realized with each. We also summarize the results of all the techniques evaluated, both for estimation error and relative time to impute a missing value using different methods. Finally, the report provides an example of analyzing the sufficiency of a given level of estimation performance relative to the requirements of a particular end application.

Our main conclusion from this work is that a successful system for filling missing values requires a combination of imputation methods. Different methods perform better for different gap sizes and likely the cut-over points will vary across individual sensor locations. However, it is not sufficient to provide just the method with the best estimated error for a given location. There are reasons for using a suboptimal technique in certain situations. One is that, in some cases, estimation accuracy may be traded for resource usage. In the face of “bursts” of gaps, a lower-accuracy method that consumes fewer resources may be necessary. A second reason is that the best method may require inputs that are unavailable because of simultaneous missing values at multiple locations.

This insight helps set the course of the work ahead. Multiple imputation methods are required, some of which have significant configuration requirements. The approach so far has been a largely manual process, which will not scale to a full highway network. More automated approaches to configuration and training are needed, as well as assessment of which methods work best at particular stations or particular station types (for example, stations with no downstream station, or stations near splits and merges in the network). Further, instrumentation to monitor the performance of the models must be provided, and they must be reconfigured or retrained on more recent data. Finally, there is a need to provide for the dynamic selection of imputation method for each particular missing value, based on expected accuracy, resource requirements and availability of correlated values.

2.0 IMPUTATION STRATEGY

When imputing missing values, one tries to answer the question of what the value would have been if conditions preventing the observation were not present. There are various paradigms one can follow to reason about missing values in the transportation domain: (1) Can a missing reading in a station be inferred from previous observations of that particular station? (2) Can one infer the missing value by looking at available values in other stations? (3) Should one consider temporally co-occurring measurements or should one reason about the immediate past? (4) Should estimation occur in a single evaluation or multiple evaluations? These questions open up numerous possibilities for imputation, nicely categorized by Ni et al. [18] in a multidimensional framework based on domain, methodology and imputed quantity. In Ni's framework, an imputation strategy consists of a choice of methodology over a particular domain using a specific input parameter.

In the imputation literature it is frequently reported that the American Association of State Highway and Transportation Officials (AASHTO) does not recommend the imputation of erroneous or missing values in traffic data programs [7][18][21][23]. The motivation for this recommendation is the lack of ability to quantify introduced errors by the imputed values. As discussed in Section 3.0, the effect of imputation can be measured in terms of the application it serves. In addition, a data archive can track which values were imputed as meta-data, so that applications may use the original non-imputed data if desired. Suggestions to amend the AASHTO guidelines have also been reported in the literature, primarily because of the usefulness of imputation as perceived by practitioners [21].

In addition to Ni's framework, one can expand the categorization of imputation methods for station data looking at two dimensions: time and space. Temporal and spatial methods are illustrated with a high-level abstraction in Figure 2.1 and Figure 2.2. In both, the station of interest is labeled "B." Methods may rely on historical data to parameterize mechanisms for online use, but the scope of the historical measurements used differs. A temporal method, such as using the time-of-day historical average, will only require previous knowledge of the measured quantity. The scope of this method is shown as a dashed, blue rectangle in Figure 2.1. For fitting, spatial methods will require historical measurements of more than one station, as shown in a solid, red rectangle in Figure 2.1. At evaluation time, spatial methods require access in the same time frame to correlated spatial sources, shown as a solid box in Figure 2.2, while temporal methods require historical access to previous measurements, shown as a dashed box in Figure 2.2. Hybrid approaches may combine both temporal and spatial extents.

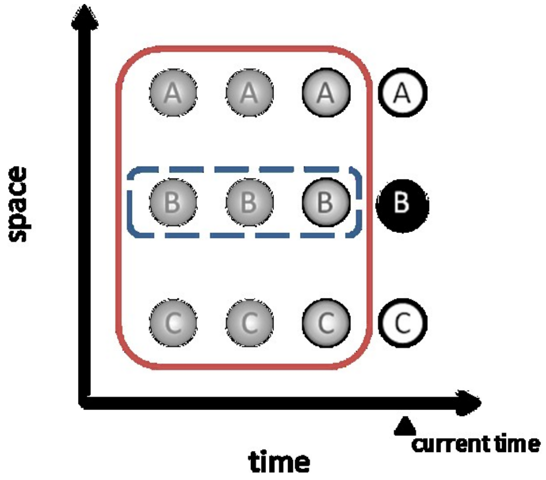


Figure 2.1 Archival lookup abstraction.

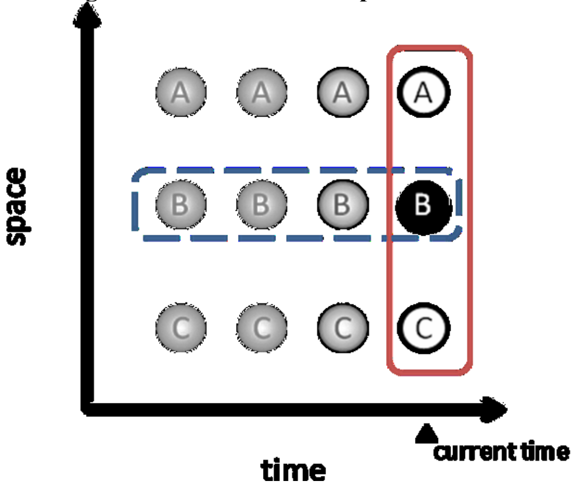


Figure 2.2 Required input abstraction.

Our particular interest is the enhancement of live streams of traffic measurements in the support of (near) real-time information products, such as speed maps. In this project, we focus on methods that are strictly usable for online imputation. In particular, these methods can only rely on data up to the time of the measurement of interest. In Figure 2.2, the set of available data excludes measurements to the right of the current time. Of online methods we consider:

Self-contained Methods – Self-contained methods assume that the observed behavior of a sensor or station can be expressed completely in terms of itself. In general, these methods work by finding historical properties from which a quantity can be inferred and used to replace missing values. However, they are not robust when there are long gaps in the data from a sensor station. A simple example of such a method is replacing a missing value for a sensor by the historical mean value for that sensor.

Correlated Methods – Correlated methods can ameliorate the effect of long periods of failure if the values used as input for these methods come from other sources likely to be online. In general, either contiguous-lane measurements or measurements from other stations can be used to estimate the value of a missing quantity. The simplest approach used to incorporate information from other stations consists of averaging the measurements of the upstream and downstream detectors, and is briefly mentioned in the literature [18][19]. This simple method assumes station correlation and equal contribution from neighboring stations, and is probably the simplest correlation model available, in a way analogous to historical mean imputation.

3.0 IMPUTATION METHODS

There are a wide variety of methods that are applicable to imputation of loop detector data, ranging from straightforward time-of-day imputation to more complex statistical techniques. We describe three types of techniques in this section: time-of-day imputation, regression — including linear and nonlinear regression — and Monte Carlo sampling. Later sections describe the experimental results of applying these techniques.

3.1 TIME-OF-DAY HISTORICAL MEAN

A simple mechanism for imputing missing data is to replace a missing value with a historical average of that value. This technique is known as historical mean imputation. The underlying assumption is that the values observed for a particular quantity over time are the result of a probability distribution, whose expected value is the mean. For Gaussian distributions, the expected value is a simple arithmetic average.

Time of day is an important consideration for traffic data imputation as traffic measurements such as speeds and occupancies are likely to follow highly different patterns in peak versus off-peak periods. An approach that addresses this issue consists in maintaining averages of the previous n observations at time t for previous days. This method is commonly referred to as time-of-day (TOD) historical average [7][18] [19][21][23][29].

For example, if a reading is missing at 8:20:40 a.m., the mean of the previous n available observations at 8:20:40 a.m. is used to impute the missing value. Different design considerations can be heuristically built into such a model, for example, maintaining separate statistics for holidays, weekends, midweek days and incident conditions. During implementation, one must select a suitable choice of the number of days to include in the historical mean calculation. Too small a number may fail to minimize the effects of an outlier value. Too large a number can obscure seasonal variation (and can increase computational expense). For example, Conklin et al. [7] found that using 30 days of historical data to compute average volume yielded the best results on their experimental data sets, having searched between five and 47 days.

3.2 ROLL-FORWARD

A very simple imputation strategy is to repeat the most recent value observed when the current value is not present. This practice is commonly used and follows from traffic-flow properties assumed not to change dramatically over small periods of time. Roll-forward is expected to be a very effective imputation strategy for short gaps, but not an adequate method to use for long gaps, in particular during transition periods. The inadequacy of roll-forward when used for hour-long gaps is shown in Figure 3.1, which illustrates speed data on sensor station 295.18 on I-5

North during the morning peak period. In this case, over-prediction is most noticeable between 7:00 and 8:00.

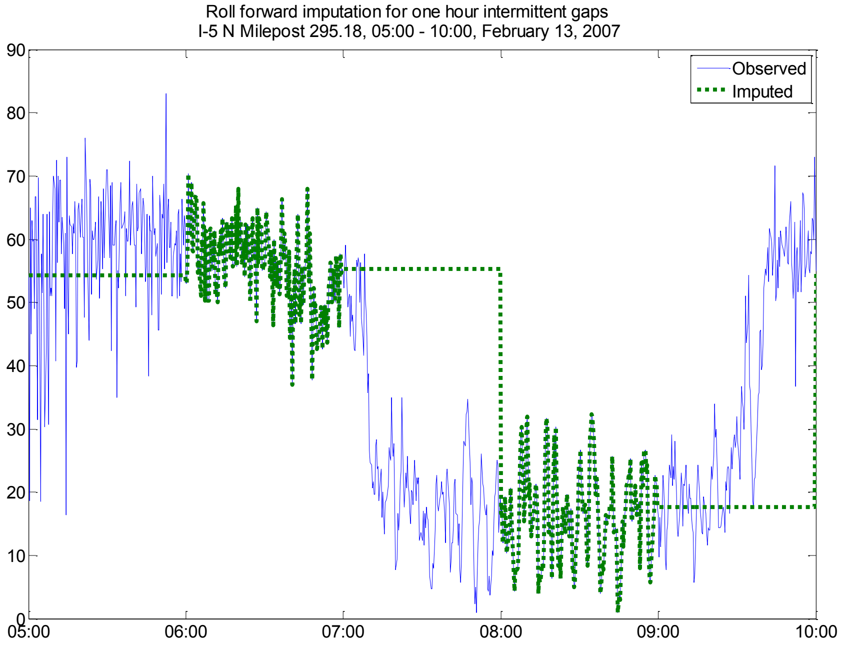


Figure 3.1 Roll-forward inadequacy during transition periods

3.3 REGRESSION

Regression analysis involves finding a functional description of an observed data collection, usually with the intention of predicting new values. This loose definition does not specify anything about the functional description, its inputs, or the mechanism for parameterizing the relationship. One could use regression to impute missing traffic using either linear or nonlinear functions. Inputs to the regression could be data from nearby stations, very recent data (last 30 minutes), or even historical data. In this section, we discuss both linear and nonlinear regression in the context of spatial correlations. We use data from nearby upstream and downstream stations as inputs. Parameterization is dependent on specific models.

Figure 3.2 shows an example freeway segment. In this segment, there are three sensor stations, labeled A, B and C. At each sensor station, there are three lanes with one detector in each lane. For ATMS and ATIS products, the individual lane measurements are typically combined into a single station reading, so that for each time step we effectively have one reading for each station (A, B, and C in this example). Applying our regression setup to this example, we would build a regression model of station B based on the data from stations A and C. Thus, at a time when data from station B is missing, but data from A and C is available, we can impute the data for station B based on the data from stations A and C using the model. Simplistically, the speeds recorded at stations A and C at 4 p.m. can be used to impute the value of the speed at B at 4 p.m.

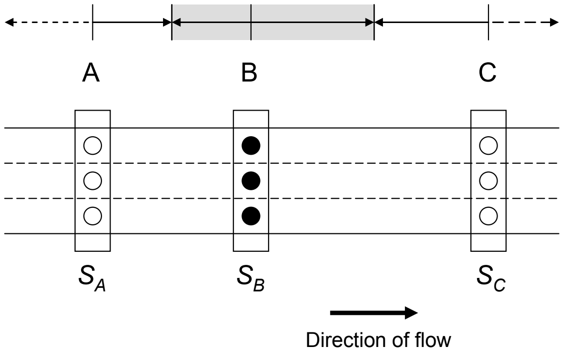


Figure 3.2 Example Freeway Segment

We note that while detectors provide point speed measurements, these measurements are often extrapolated to influence areas for products such as speed maps and travel-time estimation. A potential influence area for detector B is shown in light gray in Figure 3.2, defined from the midpoints of the location B with adjacent stations. For example, a speed map might report the speed for the length of the influence area to be the point speed reported by detector B.

Historical time-of-day imputation described in the previous subsection would impute missing data for station B from data previously received from station B. In contrast, our regression implementation imputes missing data for station B from data received from nearby stations A and C in the same time frame as the missing data.

3.3.1 Linear Regression

In this subsection, we describe a method for using linear regression on concurrent values from nearby stations to impute missing data. In particular, we use the example of imputing missing data for a station B based on nearby stations A and C as shown in Figure 3.2. Under the assumption that the relationship between stations can be expressed as a linear function, a joint Gaussian probability distribution can model the relationship between the speeds at three locations. A multidimensional Gaussian distribution is parameterized with two statistics: the mean vector μ, and the covariance matrix Σ, both of which can be estimated from historical data. If we represent the inputs (stations A and C) with the random variable X and the target (station B) with the random variable Y, the parameters of the joint distribution $p(Y,X) \equiv \mathcal{N}(\mu,\Sigma)$ can be written in block form as:

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix},$$

where Σ_{YY} is the covariance matrix of the random variable Y, Σ_{YX} is the covariance matrix of the random variables Y and X, $\Sigma_{XY} = \Sigma_{YX}{}^T$, and Σ_{XX} is the covariance matrix of the input X. One can obtain a conditional probability density $p(Y|X)$ from **Bayes' Theorem**, which will also be Gaussian. The conditional probability density is obtained by dividing the joint density by the marginal $p(X)$, which in turn can be obtained by integrating the joint probability over Y, $P(X) = \int_{-\infty}^{\infty} p(Y,X)\,dY$. Since the expected value of a random variable distributed as Gaussian is the mean, we can evaluate the conditional mean for a given input X = x. We have a target function which is a simple linear regressor, expressed as

$$E[Y|X = x] = \mu_Y + \Sigma_{YX}\Sigma_{XX}{}^{-1}(x - \mu_X).$$

3.3.2 Nonlinear Regression

We describe several methods for using nonlinear regression to impute speed data, including Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM). It is hoped that nonlinear regressions will capture the subtle variations encountered when traffic conditions switch between regimes. In our recent paper [16], we tested imputation using five-minute resolution data and an ANN framework and observed promising results that encouraged further exploration of nonlinear regression, in particular with challenging datasets such as 20-second resolution datasets, which are very noisy. According to Ni's framework, the strategy used in this paper [16] can be categorized as domain = speed, methodology = nonlinear regression, imputed quantity = speed.

ANNs can provide good approximations when properly fit; however, it is often difficult to logically understand the meaning of and draw conclusions from the model parameters. ANNs are described by the number of hidden units, their activation function (such as a sigmoid), and a collection of weights and biases. One can not reason about the problem domain (correlation between stations, for example) in terms of the weights found in an ANN. In contrast, the weights obtained through linear regression may offer intuition regarding inputs that receive negligible weights.

The GMM is defined as a weighted sum of *c* independent Gaussian components. For example, to use the concepts introduced in the previous method, let us express the joint probability distribution as a GMM:

$$p(Y,X) = \sum_{i=1}^{c} \alpha_i \mathcal{N}_i(\mu_i, \Sigma_i) ,$$

where each *i*th component has its own mean and covariance, μ_i, Σ_i, and each is weighted by a mixing coefficient α_i. The mixing coefficients sum to unity: $\sum_{i=1}^{c} \alpha_{i_i} = 1$.

Similarly to our previous exposition on linear regression, one can also derive an expression for the conditional probability distribution $p(Y|X)$ using **Bayes' theorem**. The steps are identical, but the resulting expression has a regularizing weighting function *r(x)* determined by the Gaussian components. The expression for the conditional mean is our target function, which is a nonlinear regressor of the form

$$E[Y|X = x] = \sum_{i=1}^{c} r_i(x) E_i[Y|X = x],$$

where $E_i[Y|X = x]$ is the expected value of the *i*th component of the mixture

$$E_i[Y|X = x] = \mu_{iY} + \Sigma_{iYX}\Sigma_{iXX}{}^{-1}(x - \mu_{iX}),$$

and

$$r_i(x) = \frac{\alpha_i p(x; \mu_{\mathrm{iX}}, \Sigma_{\mathrm{iXX}})}{\sum_{j=1}^{c} \alpha_j p\left(x; \mu_{\mathrm{jX}}, \Sigma_{\mathrm{jXX}}\right)}.$$

Fitting the parameters of this model presents several challenges. For a particular choice of *c*, one has to find the optimal set of mixing coefficients, and the individual parameters of each Gaussian component. This problem can be posed as an incomplete data Expectation-Maximization (EM) problem, where the class labels of each data point are unknown and inferred iteratively [9][10]. The second general problem consists of determining a *suitable* number of components. Notice the careful choice of the word *suitable*, as determining the optimal number of components for an arbitrary input still remains an open problem.

One approach to determine the best number of components consists in iterating through a range of options and performing *k-fold cross-validation* for each choice. The cross-validation method (illustrated in Figure 3.3) consists of randomly splitting the training data into *k* disjoint groups (folds), and iterating over the set *k* times, each time making one group the evaluation set and the remaining *k-1* groups the training set. One can compute the regression error on the evaluation data group for each fold and average over *k*, using the average error as a measure of suitability for the choice *c*. One can then choose the number of components as the choice among the range that yielded the smallest regression error while training.

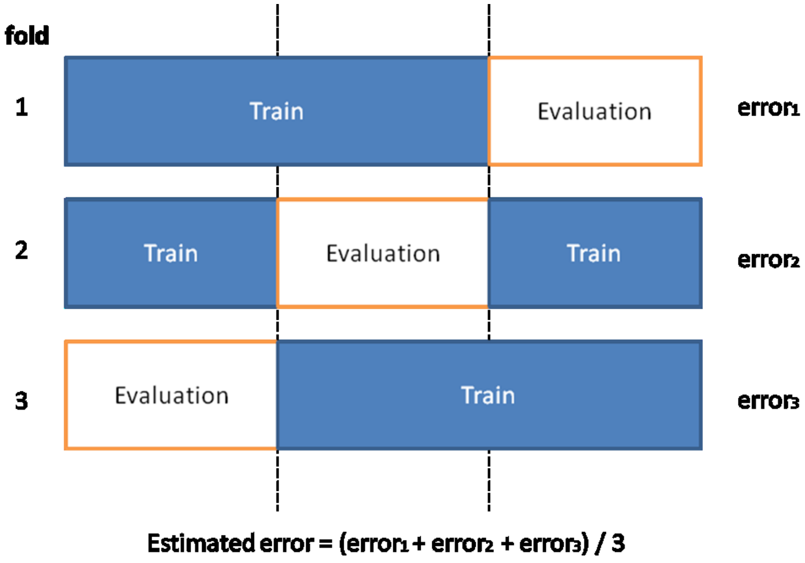


Figure 3.3 Illustration of 3-fold cross-validation.

While the concepts of finding the model parameters and choosing a suitable number of components are clear, there are a number of engineering considerations to be made. The first type of problem one encounters is the fact that the EM optimization begins with a random choice of components – this aspect means several restarts are required as one is not guaranteed to converge to a global optimum. The second type of problem comes from numerical stability: Matrices must be inverted during the fitting loops, and the initial start may have put them numerically close to a singular matrix, thus impossible to invert numerically. To cope with these problems, we call for another random restart when machine precision yields a singular matrix.

We also select the best of three runs for each fold to estimate the regressor error. Last, once a suitable number of components is selected, we select the best of three runs on the complete training set as the final model parameters. These choices are still not guaranteed to converge, and further restarts may be necessary.

3.4 MONTE CARLO SAMPLING

Randomized methods, commonly referred to as Monte Carlo methods, consist in drawing one or more samples from a particular domain and performing a deterministic computation on the samples to produce a desired output. Monte Carlo methods require the definition of a conditional probability distribution, such as the previous discussion on GMMs.

This imputation method will consist in drawing n samples from the conditional probability distribution $p(Y|X)$, by randomly selecting one of the Gaussian components and extracting the individual conditional expectation. While this method still uses nearby station measurements as inputs, the random choice of n conditional expectations and the average of those may produce reasonable results. Moreover, no further fitting is required, as this method is obtained “for free.”

4.0 GAP ANALYSIS

The need for imputation is due to missing or invalid data points. Missing or invalid data may occur in short or long intervals, depending on the cause. Communication errors can cause short-lived data gaps (gaps of one minute or less) while construction or loop-detector- cabinet damage will cause longer data gaps lasting days or even months. In addition to data gaps, imputation may be desired when invalid data is detected. Invalid data is typically data that falls outside the range of expected data values; invalid data may occur due to hardware problems at the detector location or software problems in the controller or the Advanced Traffic Management System (ATMS) itself.

Different imputation methods are applicable for different lengths of gaps. A simple roll-forward mechanism should be adequate for short gaps while more complex mechanisms will be necessary for longer gaps. In this section, we provide a brief analysis of gap patterns in the PORTAL data archive with the goal of understanding the frequency and occurrence patterns of gaps of various lengths.

4.1 GAP TYPES

Identifying invalid data is a complex issue in and of itself. The PORTAL database receives Portland-area freeway loop detector data from the Oregon Department of Transportation's (ODOT) ATMS in real time. Each data record consists of the following: timestamp, detectorid, speed, volume, occupancy and status flag. Data is provided at a 20-second granularity. That is, one record for each detector is received every 20 seconds. Data may be missing or invalid for several reasons:

- PORTAL did not receive data from the ATMS possibly due to ATMS failure, communication failure or PORTAL failure.
- ODOT did not receive data from a detector possibly due to a detector problem or communication failure.
- The data values reported fell outside of the expected range of data values. For example, a 20-second one-lane count of greater than 17 represents a flow rate of greater than 3,000 vehicles per lane per hour, which is highly unlikely [25].

For the purpose of this project, we consider three definitions of invalid or missing data as different applications may have different definitions of invalid data. In all categories, data values that were expected, but not received, by PORTAL are included in the count of invalid data. Note that PORTAL can identify data not received from the ATMS since PORTAL is supposed to receive one reading for each detector every 20 seconds.

ATMS No Data: ODOT has specified that records with a zero speed, volume, occupancy and status indicate that no data was received from the detector. In the ATMS No Data category, all data with reported zero speed, volume, occupancy and status are considered as invalid.

Data Quality Flags: The Texas Transportation Institute (TTI) has published a set of criteria for identifying potentially invalid data. The Data Quality Flags tests use those criteria to flag invalid data. In these tests, all data meeting the TTI criteria [25] listed below are considered invalid:

- Volume > 17 (20-second records)
- Occupancy > 95%
- Speed > 100 mph
- Speed = 0 when Volume > 0
- Speed > 0 when Volume = 0
- Occupancy > 0 when Volume = 0

Zero Speed: Observations of the data received from the ATMS over time have revealed that records with zero speed (but not zero volume, occupancy and status) often appear invalid. Many such records occur overnight. Theoretically, zero speeds are a valid reading; however, large numbers of zero speeds are highly likely during the low-traffic overnight hours. In the Zero Speed category, all records with zero speed and ATMS status flag other than "OK" (status = 2) are flagged as invalid.

4.2 GAP FREQUENCY

In order to develop imputation policies, it is important to understand gap patterns. The purpose of the tests described in this section is to understand the incidence of gaps of different lengths, whether there are many long gaps, many short gaps or both. In general, gaps occur either when the data is missing or when the data is flagged as invalid as described in the previous subsection. The unit of measure in the charts in this section is "gap time." That is, instead of counting the number of gaps of various lengths, we count the "gap time" attributable to gaps of different lengths. For example, four 20-second gaps contribute 80 seconds of "gap time"; one six-minute gap contributes 360 seconds of "gap time." By comparing gap time instead of number of gaps, we obtain a better understanding of the impact of gaps of different lengths. In all charts in this section, gap time is calculated over all main line loop detectors in the Portland area.

4.2.1 ATMS No Data

Figure 4.1 is a pie chart showing the gap time of data gaps due to the ATMS reporting No Data for the six months between November 2007 and April 2008. In total, the gap represented in this picture totaled 7% of all data for this time period. From this figure, we observe first that gaps of various lengths occur. The most dominant gap length is over six hours, which accounts for approximately 75% of total gap time. However, it does not appear that any gap length is rare enough to be ignored, so imputation strategies will be required for all gap lengths.

For further analysis, we compared gap patterns from three different weeks from three different months as well as daytime and nighttime gap patterns for this ATMS No Data category. We did not find significant differences between daytime and nighttime gap patterns; however, significant month-to-month differences were observed. Figure 4.2 and Figure 4.3 show the distribution of gap time of gaps due to the ATMS reporting No Data for a week in October 2007 and a week in February 2008, respectively. In total, the gaps represented in these pictures totaled 6% of all data for the October period and 7% of all data for the February period. As with the November 2007-April 2008 data, gaps of length greater than six hours are dominant; however, the number of gaps between six and 24 hours varies greatly between the two weeks.

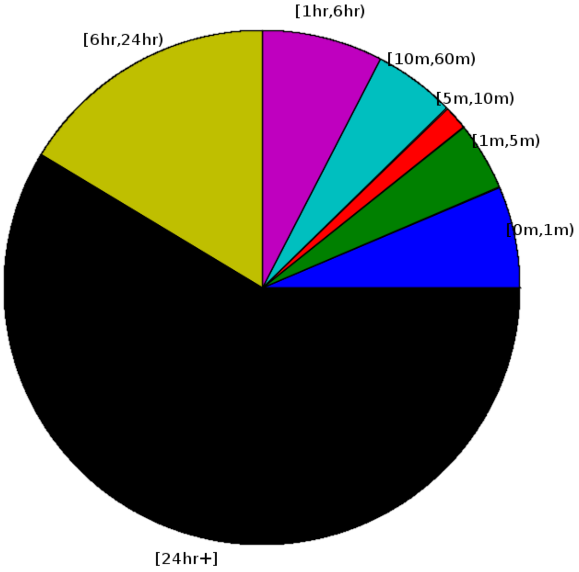
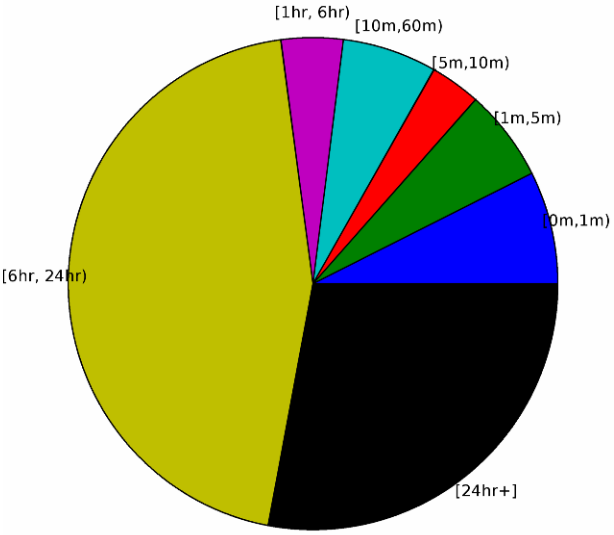


Figure 4.1 Proportion of Gap Time by Gap Duration ATMS No Data - Nov 2007 - Apr 2008



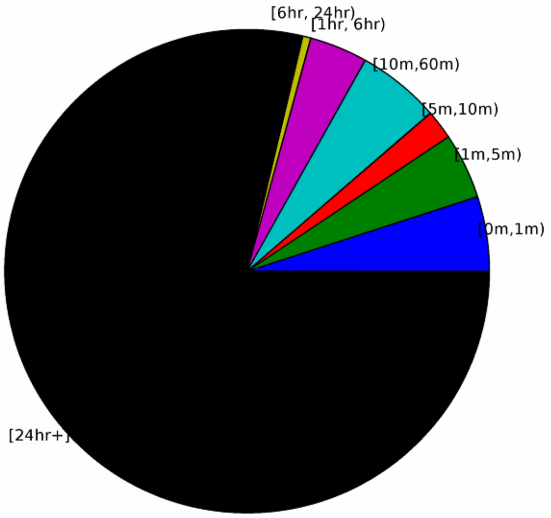


Figure 4.2 Proportion of Gap Time by Gap Duration ATMS No Data – October 2007

Figure 4.3 Proportion of Gap Time by Gap Duration ATMS No Data - February 2008

4.2.2 Data Quality Flags

Figure 4.4 shows gaps of various lengths as a percentage of total gap time, and gaps are defined as data failing the TTI data quality criteria described above along with missing data from November 2007-April 2008. As with the ATMS No Data definition for data gaps, we observe gaps of all lengths except for 24-hour gaps. Thus, no gap length dominates, again suggesting that a variety of imputation methods will be required. For this gap definition, shorter gaps are more dominant, with over 50% of the gaps for the November 2007-April 2008 period being less than five minutes. Figure 4.3 is a similar plot only for one week in October 2007. As with the ATMS No Data gaps, we observe significant month-to-month differences between the distributions of gap duration. In particular, in the November 2007-April 2008 period, there is a much larger proportion of gaps of less than one minute than in October 2007. Total gap time for November 2007-April 2008 and October 2007 is 3% and 9%, respectively. This is a significant variation in gap time due to failure of data quality tests. Nighttime versus daytime gap patterns were compared for this gap definition. It appears that there are more gaps of less than one

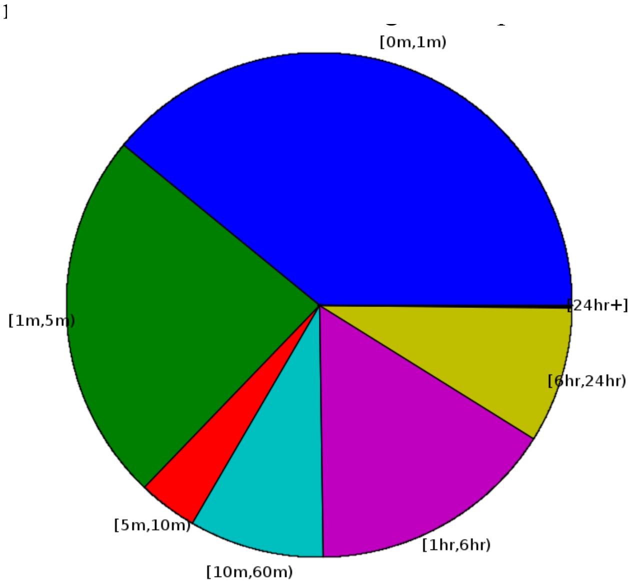


Figure 4.4 Proportion of Gap Time by Gap Duration Data Quality Flags - Nov 2007 - Apr 2008

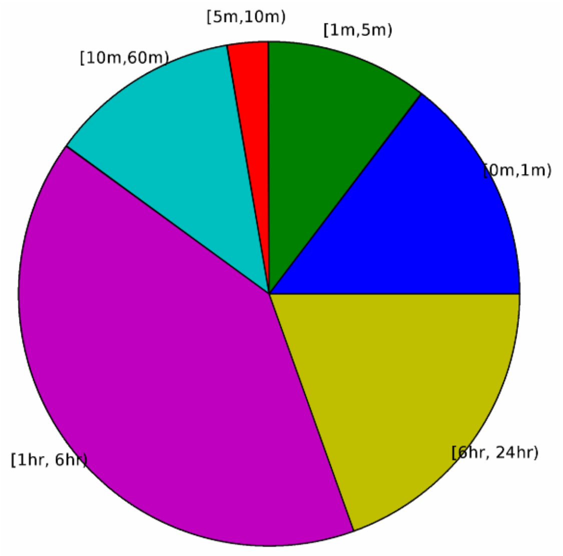


Figure 4.5 Proportion of Gap Time by Gap Duration Data Quality Flags - October 2007

4.2.3 Zero Speed

The final data-gap definition involves declaring data with a zero speed and a status flag other than "OK" invalid. Figure 4.6 shows gap time for November 2007-April 2008 for this definition of data gap. The patterns for the zero speed gaps are similar to the patterns for the gaps due to data-quality-flag failure. This similarity is likely due to the data quality test that declares data invalid if speed = 0 and volume > 0; thus both tests declare similar data invalid.

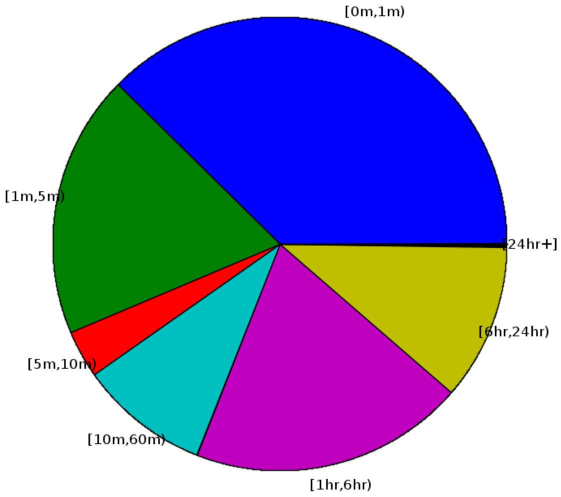


Figure 4.6 Proprtion of Gap Time By Gap Duration Zero Speed - Nov 2007 - Apr 2008

4.2.4 Summary

In summary, the gap distribution analysis shows the presence of gaps of all lengths, with no one gap length dominating. Roll-forward methods provide poorer estimates as gap length increases. While such methods are simple and inexpensive, they will have to be augmented with other imputation methods to handle gaps of other lengths.

5.0 EXPERIMENTAL FRAMEWORK

To test the performance of various imputation strategies, we selected a portion of the I-5 NB corridor. The selected segment is shown in Figure 5.1; the source image comes from Google Maps. The chosen segment includes five detector stations, labeled A through E. Traffic flows towards the top of Figure 5.1 and thus flows from station A to station E. The average detector spacing in this section is one mile. Station mileposts and descriptions for this segment are given in Table 5.1. A primary consideration in selecting this section was that it had a relatively complete set of data values. Completeness is important because we want to introduce gaps artificially, both so we can control their periodicity and duration and have the observed values to compare with our imputed values.

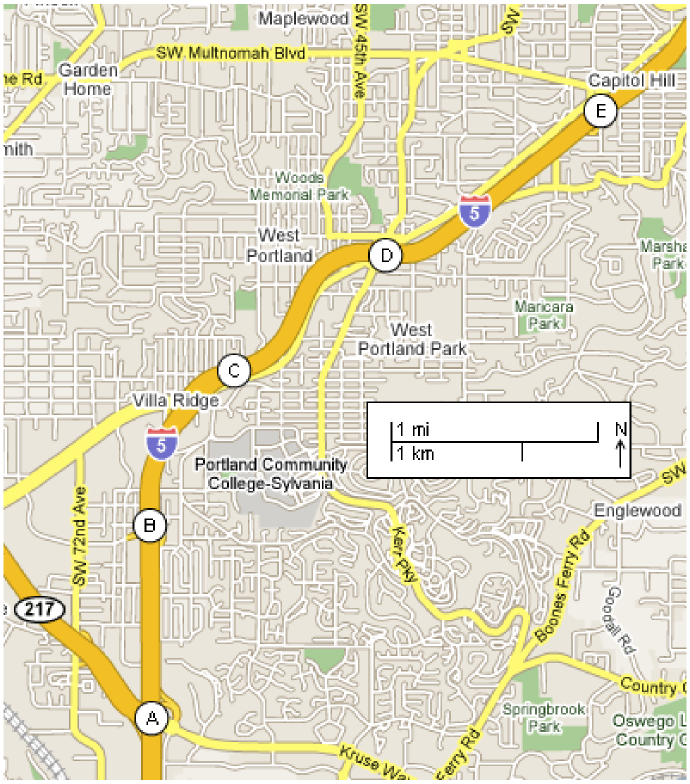


Figure 5.1 Experimental Segment - I-5 NB

Table 5.1 Experimental Segment – Station Descriptions

LABEL	DESCRIPTION	MILEPOST
A	Kruse Way	292.18
B	Haines Way	293.18
C	Pacific Highway	293.74
D	Capital Highway	295.18
E	Spring Garden	296.26

Our experiments impute speed for station C (as shown in Figure 5.1) at a 20-second resolution. Before presenting the experimental results, we explore the high-level traffic dynamics of this segment for one day. Figure 5.2 shows a surface speed plot for the experimental segment. This plot shows a change in traffic regime occurring during the morning rush hour, with a noticeable speed drop roughly between mileposts 293 and 300. Other surface plots for similar time periods (non-holiday weekdays at the same time of day) show that there is a recurrent bottleneck near milepost 300 on I-5 NB. (This bottleneck is, in fact, caused by the “Terwilliger Curves.”) We examine imputation for mid-weekday mornings from 5-10 a.m. The chosen time period includes regime changes (from free flow to congestion and back) to illustrate the imputation challenges one encounters. Notice that at this level of abstraction, lane measurements are aggregated to a station measurement, which is the target imputation quantity for this study.

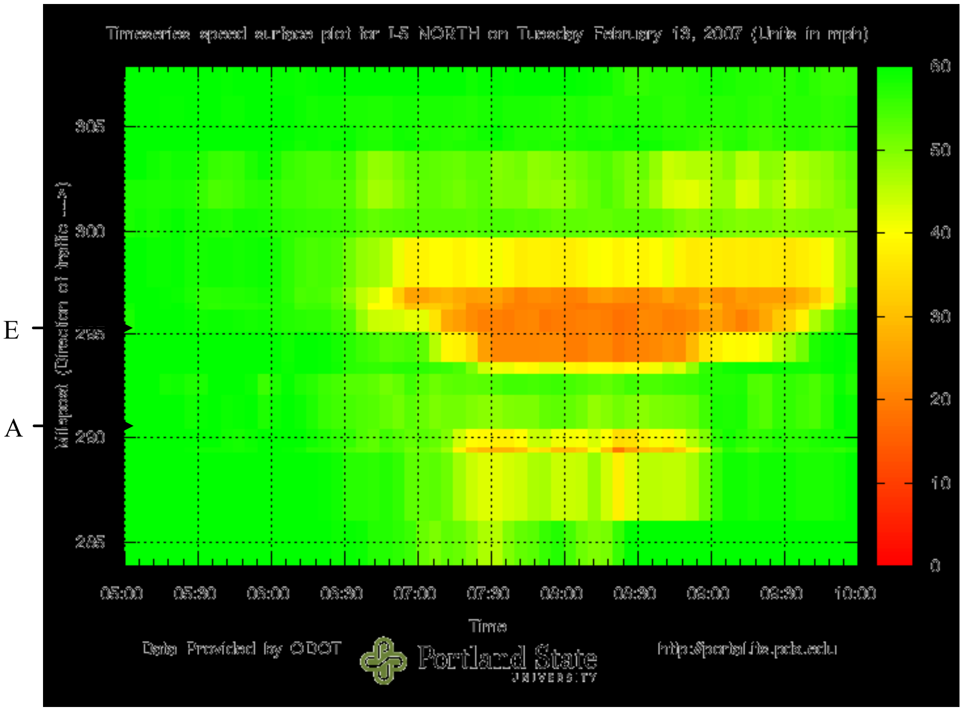


Figure 5.2 Surface Plot of I-5 NB - Feb 13, 2007 5-10 a.m.

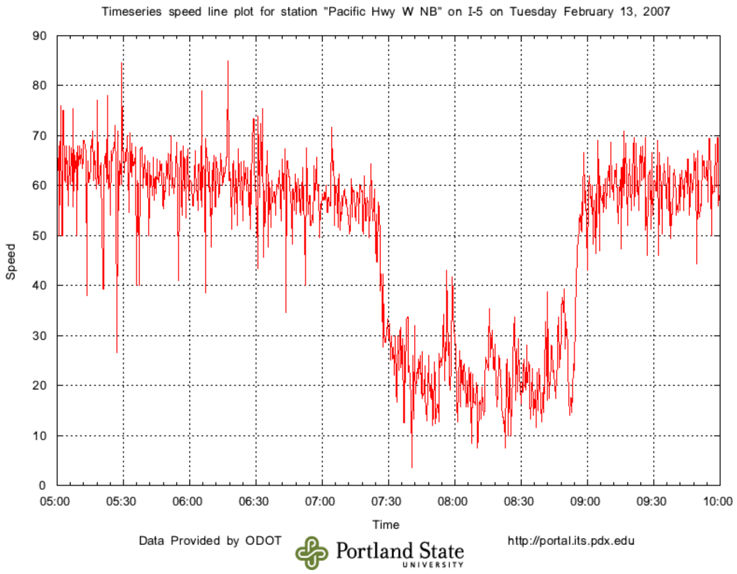


Figure 5.3 Time Series Speed Plot. Station C. Feb 13, 2007, 5-10 a.m.

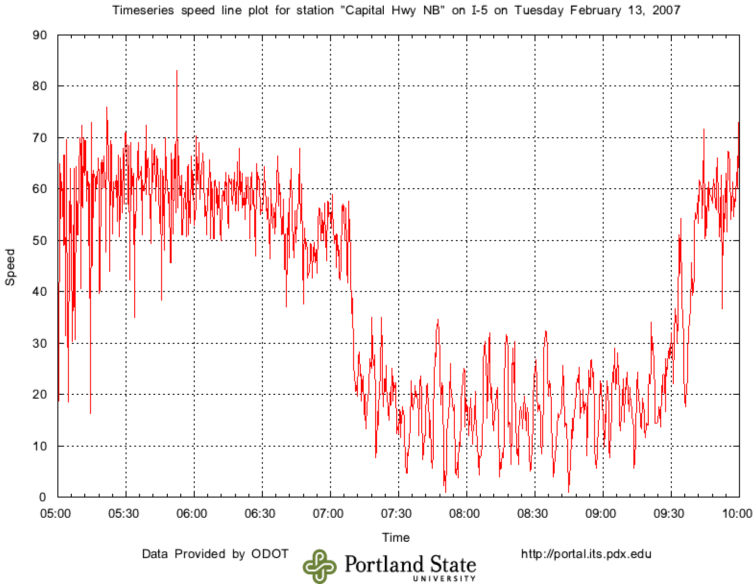


Figure 5.4 Time Series Speed Plot. Station D. Feb 13, 2007, 5-10 a.m.

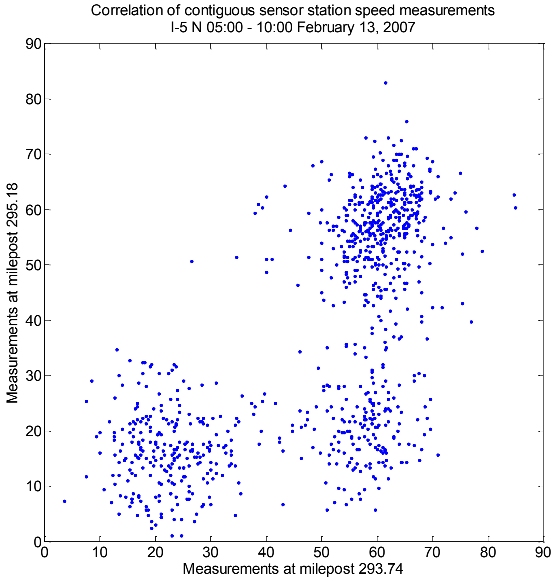


Figure 5.5 Correlation of speed measurements – two contiguous stations, C and D.

In the context of imputation, it is also convenient to examine whether the time series of speed measurements of stations near the studied station are correlated with speed measurements at the target station (station C). Such exploration also can be performed heuristically. Comparing time series of contiguous stations can help one understand if data from consecutive stations is correlated; if one observes matching trends, such as a decrease or increase in overall speed around the same time, then one can conjecture that those stations may be correlated. Figure 5.3 and Figure 5.4 show timeseries speed plots for station C and station D, respectively, for Feb 13, 2007 from 5-10 a.m. A quick visual inspection indicates that speeds at these two stations are correlated, with the congestion lasting somewhat longer at the upstream station (D). Furthermore, scatter plots such as the one in Figure 5.5 suggests a nonlinear correlation.

Several possible inputs are available for the sample scenario. The question is whether an imputation strategy should consider data from an upstream station, a downstream station, both, or multiple upstream or downstream stations. Several configurations can be assembled and compared to each other. Configurations are described in Table 5.2.

Table 5.2 Sample configurations for target output "C"

Inputs	Name	Intuition
B,D	Neighbors-1	Conditions in station C can be expressed in terms of its immediate neighbors.
A,B,D,E	Neighbors-2	Conditions in station C can be expressed in terms of its immediate neighbors and their neighbors.
D	Upstream-1	Conditions in station C can be expressed by looking one station upstream.
D,E	Upstream-2	Conditions in station C can be expressed by looking two stations upstream.
B	Downstream-1	Conditions in station C can be expressed by looking one station downstream.
A,B	Downstream-2	Conditions in station C can be expressed by looking two stations downstream.

Having described the experimental corridor and visually analyzed the traffic dynamics of this station, including correlations with recently recorded data and data from nearby stations, we proceed to present the results of imputation experiments on this corridor.

6.0 EXPERIMENTAL RESULTS

In this section, we illustrate roll-forward in detail as an exemplar of a temporal-based technique, and a GMM nonlinear regression as an example of a spatial-correlated technique, providing model choices and specific results for different choices of inputs from the described example corridor. Further discussion compares best results with the output of other methods, namely, simple linear regression and another nonlinear regression using ANNs. We also recommend a comparison strategy that helps select a subset of attempted methods and configurations.

6.1 EXPERIMENTAL SETUP

To implement the described techniques, we used the MATLAB environment on a desktop machine with a Pentium 4 processor and, 1 GB of RAM, running Windows XP. Datasets were obtained from the Portland ADUS (PORTAL) [3]. The datasets were extracted from February and March 2007 data, during the morning peak period 5-10 a.m., using 15 days for the training set and 10 days for the test set.

6.2 ROLL-FORWARD

Roll-forward is a simple implementation strategy that is expected to be effective for gaps of short duration. Further, roll-forward is computationally cheap compared to other imputation methods. Section 4.0 demonstrated that data gaps of all lengths, from short 20-second gaps to gaps greater than 24 hours, occur in the PORTAL data and that there is no particular gap length that dominates. As a consequence, imputation strategies for gaps of all durations are required. It seems clear that roll-forward will be effective for short gaps, but will break down for longer gaps. In these experiments, we analyze the effectiveness of roll-forward for gaps with varying lengths - from 20 seconds to one hour.

As stated above, it is believed that roll-forward will be effective for short gaps due to correlation between consecutive speed readings. Figure 6.1 shows the correlation between consecutive speed readings for the detector station at milepost 295.18 on I-5 NB. In other words, this plot shows the speed measured at time *t* (horizontal axis) vs. the speed measured at time *t + 20* seconds (vertical axis) for all times, *t*, between 5-10 a.m. on Feb 13, 2007, at milepost 295.18. Figure 6.1 demonstrates that, as expected, consecutive speed readings are highly correlated, supporting the conjecture that roll-forward will be an effective imputation method for short gaps.

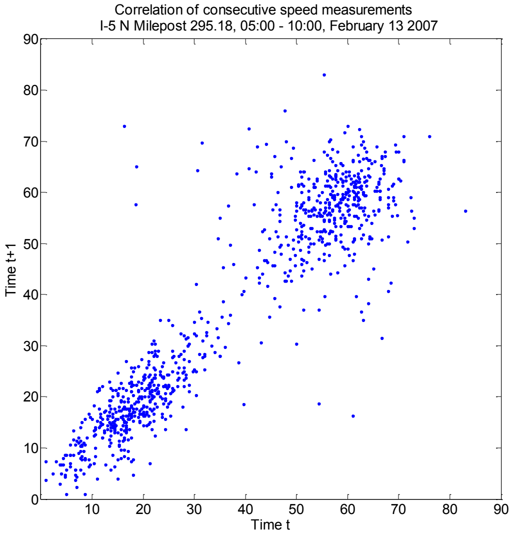


Figure 6.1 Correlation of consecutive speed measurements – time *t* and time *t+20*.

To understand the gap durations for which roll-forward is effective, we analyze the accuracy of roll-forward on gap lengths from 20 seconds up to 1 hour, a series of increasingly challenging gaps for roll-forward. We begin with a time series of speed measurements for 5-10 a.m. on Feb 13, 2007, for the station at milepost 295.18 on I-5 NB; this data set was selected because of its relative completeness. We induce gaps of lengths 20 seconds, one minute, five minutes, 15 minutes, 30 minutes and one hour on this time series. The gaps are introduced synthetically at regular intervals with alternating patterns of good and invalid data. For one-minute gaps, we alternate one minute of good data with a one-minute data gap. Gaps are induced by replacing observations with null values. We considered two disjoint gap patterns of the same length for each gap duration (i.e., for intermittent patterns of length 1 minute, we induced two patterns: preserve odd minutes and preserve even minutes). This selection helped to control for artifacts of gap placement. Each data value appears in one of the two gap patterns for each duration. Finally, the combined gap duration is the same across all cases.

Table 6.1 shows the mean square error (MSE) of the imputed data for the selected gap lengths. The mean square error quantifies the amount of difference between estimated and real values. The square root of the MSQ has the same units as the quantity being estimated. This table reports MSE for both of the disjoint gap patterns (as described in the previous paragraph) as well as the average MSE of the two patterns. Unsurprisingly, as shown in Table 6.1, roll-forward breaks down as the length of gaps increases, especially when the gaps occur during transition periods. The average MSE ranges from 63.54 for 20-second gaps to 360.06 for one-hour gaps. These numbers indicate that the average errors in terms of miles per hour are ~8 mph for 20-second gaps and ~19 mph for one-hour gaps. Figure 6.2 illustrates how the MSE worsens as gap lengths increase. We suggest that any method that attempts to address longer gap lengths should be at least as accurate for those gap lengths as roll-forward is. For example, if an alternative method yields an estimated mean square error of 170 and is not related to gap lengths, it probably could safely be invoked for this station when gap lengths exceed 15 minutes, as it is likely to provide better estimates.

Table 6.1 Gap length and MSE; I-5 NB, milepost 295.18, Feb 13, 2007, 5-10 a.m.

Gap length	MSE, configuration 1	MSE, configuration 2	Average MSE
20 seconds	62.06	65.02	63.54
1 minute	70.78	82.01	76.39
5 minutes	139.96	108.14	124.05
15 minutes	120.76	142.95	131.86
30 minutes	204.47	222.67	213.56
1 hour	68.58	651.54	360.06

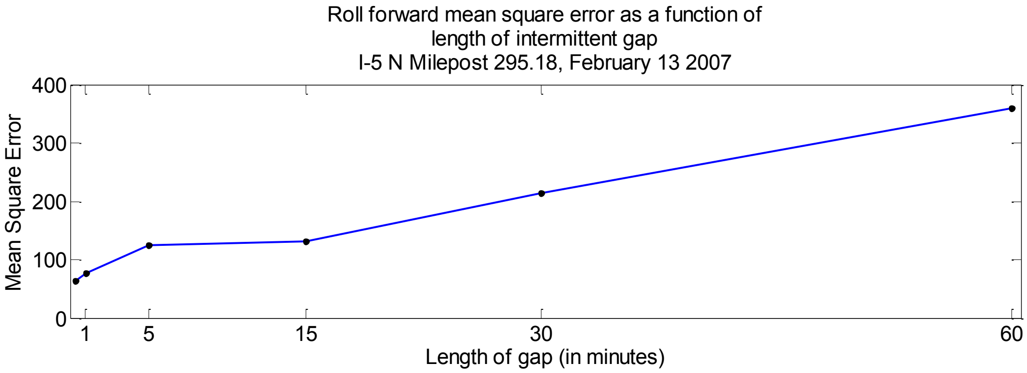


Figure 6.2 Gap length vs. MSE.

Figure 6.3 shows the observed and imputed speed time series for one-hour intermittent gaps for 5-10 a.m. on Feb 13, 2007, at milepost 295.18, I-5 NB. In this figure, from 7-8 a.m., the imputed speed is close to 55 mph, while the observed speed is much lower. Imputation error also occurs later, during the period from 9:30-10 a.m., as rush hour ends and congestion dissipates. Figure 6.3 corresponds to configuration 2, which exhibits the worst behavior for roll-forward during this time period. Under configuration 2, roll-forward has "unlucky" choices of the last value to roll, in particular since the gaps occur during transition periods.

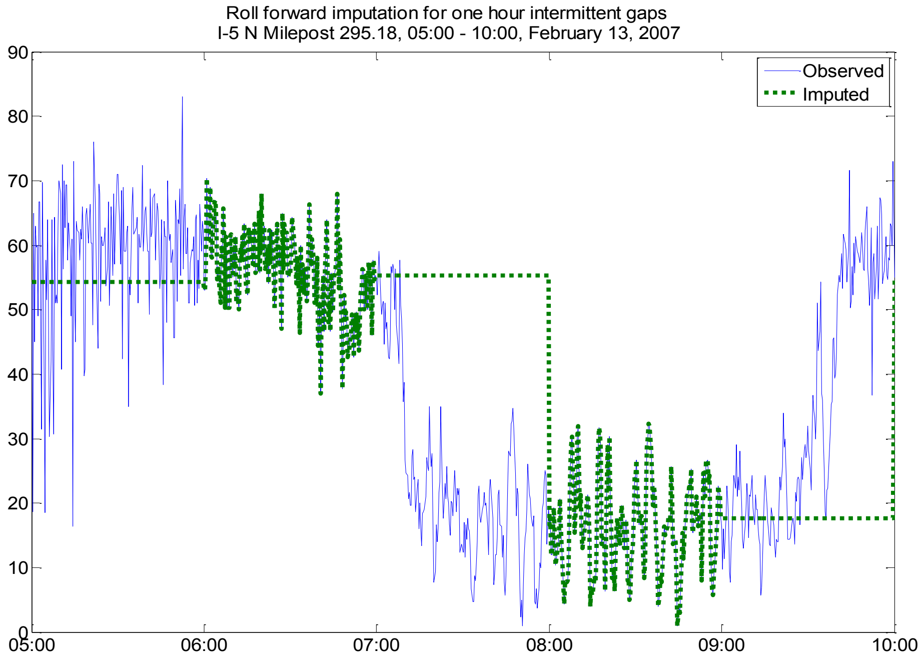


Figure 6.3 Roll-forward imputation for hour-long gaps, Configuration 2.

6.3 NONLINEAR REGRESSION

Nonlinear regression is based on constructing a nonlinear model based on data observations. As discussed before, several nonlinear models can be used to produce a regression function. To illustrate the process, we choose the GMM described in Section 3.3.2.

To address the possibility of divergence during the training set, we carefully select the starting points for the EM fitting process. First, we invoke a clustering algorithm, called *k-means* clustering [10]. K-means attempts to find the centroids of *k* clusters such that the variance of cluster elements is minimized. For a Gaussian mixture of *k* components, the centroids found by *k-means* become the initial parameters of the mixture's parameter fitting process. This intermediate step provides a better starting point to EM, improving its opportunity to converge to a suitable set of parameters that best represents the training set. Convergence on the estimator on the training set is shown in Figure 6.4. Speed data for morning periods for the 15 days in the training data is shown at a 20-second granularity. This visualization is not a complete time series, as only the morning peaks are used; however, model convergence can be appreciated by observing that the upward and downward trends are captured. Poor convergence, for example, could be manifested by a model that predicts a constant value.

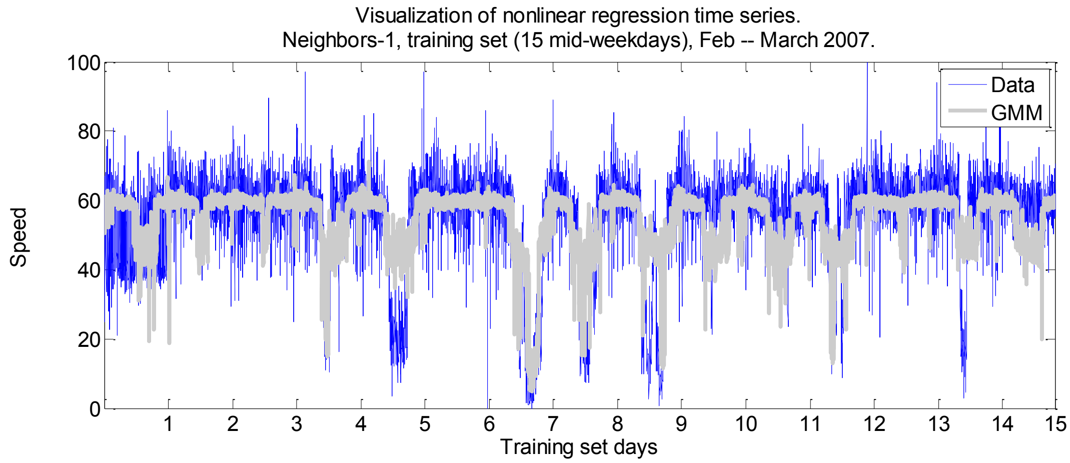


Figure 6.4 Visualization of nonlinear regression time series on training set.

Figure 6.5 shows the estimated MSE of a five-fold cross validation procedure on training data for different numbers of components. For the example shown, seven was a suitable number of components. The criterion is not just to choose the smallest error, but also to favor smaller models to avoid overfitting. In general, one can select, based on this criteria, by incrementally evaluating the regression error and not selecting a “best so far” number of components unless it improves the current error by more than 5%. In this example, seven components reduced error by more than 5%, but adding an eighth component did not reduce error by more than 5% and the ninth component actually increased error. For each of the six configurations listed in Table 5.2, Table 6.2 shows the selected number of components, the time required to select the number of components (searches were performed from two to 12 components), and finally the time to fit the model once the number of components was selected. The largest cost in terms of time investment comes from exploring candidate numbers of components. Fitting time can also be expensive and is related primarily to the number of components and inputs considered, but in this case was much less than component selection. Such behavior is expected in GMM model development.

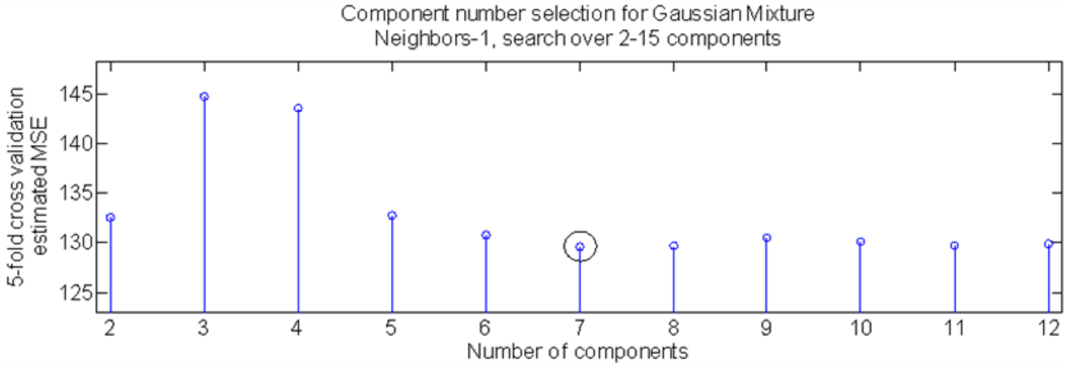


Figure 6.5 Component number selection for GMM

To fit the best possible model, we fit three models for each fold from the cross-validation partition, and then we choose the best of those three. Convergence is not guaranteed and sometimes may take longer or fail, resulting in noninvertible matrices or other precision issues. Downstream-1, for instance, failed to converge for model sizes larger than seven. Convergence fails due to numerical instability, redundant components, or empty components being formed and thus yielding noninvertible matrices.

Table 6.2 Number of suitable Gaussian components per configuration.

Configuration	Number of components	Exploration time (minutes)	Time to fit best model (minutes)
Neighbors-1	7	89.5	1.8
Neighbors-2	7	42	1.0
Upstream-1	11	319.4	22.8
Upstream-2	11	128.4	6.5
Downstream-1	5	37.4 *	3.0
Downstream-2	5	126.9	1.0

* Searched over 2-7 components only, as larger models fail to converge.

The ability to build several configurations serves the purpose of providing a menu of options, so one can fall back when all the inputs necessary for the best model are not present. Table 6.3 shows the performance of GMM over different configurations. From this table, one would conclude that station C is best modeled with configuration Neighbors-1; however, other options such as Upstream-2 and Downstream-1 may provide good "fall back" models if the required inputs for Neighbors-1 are unavailable. Table 6.3 suggests that for station C, looking at models beyond its immediate neighbors does not provide a significant advantage in terms of MSE.

Table 6.3 Performance of GMM over different configurations

	Neighbors-1	Neighbors-2	Upstream-1	Upstream-2	Downstream-1	Downstream-2
Training set MSE	128.89	120.15	149.95	146.77	171.41	158.04
Training set error variance	128.87	120.03	149.96	146.78	171.35	158.04
Test set MSE	54.84	55.77	60.91	60.81	61.74	64.84
Test set error variance	50.98	53.91	55.99	56.53	42.43	44.83

6.4 COMPARISON OF IMPUTATION METHODS

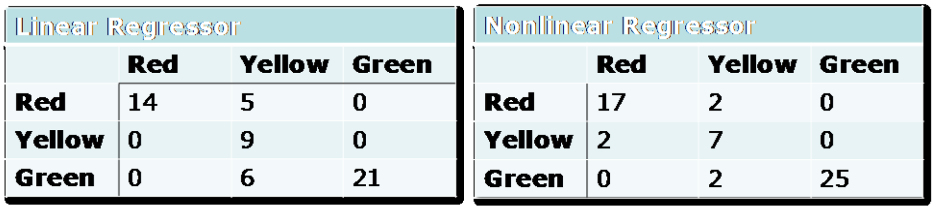
To choose an overall strategy, we compare the best model and architecture for each method and estimate the online evaluation cost. One may want to consider these alternatives in order to trade computational load and accept an estimated accuracy loss. Table 6.4 provides an example comparison under which, in terms of performance on a test set measured as MSE, the nonlinear regression mechanism yields the best performance. Estimating actual computation time from our MATLAB prototype is not easy, but can certainly be reasoned about. A TOD lookup is no more expensive than a database single-value lookup. If we take the computation time required to evaluate all test elements under a linear regression model as a baseline "x," Monte Carlo involves more computations (as the conditional expression of each Gaussian component involves matrix multiplications) and takes approximately 20 times more evaluation time than linear regression. The nonlinear regression evaluation takes up to 70 times more time than the baseline. To give an example of the data volumes processed, the total test size was around 13,000 data points and it took the nonlinear regressor 14 seconds to evaluate them all. We acknowledge that these time measurements are not exhaustive, but do expose variations in the cost of evaluation time for each method. These costs should be considered in an online environment, as one could potentially invoke a very expensive method at the same time for hundreds of stations. Having a mapping of accuracy and expense can be used for dynamic optimization of resource usage.

Table 6.4 Comparison of imputation methods.

Model	MSE of best model	Best configuration	Evaluation time
Time of day historical mean	106.1	*	Table lookup
Monte Carlo conditional sampling	99.0	Upstream-2	20x
Linear regression	58.3	Neighbors-2	x
Nonlinear regression	54.84	Neighbors-1	70x

* TOD relies only on historical means and does not look at other stations

Table 6.4 compares imputation strategies based on MSE. The gap pattern is a complete block out of one station. An alternative evaluation mechanism focuses not only on such statistical measures, but on the final effect a particular imputation approach has on a target application. For example, the intended use of an imputation mechanism may be an online congestion map. In this case, we assume the choice of method is not based on attenuating the variance; rather, it is based on whether the imputed data causes the "correct" color to be displayed on the speed map. More specifically, consider a congestion map that displays color-coded speeds: red for 0-25 mph, yellow for 25-50 mph and green for > 50 mph. If the imputed speed is 26 mph when the real data would have been 48 mph, the imputed data is "correct" as both 26 mph and 48 mph fall in the 25-50 mph range for the yellow color. This type of accuracy can be evaluated using confusion matrices that summarize the number of correct predictions (which appear in the main diagonal) and errors. Figure 6.6 compares a linear and nonlinear regressor (originally reported by Fernandez-Moctezuma et al. [16]), where the nonlinear regressor provides 9% better accuracy than the linear one. The colors red, yellow, and green correspond to speed cutoffs of 0-25 mph, 25-50 mph, and 50+ mph. Notice how no critical errors (i.e., predicting free flow when conditions are congested) are found with either method.



Linear Regressor

	Red	Yellow	Green
Red	14	5	0
Yellow	0	9	0
Green	0	6	21

Nonlinear Regressor

	Red	Yellow	Green
Red	17	2	0
Yellow	2	7	0
Green	0	2	25

Figure 6.6 Confusion matrices – linear vs. nonlinear regression.

In addition to comparing performance across methods, some design choices involving configurations for multiple input methods are required. Skyline plots provide a visual representation of the design space, as determined by computational expense and accuracy. An illustration of a skyline plot is presented in Figure 6.7. Notice the skyline dashed line –

configurations below it are clearly dominated by others, so implementing them in a production environment can be avoided.

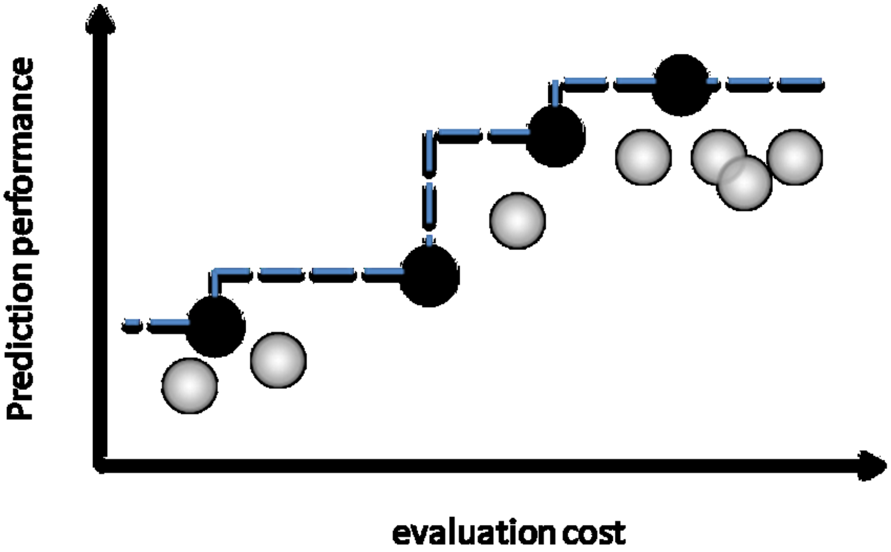


Figure 6.7 Illustration of a skyline plot.

7.0 CONCLUSIONS AND FUTURE WORK

This project has conducted an initial evaluation of several alternative imputation methods applicable to (near) real-time data imputation. Data for the study was obtained from PORTAL, the region's transportation data archive. The results demonstrate the strengths and weakness of various imputation methodologies. In addition, an analysis of the distribution of lengths of data gaps in PORTAL data was analyzed. The gap analysis results did not show a consistent pattern of gap duration across different conditions; however, long gaps (> 1 hour in duration) sometimes made up more than half of the missing data. The analysis of imputation methods indicated that the accuracy of the roll-forward heuristic decreases as gap length increases. Other imputation methods were evaluated to see if they could provide improvements for larger gap durations, with the roll-forward error providing the target for improvement. Linear and nonlinear regression techniques were tested as possible imputation methods and describe methodologies for choosing appropriate model configurations. Different choices (involving upstream stations, downstream stations or both) were examined. Our main conclusion from this work is that a successful system for filling missing values will require a combination of imputation methods. Different methods perform better for different gap sizes. However, it is not sufficient to provide just the method with the best estimated error for a given location. In some cases we may need to trade estimation accuracy for resource usage or due to unavailability of data.

8.0 REFERENCES

[1] S. Ahn, R.L. Bertini, B. Auffray, and J.H. Ross. Evaluating the Benefits of a System-Wide Adaptive Ramp-Metering Strategy in Portland, Oregon. *TRB 2007*.

[2] Paul D. Allison. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA, 2001.

[3] R.L. Bertini, S. Hansen, A. Byrd and T. Yin. PORTAL: Experience Implementing the ITS Archived Data User Service in Portland, Oregon. *Transportation Research Record: Journal of the Transportation Research Board*, Washington, D.C., 2004.

[4] Chao Chen, Jaimyoung Kwon, John Rice, Alexander Skabardonis, and Pravin Varaiya. Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems. *TRB 2003*.

[5] Mei Chen, Jingxin Xia, and Rongfang Liu. Developing a Strategy for Imputing Missing Volume Data. *TRB 2006*.

[6] B. Coiffman. Estimating Travel Times and Vehicle Trajectories on Freeways Using Dual Loop Detectors. *Transportation Research: Part A,* 2002, vol 36, no 4, 2002.

[7] James Howard Conklin, and William T. Scherer. Data Imputation Strategies for Transportation Management Systems. *Research Report, Center for Transportation Studies at the University of Virginia.* No. UVACTS-13-0-80. May, 2003.

[8] Carlos F. Daganzo. *Fundamentals of Transportation and Traffic Operations*. Elsevier, Oxford, UK, 1997.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, no. 1. (1977), pp. 1-38.

[10] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (Second Edition)*. New York, N.Y. Wiley-Interscience, 2001, pp. 282—333.

[11] Michael Falk, Frank Marhon, René Michel, Daniel Hoffman, and Maria Macke (eds.). *A First Course on Time Series Analysis: Examples with SAS*. University of Wuzburg, 2006. Version 2006.Sep.01.

[12] David L. Gold, Shawn M. Turner, Byron J. Gajewski, and Clifford Spiegelman. Imputing Missing Values in ITS Data Archives for Intervals Under 5 Minutes. *TRB 2001*.

[13] H. Hadj-Salem, J.M. Blosseville, and M. Papageorgiu. ALINEA: A Local Feedback Control Law for On-Ramp Metering; A Real-Life Study. In *Third International Conference on Road Traffic Control*, London, UK, 1990.

[14] D. Levinson and H. Huo. Effectiveness of Variable Message Signs. *TRB 2003*.

[15] D. Maier et al. NiagaraST. Available at http://datalab.cs.pdx.edu/niagara/. Last retrieved October 30, 2007.

[16] Rafael J. Fernández-Moctezuma, Kristin Tufte, David Maier, and Robert Bertini. Toward Management and Imputation of Unavailable Data in Online Advanced Traveler Information Systems. In *Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference*, Seattle, WA, USA, Sept. 30 - Oct. 3 2007

[17] Daiheng Ni and John D. Leonard II. Markov Chain Monte Carlo Multiple Imputation for Incomplete ITS Data Using Bayesian Networks. *TRB 2005*.

[18] Daiheng Ni, John D. Leonard II, Angshuman Guin, and Chunxia Feng. Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data. *Journal of Transportation Engineering*, vol. 131, no. 12, pp. 931—938, December 2005.

[19] Linh N. Nguyen, and William T. Scherer. Imputation Techniques to Account fo Missing Data in Support of Intelligent Transportation Systems. *Research Report, Center for Transportation Studies at the University of Virginia.* No. UVACTS-13-0-78. May, 2003.

[20] Brian L. Smith, and James H. Conklin. The Use of Local Lane Distribution Patterns to Estimate Missing Data Values from Traffic Monitoring Systems. *Transportation Research Record*, vol. 1811, pp. 50—56, 2002.

[21] Brian L. Smith, William T. Scherer, and James H. Conklin. Exploring Imputation Techniques for Missing Data in Transportation Management Systems. *TRB 2003*.

[22] Brian L. Smith, Simona Babiceanu. An Investigation of Extraction Transformation and Loading (ETL) Techniques for Traffic Data Warehouses. *TRB 2004*.

[23] Satish Sharma, Pawan Lingras, Ming Zhong. Effect of Missing Value Imputations on Traffic Parameters Estimations from Permanent Traffic Counts. *TRB 2003*.

[24] P.A. Tucker, D. Maier, T. Sheard, and P. Stephens. Using Punctuation Schemes to Characterize Strategies for Querying over Data Streams. *Transactions on Knowledge and Data Engineering*, Vol. 19, No. 9. September 2007.

[25] S. Turner. *Guidelines for Developing ITS Data Archiving Systems.* Report 2127-3. FHWA, U.S. Department of Transportation, Texas Department of Transportation and Texas Transportation Institute, 2001.

[26] R. Wang, M. Goto, and H. Nakamura. Validation of an Improved Method to Estimate Expressway Travel Time by the Combination of Detector and Probe Data. *Journal of the Eastern Asia Society for Transportation Studies*, Vol.5, October 2003.

[27] Yang C. Yuan. Multiple Imputation for Missing Data: Concepts and New Development. *SUGI Proceedings*, 2000.

[28] M. Zhang, T. Kim, X. Nie, W. Jin, L. Chu, and W. Recker. Evaluation of On-tamp Control Algorithms. *California PATH Research Report*, University of California, Berkeley. UCB-ITS-PRR-2001-36, December 2001.

[29] Ming Zhong, Satish Sharma, and Pawan Lingras. Matching Patterns for Updating Missing Values of Traffic Counts. *Transportation Planning and Technology,* vol. 29, no. 2, pp. 141—156. April 2006.

[30] Ming Zhong, Satish Sharma, and Pawan Lingras. Genetically Designed Models for Accurate Imputations of Missing Traffic Counts. *TRB 2004*.

[31] Ming Zhong, Satish Sharma, and Zhaobin Liu. Assessing Imputation Accuracy based on Traffic Count Data from Different Jurisdictions: Alberta and Saskatchewan Examples. *TRB 2005*.

[32] U.S. Department of Transportation, Federal Highway Administration; Institute of Transportation Engineers. *Intelligent Transportation Primer*. Institute of Transportation Engineers, Washington, D.C., 2000.

OTREC
OREGON TRANSPORTATION RESEARCH
AND EDUCATION CONSORTIUM

P.O. Box 751
Portland, OR 97207
www.otrec.us

OTREC is dedicated to stimulating and conducting collaborative multi-disciplinary research on multi-modal surface transportation issues, educating a diverse array of current practitioners and future leaders in the transportation field, and encouraging implementation of relevant research results.