**Transportation Research Record**

**IMPUTATION OF MISSING TRANSFER PASSENGER FLOW WITH SELF-MEASURING MULTI-TASK GAUSSIAN PROCESS**

--Manuscript Draft--

<table>
<thead>
<tr>
<th>Full Title:</th>
<th>IMPUTATION OF MISSING TRANSFER PASSENGER FLOW WITH SELF-MEASURING MULTI-TASK GAUSSIAN PROCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract:</td>
<td>Transportation data is of great importance for intelligent transportation control and management. Since current big data technologies can provide vast amounts of information on the status of traffic systems, data collection may be interrupted by technical failures and other practical issues. There is however, an increasing demand to carry out traffic analysis when a significant amount of data is missing. This study introduces the use of a self-measuring multi-task Gaussian process (SM-MTGP) method for imputing missing data. Particularly, the study focuses on the transfer passenger flow at a railway station collected by WiFi sensors. SM-MTGP exploits the temporal correlation between observations. First, correlations between tasks and inputs are learned simultaneously with the constructed two-way array. Second, covariances of features of these two aspects are measured nonlinearly with selected kernel functions. Third, additional knowledge is provided by the responses, helping get more insights into the similarities of the observations. The performance of the proposed method is assessed under several different missing data scenarios for a large-demand railway station in Melbourne using 6-months of real data. Test results show that: (i) the SM-MTGP has the lowest imputation errors; (ii) the SM-MTGP presents substantial improvement in reducing RMSE values by 60% over the base model; (iii) the SM-MTGP in particular outperforms the conventional approaches with large missing ratios. It is motivating to find the proposed method can achieve improved performance, our on-going effort is given on incorporating other features into this algorithm that can make further application on large-scale transit network analysis.</td>
</tr>
<tr>
<td>Manu Script Number:</td>
<td>19-04491R1</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Presentation &amp; Publication</td>
</tr>
<tr>
<td>Order of Authors:</td>
<td>Wenhua Jiang</td>
</tr>
<tr>
<td></td>
<td>Nan Zheng</td>
</tr>
<tr>
<td></td>
<td>Paul Reichl</td>
</tr>
<tr>
<td></td>
<td>Inhi Kim</td>
</tr>
</tbody>
</table>
IMPUTATION OF MISSING TRANSFER PASSENGER FLOW WITH SELF-MEASURING MULTI-TASK GAUSSIAN PROCESS

Wenhua Jiang
Institute of Transport Studies
23 College Walk, Monash University, Clayton, 3800 Victoria, Australia
Email: wenhua.jiang@monash.edu

Nan Zheng
Institute of Transport Studies
23 College Walk, Monash University, Clayton, 3800 Victoria, Australia
Email: Nan.Zheng@monash.edu

Paul Reichl
Institute of Railway Technology
17 College Walk, Monash University, Clayton, 3800 Victoria, Australia
Email: Paul.Reichl@monash.edu

Inhi Kim, Corresponding Author
Institute of Transport Studies
23 College Walk, Monash University, Clayton, 3800 Victoria, Australia
Email: inhi.kim@monash.edu

Word count: 4,750 words text + 4 tables x 250 words (each) = 5,750 words
Submission Date: 08-01-2018
ABSTRACT
Transportation data is of great importance for intelligent transportation control and management. Since current big data technologies can provide vast amount of information on the status of traffic systems, data collection may be interrupted by technical failures and other practical issues. There is however, an increasing demand to carry out traffic analysis when a significant amount of data is missing. This study introduces the use of a self-measuring multi-task Gaussian process (SM-MTGP) method for imputing missing data. Particularly, the study focuses on the transfer passenger flow at a railway station collected by WiFi sensors. SM-MTGP exploits the temporal correlation between observations. First, correlations between tasks and inputs are learned simultaneously with the constructed two-way array. Second, covariances of features of these two aspects are measured nonlinearly with selected kernel functions. Third, additional knowledge is provided by the responses, helping get more insights into the similarities of the observations. The performance of the proposed method is assessed under several different missing data scenarios for a large-demand railway station in Melbourne using 6-months of real data. Test results show that: (i) the SM-MTGP has the lowest imputation errors; (ii) the SM-MTGP presents substantial improvement in reducing RMSE values by 60% over the base model; (iii) the SM-MTGP in particular outperforms the conventional approaches with large missing ratios. It is motivating to find the proposed method can achieve improved performance, our on-going effort is given on incorporating other features into this algorithm that can make further application on large-scale transit network analysis.

Keywords: WiFi, Missing data, Imputation, Passenger transfer flow, Self-measuring multi-task Gaussian process
INTRODUCTION

Transportation data plays a critical role in intelligent transportation system management which has received strong interest among transportation practitioners and researchers. Traditional methods for travel data collection, including manual counts and surveys, can be limiting as they tend to be labor intensive, time-consuming and difficult to implement. Recently, with the development of advanced technologies, many sensing sources, have been employed for collecting traffic data, such as WiFi sensors, smart cards, Bluetooth, GPS and other technologies. These methods are gaining increasing popularity due to their distinct advantages such as cost-effective and rich data availability. Among these methods, WiFi technologies have been widely used in a range of transportation applications, including origin-destination (OD) matrices estimation, travel time measurement, and pedestrian flow analysis (1-3). However, missing data problems are inevitable during the data collection process for diverse reasons, such as sensors malfunction, connection failure or weather factors. Transportation data with missing information cannot be directly used in practical analysis if no efficient approach is applied to deal with this problem.

Much research has been devoted to handling missing data problems, aiming to make the most use of the data sets at hand for transportation system analysis. The recovery or reconstruction of the missing data is referred as data imputation in the literature. The commonly used imputation methods can be divided into three categories: interpolation, prediction and statistical methods (4). The interpolation-based methods are one of the most extensively utilized methods, which rely on historical data. The main work principle of this approach is estimating the missing points according to the average value of historical data at the targeted time period (5,6). Interpolation-based approaches assume that traffic patterns would be recurrent and nearly identical under for example the same weekdays or the same time periods. They thus ignore the random features such as variation and fluctuation of traffic flow (7). Another typical kinds of interpolation techniques are neighboring-based models, which use the information from one or more of the neighboring sites or neighboring states to achieve the estimation of the corrupted data. K-nearest neighbors (kNN) is a widely adopted method falling into this category, which is to find the nearest neighbors of targeted data points by using a specific distance metric to impute missing entries (8,9).

Regarding prediction models, autoregressive integrated moving average (ARIMA) (10) and Kalman filter methods (11-13) are the most common traffic forecasting models, which impute missing values by learning a linear estimator on historical traffic data of the modeled time series. However, these approaches do not yield good performance when a significant amount of data are missing (14-16). In addition, the observed entries after the missed points cannot be fully utilized in prediction models, which further decrease the amount of input to the imputation model and make the imputation results less accurate and ineffective (7).

Statistical learning approaches such as the probabilistic principle component analysis (PPCA) have been found to achieve good performance in managing missing data and can outperform conventional imputation methods under certain cases (4,17). The basic idea of PPCA is to assume a special distribution of the observed data and impute the missing data with values that fit the assumed distribution best. However, it is not useful when missing points occur over a whole row or column based on the structure of a two-dimension array (7). Matrix and tensor based methods have gained increasing interest due to their good performance of imputing traffic datasets (16,18). The basic framework of these methods is to minimize the reconstruction error by obtaining a suitable low-rank approximation matrix or tensor. However, as noted by Rodrigues et al. (19), the limitation of such methods is that they fail to measure imputation uncertainty.

In recent years, more advanced techniques such as, neural networks and deep learning approaches have been developed to handle missing values. Such methods have shown the ability to
produce relatively good results in filling missing transportation data when sufficient data is provided \((20,21)\). However, these approaches mainly focus on predictive ability without capturing imputation uncertainty, which can reduce the quality of the resultant data imputation \((16,19)\). Moreover, these methods which belong to the machine learning family usually require large-scale and detailed input to obtain (train) an appropriate prediction model which is typically unavailable or costly to access in practice. Based on the review, the limitations of existing imputation methods can be summarized as follows: first, potential useful information is not efficiently used in the modeling process. For example, global information is not fully utilized in some time series prediction models which only use information from one or several previous steps. Second, methods considering temporal correlation to achieve imputation usually rely on an assumption that linear relationships exist between observed variables and latent variables. Third, most imputation techniques fail to measure the uncertainty of the imputed values.

Despite all the challenges and issues for imputation of traffic data as mentioned above, there does exist other alternatives that may improve the accuracy and reliability of reconstructing the missing data. For this end, we propose the use of a self-measuring multi-task Gaussian process (SM-MTGP) method \((22,23)\) to investigate the temporal relatedness of a set of time-series observations of transfer passenger flow obtained at a railway station. The advantages of the proposed method can be summarized as follows: first, SM-MTGP is able to combine global information (features) from different sources (tasks and inputs) to measure similarities in a joint way. Second, the dependencies of various tasks and inputs is explored via covariance functions under the multi-task Gaussian Process’s (MTGP) framework which is a widely used non-parametric and non-linear model to learn complex processes. Third, in SM-MTGP, correlations between responses are captured to provide additional information for enhancing imputation accuracy. As our experiment shows, SM-MTGP is able to produce more reliable results and significantly outperform other conventional methods.

This paper consists of the following sections. Section 2 describes the application of SM-MTGP algorithm to learn the unobserved missing values in a time-series dataset. Section 3 presents the imputation comparison results of various methods based on a real-world dataset of transfer passenger flow from a particular railway station. Finally, conclusion and discussion to this research are given in Section 4.

**METHODOLOGIES**

In this section, we elaborate how to model a time-series observations of transfer passenger flow over a few months, using SM-MTGP. In general, MTGP improves learning efficiency by investigating the relevance of multiple tasks in parallel. First, the idea behind MTGP is reviewed following, then the application of SM-MTGP algorithm to imputation of missing entries is described.

**Brief review of MTGP**

First, assume that we have \(Q\) tasks and a set of observations \(Y = \{y_{i1}, y_{i2}, ..., y_{iD}\}, i = 1,2, ..., Q\), for each corresponding task at \(D\) various inputs, where \(y_{ij}\) is the response for \(i^{th}\) task given the input \(s_j\). The goal of multi-task learning is to predict unobserved responses (historical missing values or observations may happen in the future), for a certain task using information of all tasks \((23)\). The relations between certain tasks is defined with a covariance function \(K_{lk} = k^f(y_l,y_k)\), where \(k^f\) is a covariance function indicates similarities of features of \(l^{th}\) task and \(k^{th}\) task \((24)\). Moreover, the response is assumed to be generated from an unknown function of input \(s\), which follows a multivariate Gaussian distribution (known as a Gaussian process). The covariance of
input $s_i$ and $s_j$ is defined as $k^s(s_i, s_j) \ (24,25)$. By definition, a Gaussian process is a process of multivariate random variables follows multivariate normal distribution, which is fully characterized by its mean function $m(x)$ and its covariance function $k(x,x')$, detailed interpretation can be found in (25).

**FIGURE 1** Vectorization of matrix $Y$

**SM-MTGP application for missing data imputation**

When the MTGP model is introduced to the imputation of missing values of transfer passenger flow, the shared information of tasks is considered in terms of the temporal relatedness of various days. Transfer passenger flow over $Q$ days can be treated as $Q$ tasks, and the number of sampling time intervals $D$ per day represents $D$ distinct inputs. We define a matrix $Y = \{y_{ij}\}$ ($i = 1, 2, ..., Q; j = 1, 2, ..., D$), where $y_{ij}$ is number of transfer passengers for the $i^{th}$ day (task) on the $j^{th}$ time interval (input). By stacking the column vectors of $Y \in Q \times D$, a $Q \times D$ dimension vector $y = \text{vec}(Y)$ is obtained (Figure 1). We also define $M$ observed entries in $Y$ by creating a $M$ dimensional vector $y_{obs} = P y \in \mathbb{R}^M$ in which $P \in \{0,1\}^{M \times (Q \times D)}$ is a index matrix used for removing unobserved elements in $Y$.

The observed passenger flows of each grid of matrix $Y$ are standardized as $\tilde{Y}$ given Equation (1), where $\mu$ and $sd$ are the mean and standard deviation of the observed passenger numbers, respectively.

$$\tilde{y}_{ij} = \frac{y_{ij} - \mu}{sd} \quad (1)$$

The MTGP model of $\tilde{Y}$ can be described as Equation (2), in which $m_{ij}$ is the expected value of the element $\tilde{y}_{ij}$, and $\varepsilon$ is an additive Gaussian noise with variance $\sigma^2$.

$$\tilde{y}_{ij} = m_{ij} + \varepsilon, \ \varepsilon \sim N(0, \sigma^2) \quad (2)$$

As pointed out by Yu et al. (26), $m_{ij}$ follows a tensor Gaussian process, as described in Equation (3). $\Sigma_Q$ and $\Sigma_D$ are covariance matrices that represent similarities over $Q$ days and $D$ time intervals, and $\otimes$ denotes the Kronecker product. By applying SM-MTGP, the covariance matrices
\[ \Sigma_Q \] are defined as a product of kernel of days (tasks) features \( K_Q^f \) and the self-measuring kernel \( G_Q^m \), and \( \Sigma_D \) are defined as a product of the kernel of time intervals (inputs) features \( K_D^f \) and self-measuring kernel \( G_D^m \), as shown in Equation (4). The calculation of self-measuring kernel requires a complete dataset, therefore, the missing entries in \( Y \) are initially imputed by column mean of weekdays and weekends. That is, the missing values at a particular time interval on weekdays are filled by mean of all the observed elements of that time interval on weekdays, and the missing entries on weekends are imputed in the same manner.

\[
m \sim N(0, \Sigma_Q \otimes \Sigma_D) \quad (3)
\]

\[
\Sigma_Q = K_Q^f G_Q^m, \quad \Sigma_D = K_D^f G_D^m \quad (4)
\]

Let us define

\[
K_Q^f = k(y_i, y_j) \in \mathbb{R}^{Q \times Q}, \quad G_Q^m = g(y_i, y_j) \in \mathbb{R}^{Q \times Q}, \quad (5)
\]

\[
K_D^f = k(y_h, y_l) \in \mathbb{R}^{D \times D}, \quad G_D^m = g(y_h, y_l) \in \mathbb{R}^{D \times D} \quad (6)
\]

where \( k(y_i, y_j) \) and \( k(y_h, y_l) \) indicate covariances of features of \( i^{th} \) day and \( j^{th} \) day, and covariances of features of \( h^{th} \) time interval and \( l^{th} \) time interval, respectively. Similarly, \( g(y_i, y_j) \) and \( g(y_h, y_l) \) measure covariances of self-measuring observations of \( i^{th} \) day and \( j^{th} \) day, and covariances of self-measuring observations of \( h^{th} \) time interval and \( l^{th} \) time interval, see in Equations (5) and (6). Regarding self-measuring observations, the real number of transfer passenger flows are used when the values are observed, and the initially imputed weekday or weekend mean of passenger flows are used when the values are missing. By following the principle of MTGP, the joint distribution of \( \tilde{Y} \) can be described as Equation (7).

\[
\int p(\tilde{Y}|M, 0, \sigma^2) \, p(M|\Sigma_Q, \Sigma_D) \, dM = N(\tilde{Y}|\mathbf{0}, \Phi) \quad (7)
\]

where \( \Phi = \Sigma_Q \otimes \Sigma_D + \sigma^2 \mathbf{I} \). Using a Gaussian process framework given the observed number of transfer passengers, the unobserved passenger flows in \( Y \) can be derived by predictive Equation (8).

\[
E[\tilde{y}_{ab}|\tilde{y}_{obs}, \Sigma_Q, \Sigma_D] = (\Sigma_{Qa} \otimes \mathbf{K}_{Db})_{obs}^T \, \Phi_{obs}^{-1} \tilde{y}_{obs} \quad (8)
\]

where \( \Phi_{obs} = \mathbf{P} \Phi \mathbf{P}^T \in \mathbb{R}^{M \times M} \) is a covariance matrix over the observed transfer passenger flows in \( Y \). \( \Sigma_{Qa} \) denotes \( a^{th} \) column vector in \( \Sigma_Q \), which measures the similarities between \( a^{th} \) day and all the other days among \( Q \) days, and \( \mathbf{K}_{Db} \) indicates \( b^{th} \) column vector of \( 
\]

EXPERIMENTS

In this section, we test the performance of the proposed SM-MTGP method for imputing missing values of transfer passenger flow for a large-demand railway station in Melbourne. First, we introduce the utilized dataset in the experiments. Then, the comparison results of various
imputation techniques under several scenarios covering different missing data patterns are reported, and the possible reasons for the imputation differences are also given.

**Dataset**
The data analysed in this paper includes around 6-months of passenger flow data (March 8, 2017 to September 15, 2017), which were collected by WiFi sensors at Richmond railway station, Melbourne, Australia. Richmond station is located near the Melbourne CBD and Melbourne's sporting precinct. It is a connection of all of Melbourne's eastern and southeastern rail lines (8 lines) (Figure 2). It also serves for Melbourne's sports events. It has 10 platforms. The deployed 12 WiFi sensors are distributed at four platforms 7-8, platforms 9-10 and two sided underpasses (Figure 3). In our experiment, we focus on transfer passengers between platforms 1-6 to platforms 7-8 through the left underpass to get more insights into transfer passenger flow of Richmond station.

---

**FIGURE 2** Map of location and train lines of Richmond station (map source from Google map and Public Transport Victoria)
FIGURE 3  Map of 12 WiFi sensors distribution (map source from Google map).

Data collected by WiFi sensors contains a certain amount of extraneous data, including devices of non-passengers (drivers or pedestrians outside the station), pass-through passengers, devices of the station staff, and nonmobile devices (1-3). Therefore, several filtering algorithms have been applied to remove these noisy readings. Another challenge of WiFi data is that passengers may not carry a detectable smart device or some may carry more than one. Besides, passengers who carry smart devices may turn off their WiFi function. To overcome these issues, we refer to the results of a survey conducted by Faculty of Arts at Monash University, which show that 98% passengers carry mobile devices and around 70% switch on the WiFi function of their devices. The WiFi data is aggregated to 60 minutes in our study. Given proper preprocessing of the data, thus, we got a two-way array data set of 192 rows (days) and 24 columns (24 time intervals per day) with 1507 missing entries. Some whole days are missing in our initial dataset, due to the connection failure and some sensor issues.

Comparison of imputation performance

With the purpose of evaluating the imputation performance of SM-MTGP, we compared it with several other frequently-used imputation methods for different ratios. For a given missing ratio, different missing scenarios were implemented, which we defined as discrete missing pattern and mixed missing pattern. For discrete missing patterns, the artificially removed data were generated randomly from the initial dataset with an assumed missing ratio. Moreover, in order to evaluate the performance of of SM-MTGP with whole day missing, the mixed missing patterns were considered, in which one missing day, two and four random missing days were tested in the same missing ratio (Figure 4). The experiments were conducted with the following missing ratios: 10%, 25%, and 50%. The missing ratio in our experiment is defined as the number of artificially removed data divided by the number of observed entries, corresponding to overall missing rates (the number of missing points divided by the number of total points) of 39%, 50%, and 66%, respectively. To mitigate the randomness impact of the sampling, for each scenario of a given missing ratio, we generated 10 different samples of missing data, imputed all these 10 samples and finally calculated average impute errors.
The Root Mean Square Error (RMSE) given in Equation (9) is used as the evaluation criteria, which measures the error between the real values $y_{\text{real}}$ and the imputed values $y_{\text{imp}}$. To test the performance of various imputation techniques, we used the basic column mean imputation method as our base model, compared with methods as following: (a) kNN method, in which the median values of the k nearest neighbors is used, the distance is measured by the contribution of variable (27). (b) SVD-impute method, in which the missing values in the initial data set has been substituted by column average as SVD requires a complete matrix (5). (c) Kalman filter method, in which time series arima model is used as state space model (28). (d) SM-MTGP method, in which the RBF kernel $k(x,x') = g(x,x') = \exp(-\lambda \|x - x'\|^2)$ is utilised as covariance function, the values of parameters were determined through N-fold cross validation, we performed $n = 4$, 4, and 3 for 10%, 25% and 50% missing rates through Bayesian Optimization algorithm using GPyOpt (29).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{\text{real}} - y_{\text{imp}})^2} \quad (9)$$

The RMSE results of discrete missing data for different missing ratios with various imputation algorithms are obtained (Table 1). The results are shown as RMSE and percentage of improvement compared with the base model. As expected, the RMSEs of all the experimental methods increase when the percentage of the missing data increases. The base models (mean-based imputation) produce relative stable results along different missing ratios, due to the average process using all the measurements. SVD-impute outperforms kNN and Kalman filter under small missing ratio, reflecting the benefits of mean initialization as it provides more information of the whole dataset. However, it is not useful when missing ratio increases to 50%, this may because initialization process is risky with introducing biases under large missing ratio. It is noticed that the improvements in RMSE by SM-MTGP is 64%, 60% and 56% for all these three different missing rates, respectively. SM-MTGP methods significantly outperform other imputation methods. One possible explanation is that SM-MTGP incorporates global information from features of sampling days and time intervals and information from responses. Additionally, the relatedness among various days and time intervals are learned nonlinearly using kernel function which also help obtain improvements of imputation performance.

**TABLE 1** RMSE Results for Different Missing Ratios of Discrete missing data

<table>
<thead>
<tr>
<th>Missing data</th>
<th>Mean</th>
<th>kNN</th>
<th>SVD-impute</th>
<th>Kalman filter</th>
<th>SM-MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>9.892</td>
<td>7.754</td>
<td>5.636</td>
<td>8.007</td>
<td>3.587</td>
</tr>
<tr>
<td></td>
<td>(22%)</td>
<td>(43%)</td>
<td>(19%)</td>
<td>(64%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(15%)</td>
<td>(29%)</td>
<td>(8%)</td>
<td>(60%)</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>10.118</td>
<td>10.230</td>
<td>Not work</td>
<td>9.503</td>
<td>4.453</td>
</tr>
<tr>
<td></td>
<td>(-1%)</td>
<td>---</td>
<td>(6%)</td>
<td>(56%)</td>
<td></td>
</tr>
</tbody>
</table>

The results of three mixed missing patterns (one random day, two and four random days missing) under different missing ratios are also reported (Tables 2, 3, and 4). Similarly, the imputation errors show a gradual increasing trend as the missing ratio increases. Furthermore, we
can observe that, the imputation performance shows a decreasing trend as the number of whole missing day increase. In addition, compared with discrete missing patterns, performance decays in mixed missing patterns with the same missing ratio. This is mainly because empty entries are distributed uniformly under discrete missing ratios, allowing for more local information to recover the missing data. From the results, the SM-MTGP method is still able to obtain better performance compared with all the other methods, leading to improvements in RMSE around 60%. We can see that, the SM-MTGP can produce more efficient results under all these experimental mixed missing scenarios.

### TABLE 2 RMSE Results for Different Missing Ratios of Mixed Missing Data with One Random Day Missing

<table>
<thead>
<tr>
<th>Missing data</th>
<th>Mean</th>
<th>kNN</th>
<th>SVD-impute</th>
<th>Kalman filter</th>
<th>SM-MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>10.081</td>
<td>7.873</td>
<td>6.503</td>
<td>9.344</td>
<td>3.765</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(22%)</td>
<td>(35%)</td>
<td>(7%)</td>
<td>(63%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9%)</td>
<td>(30%)</td>
<td>(9%)</td>
<td>(56%)</td>
</tr>
<tr>
<td>50%</td>
<td>10.281</td>
<td>10.401</td>
<td>Not work</td>
<td>9.534</td>
<td>4.401</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-1%)</td>
<td>(7%)</td>
<td>(57%)</td>
</tr>
</tbody>
</table>

### TABLE 3 RMSE Results for Different Missing Ratios of Mixed Missing Data with Two Random Days Missing

<table>
<thead>
<tr>
<th>Missing data</th>
<th>Mean</th>
<th>kNN</th>
<th>SVD-impute</th>
<th>Kalman filter</th>
<th>SM-MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>9.971</td>
<td>8.185</td>
<td>6.469</td>
<td>8.819</td>
<td>3.815</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(18%)</td>
<td>(35%)</td>
<td>(12%)</td>
<td>(62%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9%)</td>
<td>(35%)</td>
<td>(7%)</td>
<td>(60%)</td>
</tr>
<tr>
<td>50%</td>
<td>10.029</td>
<td>10.140</td>
<td>Not work</td>
<td>9.458</td>
<td>4.201</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-1%)</td>
<td>(6%)</td>
<td>(58%)</td>
</tr>
</tbody>
</table>

### TABLE 4 RMSE Results for Different Missing Ratios of Mixed Missing Data with Four Random Days Missing

<table>
<thead>
<tr>
<th>Missing data</th>
<th>Mean</th>
<th>kNN</th>
<th>SVD-impute</th>
<th>Kalman filter</th>
<th>SM-MTGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>10.379</td>
<td>8.839</td>
<td>7.721</td>
<td>8.974</td>
<td>4.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(15%)</td>
<td>(26%)</td>
<td>(14%)</td>
<td>(61%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(12%)</td>
<td>(29%)</td>
<td>(9%)</td>
<td>(58%)</td>
</tr>
<tr>
<td>50%</td>
<td>10.215</td>
<td>10.218</td>
<td>Not work</td>
<td>9.642</td>
<td>4.592</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0%)</td>
<td>---</td>
<td>(6%)</td>
<td>(55%)</td>
</tr>
</tbody>
</table>

### CONCLUSION AND DISCUSSION

This paper proposed the use of a self-measuring multi-task Gaussian process (SM-MTGP) strategy for imputing missing values of transfer passenger flow at a railway station. SM-MTGP is able to capture and exploit more information from the sampling days (tasks), time intervals (inputs) and responses, which lead to high imputation accuracy. Specifically, the proposed method is advantageous in: 1) leveraging global information (features) from all tasks and all inputs based on a two-way array jointly, 2) utilizing kernel functions to capture nonlinear patterns in the
observations; 3) capturing correlations between responses to improve imputation results. The
methods based on a real-world passenger flow dataset. As the distribution of missing data can vary
substantially under different contexts, therefore, experiments under different missing scenarios
were performed in our study, including discrete missing patterns and mixed missing patterns with
several assumed missing ratios. Our results show that: (a) in general, the SM-MTGP leads to better
results compared to other experimental methods; (b) the imputation accuracy can achieve around
60% improvement in RMSE in all the tested missing scenarios compared with the base model; (c)
the SM-MTGP significantly outperforms other methods under the large missing ratio. The
outcome of this study could have broad application in missing data imputation in both road traffic
and public transport for subsequent intelligent transportation system analysis.

In this study, the RBF kernel was selected as the covariance function, but there might be
more suitable kernel functions that can improve imputation accuracy. Future studies should be
conducted for the investigation of the following aspects: 1) the choice of kernel functions, as they
measure the similarities of different aspects and could affect the imputation results and 2)
consideration of other features such as weather condition into the algorithm.

ACKNOWLEDGEMENTS
This work has been supported by the Rail Manufacturing Cooperative Research Centre (funded
jointly by participating rail organisations and the Australian Federal Government’s Business
Cooperative Research Centres Program) through Project R3.7.11 – Short-term rail passenger flow
forecasting application. The authors would like to thank the Institute of Railway Technology at
Monash University, YPB Systems Ltd and Metro Trains Melbourne, for the provision of WiFi
data. Suvey report provided by Prof. Brett Hutchins from Faulty of Arts, Monash University was
also greatly appreciated.

The authors confirm contribution to the paper as follows: study conception and design: Wenhua
Jiang, Nan Zheng, Inhi Kim; data collection: Paul Reichl; analysis and interpretation of results:
Wenhua Jiang, Nan Zheng, Inhi Kim, Paul Reichl; draft manuscript preparation: Wenhua Jiang.
All authors reviewed the results and approved the final version of the manuscript.

REFERENCES
   Destination Information from Bluetooth and Wi-Fi Sensing for Transit.
   Transportation Research Record: Journal of the Transportation Research Board,
   No. 2595, 2016, pp. 11–17.
2. Hidayat, A., S. Terabe, and H. Yaginuma. WiFi Scanner Technologies for
   Obtaining Travel Data about Circulator Bus Passengers: Case Study in Obuse,
   Nagano Prefecture, Japan. Transportation Research Record: Journal of the
   Transportation Research Board, 2018. 0361198118776153
   with WiFi and Bluetooth Sensors. Transportation Research Record: Journal of the
   Transportation Research Board, No. 2644, 2017, pp. 72–82.


