Assessing and Comparing Data Imputation Techniques for Item Nonresponse in Household Travel Surveys

Alikasim Budhwani¹, Tina Lin², Devin Feng², and Chris Bachmann²

Abstract
This research provides a comparative assessment of data imputation techniques for item nonresponse in household travel surveys. Using the Transportation Tomorrow Survey (TTS) data for the Region of Waterloo in Ontario, Canada, a series of synthetic datasets are generated with varying amounts of missing data, while preserving the respective proportions of missing items and missing item combinations in the original survey data. Then, the performances of six different imputation techniques are compared. The six different imputation techniques include two simple imputation techniques (mode and hot-deck), three discriminative models (logistic regression, multi-layered perceptron, support vector machines) and one generative model (autoencoder). This assessment compares these techniques, as well as the impact of the proportion of item nonresponse in the dataset through their repeated application to multiple synthetic datasets. Results show that the machine/deep learning techniques (both generative and discriminative) not previously applied to household travel survey data outperform their simple imputation counterparts. Overall, the accuracy of travel household survey data imputation is shown to depend on many factors, including the technique employed, the dimensionality of the missing item, and the hypertuning of the technique (if applicable), but not on the amount of missing data in these experiments. This research should prove beneficial to practitioners who often confront item nonresponse in their household travel survey data by providing evidence and recommendations to support the selection and implementation of a data imputation technique. The research methodology also provides a repeatable procedure for future researchers to test data imputation techniques on their own datasets.

Keywords
planning and analysis, household travel surveys

Household travel surveys provide a “snapshot” of data, collected at periodic but infrequent intervals (1). For example, the Transportation Tomorrow Survey (TTS) is a survey on how Ontarians in the Greater Toronto and Hamilton Area (GTHA) and Greater Golden Horseshoe (GGH) travel. The survey data have been collected every 5 years (2016, 2011, 2006, 2001, 1996, 1991, and 1986) (2). Less than ideal datasets are often collected as a compromise between serving the survey’s purposes and maximizing the sample size (e.g., reducing respondent burden) within the cost and time constraints (3).

Many common problems with household travel surveys stem from their small sample sizes. For modeling purposes, the total trips generated may be underrepresented by the expanded survey data (1). In relation to trip distribution, the sample sizes of household travel surveys are often not large enough to analyze results at the zonal level (1, 4). Also, for mode choice modeling, the number of observations for transit modes is likely to be too small to estimate statistically significant model parameters, except in areas with high transit use (or very large survey sample sizes); the same is true for mode

¹Department of Mechanical Engineering, University of Waterloo, Waterloo, Ontario, Canada
²Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada

Corresponding Author:
Chris Bachmann, chris.bachmann@uwaterloo.ca
shares for relatively small segments of the population (such as members of zero-vehicle, high income households) (7, 5). Even when used solely for information purposes, the small sample sizes require expansion to match regional distributions based on Census data for certain characteristics, such as household size, income group, and vehicle ownership.

Despite mitigating efforts, one of the primary reasons for small sample sizes is nonresponse. Two types of nonresponse are unit nonresponse and item nonresponse. Unit nonresponse refers to the failure to obtain questionnaires or data collection forms (such as travel diaries) for a member of the sample. Item nonresponse—the focus of this research—refers to the failure to obtain a specific item of information from a responding member of the sample. For example, the income response is usually relatively low in household travel surveys (6). Item nonresponse is often used interchangeably with the term “missing data.” Unless adjustments are made to the data, the level of nonresponse bias will depend on two factors: the proportion of the sample for whom data were not obtained, and how much the respondents differ from the nonrespondents (7). Weighting adjustment and imputation are the principal techniques for dealing with survey unit and item nonresponse, respectively. However, the transportation field has done relatively little work on the assessment of data imputation techniques for item nonresponse, and the implications of choosing alternative techniques for a given item and level of missing data are not well understood.

This research has three primary objectives. The first objective is to determine the methods of data imputation feasible for household travel surveys. While traditional statistical data imputation techniques such as hot-deck imputation, regression-based imputation, and cell mean imputation are well known to the transportation community, techniques recently developed and applied in other fields may also be applicable to household travel survey data. The second objective is to develop synthetic household travel survey datasets with variable proportions of missing data that maintain the underlying correlations of item nonresponse in an observed household travel survey. Such datasets are necessary for testing data imputation techniques, so the imputed items can be compared with their amputated values to determine the relative performance of the various techniques. The third objective is to assess the performance of the feasible data imputation techniques on the synthetically generated datasets. This assessment compares techniques, as well as the impact of the proportion of item nonresponse in the dataset through the repeated application to multiple synthetic datasets.

The remainder of this paper is organized as follows. The next section provides a literature review of data imputation techniques and the transportation applications of data imputation techniques. The household travel survey data and the data amputation method developed for generating synthetic datasets are described next, followed by the data imputation techniques to be assessed on these synthetic datasets. The results are then presented and discussed in the context of their relative performances across datasets with varying proportions of item nonresponse, before the study limitations are highlighted. The final section outlines the main findings and contributions of this research before providing a brief agenda for future research.

**Literature Review**

*Data Imputation Techniques*

When dealing with the unit nonresponse issue of travel household surveys, early work typically utilized a weight-class method as a correction technique (8). Other imputation methods for household travel surveys have included deductive imputation, mean imputation, hot-deck imputation, and regression imputation (7, 9). Deductive imputation imputes data through logical conclusions (e.g., trip distance from an origin and destination), but suffers limitation when logical conclusions cannot be made. Mean imputation replaces values with the mean values of observed cases, which can also suffer from understated variance estimates and invalid confidence intervals (9). While deductive imputation and mean imputation are typically more intuitive techniques, they are rarely used because of their limitations. While hot-deck and regression imputation have been applied to household travel surveys, the other techniques studied in this research are inspired by similar datasets from the healthcare field (6, 10). Clinical information in the healthcare field commonly contains missing values and incomplete data, which are addressed with hot-deck imputation, multilayered perceptron (MLP), and autoencoders (11, 12). Additionally, support vector machines (SVM) have also been used to impute data for activity-based diaries, which are nearly identical datasets, and therefore suffer similar challenges because of nonresponses (13).

*Other Transportation Applications of Data Imputation*

Since the application of traffic management often requires accurate and complete traffic datasets, most data imputation techniques have focused on the estimation of incomplete traffic data. The information lost reflects the shortcomings of real-world limitations, such as with malfunctioning devices, transmission errors, weather effects, and so forth. (14, 15). The traditional techniques used to impute traffic data are categorized into three categories: prediction, interpolation, and
statistical learning (15). Prediction models include autoregressive integrated moving average, Bayesian networks (BN), neural networks, and support vector regression (16–21). Interpolation techniques include temporal- and spatial-neighboring (22). Statistical learning techniques are typically principal component analysis based (23, 24). More recently, imputation methods consider spatio-temporal correlation leading to a tensor-based approach to deal with missing traffic data (24–27).

Some extensions of data imputation, such as data inference or data derivation, in transportation applications have arisen as global positioning system (GPS)-based surveys have gained popularity in mobility data collection. The limitation of GPS-based survey data often lies with the inability to detect mode and trip purpose (28–30). Similar limitations of trip purpose also exist in smart-card data (31). Common approaches for mode derivation algorithms consist of rule-based, probability-based, and machine learning methods (32,33). Early research focused on rule-based approaches which are easily interpretable but have low transferability and are unsuitable for large datasets (34–38). Probability-based methods are often more flexible with some transferability and more suitable for larger data sets (28,39–42). Machine learning methods represent several types of algorithm such as neural networks, BN, decision trees, SVM, and autoencoders, which have higher transferability than rule-based or probability-based methods and are suitable for very large datasets (29, 30, 43, 44).

Newer methods of imputation have also been proposed to deal with specific issues: missing data and endogeneity in discrete choice models, immeasurability and discontinuity in elements of collision data, and crowdsourced data that are highly prone to missing observations (45–48).

**Methods**

**Data**

Travel diary data for the Region of Waterloo (the “Region”) were retrieved from the TTS, which most recently surveyed 4.8% of the Region’s households in October 2016 (49). The TTS characterizes each sample with trip, transit (if primary mode), person, and household attributes. Each of these attributes, or variables, are referred to as “features” in this paper. Each feature is discretized by dimensions that refer to the number of unique values or categories.

Table 1 details the features in the raw dataset, their dimensions, the quantity of dimensions (bins), and the analytical unit at which the features are represented. Percentile-based bin sizes were used to discretize continuous data, so that the dimensions of the features have equal sample sizes. Some classification models benefit from the discretization of continuous variables because equal-sized bins reduce training bias favoring frequently observed categories. Additionally, homogeneity is important for most models. In the case of hierarchical splitting

| Features (label)                  | Dimensions                          | Bins | Unit |
|-----------------------------------|-------------------------------------|------|------|
| Sociodemographic                  |                                     |      |      |
| Age (age)                         | (11–19]; (19–28]; (28–35]; (35–41]; (41–47]; (47–53]; (53–58]; (58–65]; (65–72]; (72–98] | 10   | PER  |
| Sex (sex)                         | F: female; M: male                  | 2    | PER  |
| Income (hh_income)                | <15k; 15k–39k; 40k–59k; 60k–99k; 100k–124k; >124k | 6    | HH   |
| Employment (emp_stat)             | Employed; Not_employed; Work_at_Home| 3    | PER  |
| Student (stu_sat)                 | Student; Not_student                | 2    | PER  |
| Occupation (occupation)           | Retail&Service; General_Office; Manufacturing; Not_employed | 4    | PER  |
| Transportation                    |                                     |      |      |
| Licensed (driver_lic)             | Y: yes; N: no                      | 2    | PER  |
| Vehicle ownership (hh_n_vehs)     | 0; 1; 2: more than 2               | 4    | HH   |
| Transit pass (trans_pass)         | Y: yes; N: no; Other_agency (non-GRT pass) | 3    | PER  |
| Free parking at work (free_park)  | Y: yes; N: no; NA: not applicable   | 3    | PER  |
| Household                         |                                     |      |      |
| Household size (hh_size)          | (0.9–1]; (1–2]; (2–3]; (3–4]; (4–9] | 5    | HH   |
| Household employees (hh_n_emp)    | (0–0.01]; [0.01]; [0.1]; [1–2]; (2–3]; (3–4]; (4–11] | 6    | HH   |
| Dwelling type (hh_dwell_type)     | House; Townhouse; Apartment         | 3    | HH   |
| Trip-related                      |                                     |      |      |
| Trip purpose (trip_purp)          | Home-based work; home-based discretionary; non-home-based | 3    | TRP  |
| Trip time (start_time)            | Peak: [6:00–9:35]; [15:00–17:00]; Off_Peak | 2    | TRP  |
| Primary mode (mode_prime)         | Transit; non-transit                | 2    | TRP  |
| Network trip distance (trip_km)   | (0–1]; (1–2]; (2–3]; (3–4]; (4–5]; (5–6]; (6–8]; (8–11]; (11–16]; (16–56] | 10   | TRP  |
| Daily trip count (n_pers_trip)    | [1–2]; [2–3]; [3–4]; [4–6]; [6–18] | 5    | PER  |

Note: PER = Person; HH = Household; TRP = Trip; GRT = Grand River Transit.
models, bins may better describe trends within feature dimensions because there is sufficient data to indicate trends in subsequent splits. Continuous variables that were discretized include age, household size, number of vehicles in household, and travel distance. The number of bins for each continuous feature were decided pragmatically on the basis of relevance in the transportation domain (e.g., peak versus off-peak trip departure times) and socioeconomic interest (e.g., age deciles).

Figure 1 illustrates the observed missingness of the raw dataset, which contains a total of 51,104 trips conducted by households within the Region of Waterloo. The total raw dataset consists of 11,502 trips with at least one missing feature. Of the 18 features outlined in Table 1, the following seven features have incomplete observations: trip_km, age, driver_lic, trans_pass, emp_stat, occupation, and free_park. Of the missing features, hh_income and free_park have the highest proportions of incomplete observations totaling 16.1% and 7.7%, respectively.

### Data Amputation for Simulation

A data amputation procedure is used to generate missing values in the processed dataset so that the performance of techniques can be evaluated. The data amputation aims to accurately reproduce the missing combinations, as summarized in Table 2, with the respective proportions to mimic the original dataset. This procedure allows the data imputation techniques to be tested on a representative dataset.

For this study, all records with missing items in the original survey dataset were deleted. The resulting dataset with fully known observations consists of 42,549 trips amongst 9,790 households. The amputation procedure starts at the household dataset and let this dataset with no missing values be called $X^O$. Missing data (of type Missing at Random) with the same missing proportion and missing combinations for household features as the original dataset (Table 2) were generated in $X^O$ using a multivariate amputation procedure (30). This new household dataset is then merged with the personal dataset and

---

**Table 2. Missing Combinations of Raw Data (by Frequency)**

| Missing combinations | Count |
|----------------------|-------|
| ['hh_income']        | 7,348 |
| ['free_park']        | 3,189 |
| ['free_park', 'hh_income'] | 673 |
| ['occupation']       | 93   |
| ['age']              | 31   |
| ['age', 'hh_income'] | 28   |
| ['tran_pass', 'free_park'] | 26  |
| ['occupation', 'free_park'] | 18  |
| ['driver_lic']       | 15   |
| ['emp_stat']         | 14   |
| ['emp_stat', 'hh_income'] | 6   |
| ['age', 'free_park'] | 5    |
| ['tran_pass', 'hh_income'] | 4   |
| ['driver_lic', 'tran_pass', 'free_park'] | 4  |
| ['occupation', 'free_park', 'hh_income'] | 4  |
| ['driver_lic', 'free_park'] | 4  |
| ['tran_pass', 'occupation', 'hh_income'] | 3  |
| ['tran_pass', 'occupation', 'free_park', 'hh_income'] | 3  |
| ['age', 'free_park', 'hh_income'] | 3  |
| ['age', 'occupation', 'hh_income'] | 2  |
| ['age', 'occupation', 'free_park', 'hh_income'] | 2  |
| ['driver_lic', 'tran_pass', 'free_park', 'hh_income'] | 2  |
| ['driver_lic', 'occupation'] | 2  |
| ['trip_km']          | 2    |
| ['n_vehicle']        | 2    |
| ['trip_km', 'hh_income'] | 1  |
trip dataset and the amputation procedure is repeated with personal and trip features. The merged dataset with synthetically generated missing data is denoted by $X$ and is used by the imputation techniques to test their performances. Multiple datasets with synthetically generated missing values ranging from 5% missingness to 50% missingness were created using the Pymice library. A sample of a simulated data set for the 50% missingness scenario is illustrated in Figure 2.

Data Imputation Techniques

Before testing the imputation techniques, all features were dummy encoded for preparation. The techniques utilized in imputing household travel survey data can be broadly classified into three categories: simple models, discriminative models, and generative models.

Multiple imputation—another popular imputation technique—is considered, but ultimately excluded from this study. Multiple imputation is sequential single imputation where several imputed versions of the incomplete data sets are first created by considering multiple candidates for each missing data point. The imputed sets of incomplete data are then evaluated using standard statistical procedures resulting in multiple different outcomes of the statistical analyses. Finally, these results are combined into an overall statistical analysis in which the uncertainty about the missing data is integrated in the standard errors and significance tests. A limiting factor of multiple imputation is the requirement of repeated imputation of the dataset, ultimately leading to a computationally expensive technique. Furthermore, it traditionally assumes a continuous metric for missing data and imputes continuous data, therefore converting categorical values into continuous ones. This assumption can cause problems with binary variables (e.g., free_park) by imputing a value of 0.8, thus creating impractical results. This problem would increase as the amount of missingness in the dataset increases. The fundamental concern of multiple imputation is that it fails to understand the meaning of the measure it is imputing where, even if no bias is present, it degrades the meaning of the considered variables (51).

Simple Imputation. The simple imputation methods used in this research are mode imputation and hot-deck imputation. Mode and hot-decking are simple imputation methods in which the missing values are filled by standard measure or constant. These methods do not require any training, since they instantly fill in the missing values based on simplistic calculations. Therefore, the dataset is not required to be dummy encoded for these methods. Both the methods are implemented using the scikit-learn library in Python 3.

Mode. Mode imputation replaces the missing values with the most frequent value in the column (52). It is the mean equivalent for the categorical variables. This method is a relatively common alternative to listwise deleting the missing data in the dataset. In this work, this approach is considered as the baseline method.

Hot-Deck Imputation. Hot-deck imputation is another simple imputation technique for filling missing data. It is especially useful for imputing missing data in large datasets (53). In this method, a missing item or value from a record (recipient) is replaced by a value from another similar case (donor) that has complete data.

For this study, the nearest neighbor (NN) hot-deck ($k$-NN) imputation is used, where a record with a missing value (recipient) is assigned the value of the NN case.
(donor) based on a distance metric. This imputation was conducted using the KNNImpute module from the scikit-learn library. This module carries out matrix completion of the data frame by choosing the mean values of the $k$ closest samples for features where both samples are present. The distance metric used for this study was nan-Euclidean distance and only one neighbor was chosen for imputation.

Consider that a household survey case is represented by an $n$-dimensional input vector, $x = [x_1, x_2, \ldots, x_n]^T$. Furthermore, the number of missing values in $x$ is represented as a binary vector, $m$, where $m_j = 1$ if the feature is not missing, and 0 otherwise. Given two cases, $x_a$ and $x_b$, weight is defined as: $w = \frac{q}{k}$ where $k$ is the max number of missing (nan) values present in the two cases.

The nan-Euclidean distance between $x_a$ and $x_b$ is:

$$d(x_a, x_b) = \sqrt{w \sum_{j=1}^{n} d_j(x_a, x_b)^2}$$

where $d_j(x_a, x_b) = \text{distance between } x_a \text{ and } x_b \text{ on its } j^{th} \text{ attribute}$, and $w = \text{weight}$.

This distance can either be 0 or 1 as all the features are categorical and dummy encoded. The weight term, $w$, is added to the equation to give more weight to cases with fewer missing values in them.

**Discriminative Models.** Discriminative models are models used for classification and regression. They study the probability of a class label, $y$, given observed variable, $x$, denoted as $P(y|x)$ (54). For using these models as imputation techniques, an imputation scheme has been developed in this study as follows:

1. For a simulated incomplete dataset, $X$, separate the cases that do not contain any missing data (this is the complete set, $X^C$) from the set having missing values in them (this is the incomplete set, $X^I$).
2. For each feature in each combination of missing features in $X^I$, train the discriminative model using $X^C$. The dependent/target variables, $y$, are the missing features, and the independent variables are the other remaining features. There is one discriminative model for feature in each missing combination (Table 2).
3. After the models are trained, for each feature in each missing combination in $X^I$, the missing values are predicted using the respectively trained model from the previous step.
4. Overall accuracy of a feature is obtained by averaging the accuracy of all the different models used for that feature.
5. The above steps are conducted for all of the synthetic datasets.

This methodology was implemented with the scikit-learn library in Python 3.

**Logistic Regression.** This model is a generalization of the logistic regression model, which is used to model the relationship between a binary dependent variable and a set of $k$ predictor variables $(x_1, x_2, \ldots, x_k)$. The expression of binary logistic regression is:

$$\log \left( \frac{\text{prob}(E)}{1 - \text{prob}(E)} \right) = b_0 + \sum_{i=1}^{k} b_i x_i$$

where $b_i = \text{unknown coefficient}$, $b_0 = \text{intercept}$, and $\text{prob}(E) = \text{probability of event } E$.

The left-hand side of the equation is also known as the binary logit model.

The simple logistic regression model can be generalized for multi-class dependent variables (multinomial logistic regression). For this case, with $q$ representing all the classes, there will be $q - 1$ logits of the form:

$$\log \left( \frac{\text{prob}(\text{category}_j)}{\text{prob}(\text{category}_q)} \right) = b_j^{(q)} + \sum_{i=1}^{k} b_{ij} x_i, j = 1, \ldots, q - 1$$

This model can be used for imputation by considering the dependent categorical variable with the missing value and all other features as the predictors.

**Support Vector Machines (SVM).** SVM is another discriminative model used for imputation. This method takes a set of input data and predicts to which of one of the two possible classes each given input belongs. It does so by making a hyperplane that separates the data into two categories. The best separation is achieved by the hyperplane that has the largest distance to the nearest training data of any class (known as the margin), since the larger the margin, the lower the generalization error of the classifier (55). At times, the hyperplane will allow misclassifications, to prevent overfitting of the training data. In this case, the nearest training point to the hyperplane will be called a soft margin.

For a given dataset $D$:

$$D = \{(x_i, y_i)|x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}, i = 1 \ldots n \}$$

where $y_i = \text{one of the two classes to which } x_i \text{ belongs}$.

In each case, $x_i$ is a $p$-dimensional vector. The SVM classifies the cases by forming a max-margin hyperplane.
which divides the cases into two separate regions, \( y_i = 1 \) and \( y_i = -1 \).

The optimization problem of SVM is as follows:

\[
\min_{w, \xi, b} \max_{\alpha, \beta} \left( \frac{1}{2} |w|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i \right), \alpha_i, \beta_i \geq 0 \quad (4)
\]

where

\[- \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1 + \xi_i] = \text{the hyperplane formula},\]
\[w = \text{the normal vector to the hyperplane, which needs to be minimized to find hyperplane with the maximum distance from the margin},\]
\[b = \text{the offset of plane, which needs to be minimized to find a good hyperplane with the maximum distance from the margin},\]
\[y_i = \text{which of the two classes the case } x_i \text{ belongs to (1 or -1)},\]
\[\xi (\text{also known as the slack variable}) = \text{degree of misclassification of the case } x_i, \text{ which needs to be minimized to find a good classifier (hyperplane)},\]
\[C = \text{a regulariziation term which determines how much misclassification can be allowed by the hyperplane, and} \]
\[\alpha_i = \text{a non-negative Lagrange multiplier used to prevent the margins from falling on to the hyperplane.}\]

Other terms in the optimization problem are less critical hyperparameters and are left as default values (55).

For multiclass classification of \( n \) classes, \( n - 1 \) SVM iterations are performed. The original optimization problem may be expressed in a finite dimensional space. The hyperplane, however, needs to be obtained by transforming the original space to higher dimensions, which makes the separation easier in that space. To keep the problem less computationally intensive, the SVM is designed to ensure that optimization problem can be calculated easily by the kernel function in the original space. There are two main kernel functions usually used: the polynomial kernel and the radial basis function (RBF) kernel. In this study, the RBF kernel was used because of its generally higher performance (55). This model can be used for imputation by considering the dependent categorical variable (\( n \) classes) with the missing value as the regions divided by the hyperplane (\( y \)) and all other features as the predictors (\( x \)).

**Multi-Layered Perceptron (MLP).** MLP is a neural network that consists of multiple layers of nodes interconnected in a feed-forward fashion. Each node or neuron is connected to the nodes of the next layer. These connections are called weights. The common MLP structure is made up of three layers: input layer, hidden layer, and output layer. The input layer connects the input data to the \( H \) nodes in the hidden layer; the hidden layer is then connected to the neurons in the output layer (56).

For an input vector \( x \), the \( H \) nodes in the hidden layer are of the form:

\[
z_h = f \left( \sum_{j=1}^{H} w_{hj} x_j + w_{h0} \right) \quad (5)
\]

where
\[h = 1, \ldots, H = \text{hidden nodes},\]
\[z_h = \text{hidden node output},\]
\[w_{hj} = \text{input to hidden weights}, \text{ and}\]
\[w_{h0} = \text{input to hidden biases}.\]

The function \( f \) is a nonlinear function which, in most cases, is sigmoidal. These \( z_h \) are then linearly combined and again transformed using an output function \( g \) to produce a vector of outputs \( y \).

\[
y_t = g \left( \sum_{h=1}^{H} w_{th} z_h + w_{t0} \right) \quad (6)
\]

where
\[t = 1, \ldots, T = \text{total number of output nodes},\]
\[w_{th} = \text{hidden output layer weights}, \text{ and}\]
\[w_{t0} = \text{hidden output layer biases}.\]

Since this study solely focuses on categorical variables, a logistic sigmoid was used as the \( g \) function. Combining both layer equations provides the overall network function:

\[
y_t = g \left( \sum_{h=1}^{H} w_{th} f \left( \sum_{j=1}^{n} w_{hj} x_j + w_{h0} \right) + w_{t0} \right)
\]

This study utilizes a two-layered MLP network (two hidden layers) as it is sufficient to model any complex relationships between the inputs and outputs (57, 58). The network weights are found through model training on the Adam optimization method to minimize the cost function. Each hidden layer consists of 30 neurons based on a parameter sweep of hidden layer size (from 1 to 50). MLP networks can be used as an imputation technique by training the model to learn the missing features (used as outputs), using the remaining complete features as inputs (59, 60).

**Generative Model.** Generative models capture the joint probability of \( P(x, y) \) where \( x \) represents data instances and \( y \) represents labels. Generative models aim to model how data is placed in space by creating a learned representation, called latent space, to accurately generate the output data (61). Generative models provide an effective
way to train rich models to resemble a real distribution. For this study, only autoencoders are considered because of the complex modeling of many of the generative models. For using these models as imputation techniques, an imputation scheme has been developed in this study as follows:

1. For a simulated incomplete dataset, $X$, separate the cases that do not contain any missing data (this is the complete set, $X^C$) from the set have missing values in them (this is the incomplete set, $X'$).
2. Create a corrupt set, $X^R$, by generating missing values in a copy of the $X^C$ set. $X^R$ would be used as input while training the autoencoder and $X^C$ is used to check the reconstructed output from autoencoders during training.
3. The $X'$ set will be used as the test set to determine the performance of the model
4. The percentage of missing data generated in set $X^R$ corresponds to the missing data in the simulated dataset $X$.

The methodology is implemented where dataset $X$ is not dummy encoded.

**Autoencoders.** Autoencoders are a type of neural network, having the same fundamental theory, that learn a distributed representation of their input (62). They transform the data into a smaller hidden layer, called the bottleneck layer, and reconstruct the entire dataset using the inputs from the bottleneck layer. By using a hidden layer (bottleneck) smaller than the input layer, the autoencoder learns the most important features present in the data.

For this study, an autoencoder with a categorical hinge cost is created between the reconstructed layer (output) and the input data, as it had the best performance. The weights and biases of the autoencoder are trained only on the complete dataset, $X^R$, which has been corrupted. No prior imputation is needed to train the model. The autoencoders are trained with a 100 training epoch limit (63). If a new minimum cost is not reached in 100 epochs, training stops. The autoencoder was implemented using the Tensorflow and Keras libraries. A parameter sweep is performed to tune select hyperparameters of the autoencoder. In the sweep, autoencoders of one to four hidden layers and each combination of 2, 4, 10 hidden nodes per hidden layer were tested. Autoencoders with four hidden layers made up of 4, 2, 2, and 4 nodes, respectively, were determined to have the best performance.

**Measures of Effectiveness**

Since the amputated dataset can provide a ground truth scenario, an objective metric is utilized for data imputation technique comparison. This study utilizes accuracy, shown in the formula below:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

where

- $tp$ = true positive,
- $tn$ = true negative,
- $fp$ = false positive, and
- $fn$ = false negative.

**Results**

Figures 3 and 4 summarize the accuracies of the six data imputation models across a percentage of missing data varying from 5% to 50% for high dimension and low dimension features, respectively. The baseline method of mode imputation is highlighted in Figures 3 and 4 with a dashed line type. The predictive accuracies of features with high dimensionalities, shown in Figure 3, reveal that autoencoders obtain the best overall averages accuracy of 84.1% (ranging from 33% to 100%). Apart from **emp.stat**, autoencoders perform significantly higher than the other five models. The SVM model typically follows autoencoders in relation to predictive accuracy averaging at 61.7% (ranging from 19% to 100%). The MLP and regression models then follow with average accuracies of 53.8% and 50.5%, respectively (both ranging from 0% to 100%). Lastly, the baseline mode imputation and $k$-NN performs with average accuracies of 46.8% and 48.4%, respectively (both ranging from 0% to 100%).

The predictive accuracies of features with low dimensionalities shown in Figure 4 reveal SVM and MLP models typically obtain the highest overall accuracies, averaging 88.7% (ranging from 82% to 99%) and 87.7% (ranging from 78% to 98%), respectively. Unlike high dimension features, the regression model performs with higher predictive accuracy and typically follows the SVM and MLP models averaging 68.1% accuracy (ranging from 51% to 90%). Unlike features with high dimensionality, the accuracy of autoencoders is inconsistent in performance and has a lower average accuracy of 58.7% (ranging from 18% to 100%). The $k$-NN performs poorly amongst the low dimension features with an accuracy averaging 29.7% (ranging from 5% to 50%). The baseline mode imputation performs at an average accuracy of 23.7% (ranging from 0% to 48%). Amongst the models that have average accuracies of 50% or higher, the autoencoder is the only model that has a higher predictive accuracies with higher dimension features than with lower dimension features.

For both high and low dimension figures, $k$-NN and mode imputation performed with very low accuracies overall with predictions of $<5\%$ accuracy for man
features. It is also shown that the estimation accuracies of most features in all techniques do not show much impact with an increase in the percentage of missing data.

**Discussion**

Amongst the six methods assessed, autoencoders performed the best across varied missing data in datasets designed to be not at random. Furthermore, the performance, on average, for any of the models does not significantly change by the missingness of the dataset in the tested range (5%–50%), corresponding to previous findings from research on imputation in microarrays (64). Because of a low proportion of missingness in many features (e.g., *emp_stat, hh_n_vehs*), the training dataset for many machine learning models is limited. From the small training dataset, ranging the percentage of missing data

![Figure 3. Comparative model performance for high dimension features: (a) income, (b) vehicle ownership, (c) age, (d) employment, and (e) occupation.](image-url)
may not cover critical amounts of data to warrant a change in predictive accuracies.

The results from autoencoders show limitations with low dimension features. A possible explanation might be not having optimal hidden layers and/or optimal neuron combinations, which could affect the model’s ability to optimally learn the latent space of the dataset. Better hypertuning of the hidden layer and neuron parameter may help with the model performance on features with binary classes. Additionally, the potential for autoencoders could be improved with further training, such as more epochs and less bottlenecking. Methods like MLP and SVM would be better suited to deal with features with low dimensions as seen through their performances. The remaining methods (mode, hot-deck, and logistic regression) are not recommended for imputation because of their poor performances on average.

Earlier research imputed one high dimension feature, number of cars, and one low dimension feature, drivers license, by utilizing SVM with inputs such as household composition, income, age, and so forth (13). Consistent with the findings of this study, their results reveal that the SVM performs better with lower dimension features (89% accuracy) than high dimension feature (69% accuracy) (13). When compared with the results of this study, the comparable feature of number of cars would be hh_n_vehs (33%–100% accuracy). Additionally, one previous study imputed an income feature through hot-deck and regression methods, finding better results with hot-deck than regression, which is also consistent with the findings of this study (6).

Although mode imputation is the simplest and easiest form of imputing data, it lowers the variability of the data and undermines the standard deviations and variances (65, 66). While regression models utilize data from other variables used to predict the missing, it overestimates the correlations between predicted variables and its features while also underestimating variances and co-variances (65). Additionally, the hot-deck approach, commonly used in surveys, can bias the imputed results (53,67). The shortcomings of the aforementioned approaches present an opportunity for machine learning models where results are not biased. Consistent with the results of this research, machine learning methods outperform traditional statistical methods in imputing missing data, resulting in better overall accuracy because of their complex and flexible structure that allows for better pattern prediction and correlation detection (11).

This research is a promising step in using machine/deep learning techniques (both generative and discriminative) for missing data imputation for household surveys. All machine learning techniques are computationally expensive, some more so than others. Autoencoders are the fastest of the machine learning techniques to compute while the discriminative models take longer (SVM being the longest) because of the imputation scheme used for them. As the dataset size becomes larger, autoencoder time for training increases linearly; the same is true for the other neural network model, MLP. Models like SVM require computing a distance matrix, which increases exponentially in time.

**Conclusion**

In this study, the performances of six different imputation techniques are compared for household surveys. The six different imputation techniques include two simple imputation techniques (mode and hot-deck), three discriminative models (logistic regression, MLP, SVM) and one generative model (autoencoder). The assessment
and comparison of these techniques resulted in several key findings. First, simple imputation techniques used previously in transportation, such as mode and hot-decking, performed considerably worse than the newer methods compared in this study (e.g., autoencoders, MLP, and SVM). Second, the dimensionality of the item had an impact on the most accurate imputation technique—for example, autoencoders performed very well with high dimensionality items, but less so with low dimensionality items. A related third finding is that the hypertuning of machine/deep learning techniques may require tailoring for specific items or for groups of items with similar dimensionality—of course, this can be done in practice the same way the techniques were assessed in this research (i.e., create a synthetic dataset with amputations). Fourth, imputation is generally more accurate with low dimension items compared with high dimension items. And fifth, the amount of missing data did not have a large influence on the accuracy of the imputation techniques in the ranges tested.

This research should be considered a first exploratory analysis of these methods. Future investigation is needed to confirm whether the performances shown by the machine learning models in this paper are a result of the structure of this specific household travel survey data or are suitable for generalization. Additionally, this study conducted imputation on the trip file, even for household characteristics, and some data imputation techniques such as hot-deck and regression imputation may perform better when applied at the household- and person-level files. Furthermore, because of spatial aggregation biases, spatial location is excluded from the original dataset. Further investigation with detailed spatial analysis to support data is noted as an opportunity for future research. The size of 42,549 cases is relatively large and methods may differ in performance with larger or smaller datasets. Similarly, investigation needs to be done on how the performance of the models might change if the continuous features were not binned and converted into categorical variables (where applicable). Lastly, detailed studies focused on hyperparameter tuning specifically for household travel survey data are warranted to determine if the tuning techniques used in other fields and adopted in this research are optimal for these data.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: A. Budhwani, T. Lin, D. Feng, C. Bachmann; data collection: D. Feng; analysis and interpretation of results: A. Budhwani, T. Lin, C. Bachmann; draft manuscript preparation: A. Budhwani, T. Lin, D. Feng, C. Bachmann. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Devin Feng https://orcid.org/0000-0003-2724-3566

References

1. United States Federal Highway Administration. (ed.). Travel Model Validation and Reasonableness Checking Manual, 2nd ed. https://rosap.ntl.bts.gov/view/dot/55924.
2. Data Management Group. Design and Conduct of the Survey. http://dmg.utoronto.ca/pdf/tts/2016/2016TTS_Conduct.pdf.
3. de Dios Ortúzar, J., and L. G. Willumsen. Modelling Transport. Wiley, Chichester, West Sussex, 1994.
4. Transportation Research Board. Metropolitan Travel Forecasting: Current Practice and Future Direction — Special Report 288. The National Academies Press, Washington, D.C., 2007.
5. Transportation Research Board of Sciences, Engineering, and Medicine. Travel Demand Forecasting: Parameters and Techniques. The National Academies Press, Washington, D.C., 2012.
6. Xu, M., M. Taylor, and A. Page. A Comparison of Two Methods for Imputing Missing Income From Household Travel Survey Data. Doctoral dissertation. Bureau of Transport and Regional Economics, 2002.
7. Zimowski, M., R. Tourangeau, R. Ghadialy, and S. Pedlow. (ed.). Nonresponse in Household Travel Surveys. N. O. R. Center. https://rosap.ntl.bts.gov/view/dot/15587.
8. Roux, S., and J. Armoogum. Calibration Strategies to Correct Nonresponse in a National Travel Survey. Transportation Research Record: Journal of the Transportation Research Board, 2011, 2246: 1–7.
9. Armoogum, J., and J.-L. Madre. Weighting or Imputations? The Example of Nonresponses for Daily Trips in the French NPTS. https://rosap.ntl.bts.gov/view/dot/4716.
10. National Capital Region Transportation Planning Board. 2007/2008 TPB Household Travel Survey. https://www.mwcog.org/file.aspx?A=r2AyJ%2BmVviaHwovp3noNyN38xda%2FXH6uNAWdI%3D.
11. Jerez, J. M., I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martin, and L. Franco. Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem. Artificial Intelligence in Medicine, Vol. 50, No. 2, 2010, pp. 105–115. https://doi.org/https://doi.org/10.1016/j.artmed.2010.05.002.
12. Beaulieu-Jones, B. K., and J. H. Moore. Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. Proc., Pacific Symposium on
13. Yang, B., D. Janssens, D. Ruan, M. Cools, T. Bellemans, and G. Wets. A Data Imputation Method With Support Vector Machines for Activity-Based Transportation Models. In Foundations of Intelligent Systems (Wang, Y., and T. Li, eds.), Springer, Berlin, Heidelberg, 2011, pp. 249–257.

14. Huang, J., B. Mao, Y. Bai, T. Zhang, and C. Miao. An Integrated Fuzzy C-Means Method for Missing Data Imputation Using Taxi GPS Data. Sensors (Switzerland), Vol. 20, No. 7, 2020, p. 1992. https://doi.org/10.3390/s20071992.

15. Li, Y., Z. Li, and L. Li. Missing Traffic Data: Comparison of Imputation Methods. IET Intelligent Transport Systems, Vol. 8, No. 1, 2014, pp. 51–57. https://doi.org/10.1049/iet-its.2013.0052.

16. Ahmed, M. S., and A. R. Cook. Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. Transportation Research Record: Journal of the Transportation Research Board, 1979. 722: 1–9.

17. Lee, S., and D. B. Fambro. Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting. Transportation Research Record: Journal of the Transportation Research Board, 1999. 1678: 179–188.

18. Ghosh, B., B. Basu, and M. O’Mahony. Bayesian Time-Series Model for Short-Term Traffic Flow Forecasting. Journal of Transportation Engineering, Vol. 133, No. 3, 2007, pp. 180–189. https://doi.org/10.1061/(ASCE)0733-947X(2007)133:3(180).

19. Dia, H. An Object-Oriented Neural Network Approach to Short-Term Traffic Forecasting. European Journal of Operational Research, Vol. 131, No. 2, 2001, pp. 253–261. https://doi.org/10.1016/S0377-2217(00)00125-9.

20. Vlahogianni, E. I., M. G. Karlaftis, and J. C. Golas. Optimized and Meta-Optimized Neural Networks for Short-Term Traffic Flow Prediction: A Genetic Approach. Transportation Research Part C: Emerging Technologies, Vol. 13, No. 3, 2005, pp. 211–234. https://doi.org/10.1016/j.trc.2005.04.007.

21. Castro-Neto, M., Y.-S. Jeong, M.-K. Jeong, and L. D. Han. Online-SVR for Short-Term Traffic Flow Prediction Under Typical and Atypical Traffic Conditions. Expert Systems With Applications, Vol. 36, No. 3, 2009, pp. 6164–6173. https://doi.org/10.1016/j.eswa.2008.07.069.

22. Yin, W., P. Murray-Tuite, and H. Rakha. Imputing Erroneous Data of Single-Station Loop Detectors for Nonincident Conditions: Comparison Between Temporal and Spatial Methods. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, Vol. 16, No. 3, 2012, pp. 159–176. https://doi.org/10.1080/15472450 .2012.694788.

23. Qu, L., L. Li, Y. Zhang, and J. Hu. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach. IEEE Transactions on Intelligent Transportation Systems, Vol. 10, No. 3, 2009, pp. 512–522. https://doi.org/10.1109/TITS.2009.2026312.

24. Ran, B., H. Tan, Y. Wu, and P. J. Jin. Tensor Based Missing Traffic Data Completion With Spatial-Temporal Correlation. Physica A: Statistical Mechanics and its Applications, Vol. 446, 2016, pp. 54–63. https://doi.org/10.1016/j.physa.2015.09.010.

25. Tan, H., G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li. A Tensor-Based Method for Missing Traffic Data Completion. Transportation Research Part C: Emerging Technologies, Vol. 28, 2013, pp. 15–27. https://doi.org/10.1016/j.trc.2012.12.007.

26. Chen, X., Z. He, and L. Sun. A Bayesian Tensor Decomposition Approach for Spatiotemporal Traffic Data Imputation. Transportation Research Part C: Emerging Technologies, Vol. 98, 2019, pp. 73–84. https://doi.org/10.1016/j.trc.2018.11.003.

27. Asif, M. T., N. Mitrović, L. Garg, J. Dauwels, and P. Jaillet. Low-Dimensional Models for Missing Data Imputation in Road Networks. Proc., 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, IEEE, New York, 2013, pp. 3527–3531.

28. Broach, J., J. Dill, and N. W. McNeil. Travel Mode Imputation Using GPS and Accelerometer Data From a Multi-Day Travel Survey. Journal of Transport Geography, Vol. 78, 2019, pp. 194–204. https://doi.org/10.1016/j.jtrangeo.2019.06.001.

29. Feng, T., and H. J. P. Timmermans. Integrated Imputation of Activity-Travel Diaries Incorporating the Measurement of Uncertainty. Transportation Planning and Technology, Vol. 42, No. 3, 2019, pp. 274–292. https://doi.org/10.1080/03081060.2019.1576384.

30. Chen, C., S. Jiao, S. Zhang, W. Liu, L. Feng, and Y. Wang. TripImputor: Real-Time Imputing Taxi Trip Purpose Leveraging Multi-Sourced Urban Data. IEEE Transactions on Intelligent Transportation Systems, Vol. 19, No. 10, 2018, pp. 3292–3304. https://doi.org/10.1109/TITS.2017.2771231.

31. Han, G., and K. Sohn. Activity Imputation for Trip-Chains Elicited From Smart-Card Data Using a Continuous Hidden Markov Model. Transportation Research Part B: Methodological, Vol. 83, 2016, pp. 121–135. https://doi.org/10.1016/j.trb.2015.11.015.

32. Nguyen, M. H., J. Armoogum, J.-L. Madre, and C. Garcia. Reviewing Trip Purpose Imputation in GPS-Based Travel Surveys. Journal of Traffic and Transportation Engineering (English Edition), Vol. 7, No. 4, 2020, pp. 395–412. https://doi.org/10.1016/j.jtte.2020.05.004.

33. Feng, T., and H. J. P. Timmermans. Comparison of Advanced Imputation Algorithms for Detection of Transportation Mode and Activity Episode Using GPS Data. Transportation Planning and Technology, Vol. 39, No. 2, 2016, pp. 180–194. https://doi.org/10.1080/03081060 .2015.1127540.

34. Shen, L., and P. R. Stopher. A Process for Trip Purpose Imputation From Global Positioning System Data. Transportation Research Part C: Emerging Technologies, Vol. 36, 2013, pp. 261–267. https://doi.org/10.1016/j.trc.2013 .09.004.

35. Wolf, J., R. Guensler, and W. Bachman. Elimination of the Travel Diary: Experiment to Derive Trip Purpose From Global Positioning System Travel Data. Transportation Research Record: Journal of the Transportation Research Board, 2001. 1768: 125–134.
36. Bohte, W., and K. Maat. Deriving and Validating Trip Purposes and Travel Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in The Netherlands. *Transportation Research Part C: Emerging Technologies*, Vol. 17, No. 3, 2009, pp. 285–297. https://doi.org/10.1016/j.trc.2008.11.004.

37. Stopher, P., C. FitzGerald, and J. Zhang. Search for a Global Positioning System Device to Measure Person Travel. *Transportation Research Part C: Emerging Technologies*, Vol. 16, No. 3, 2008, pp. 350–369. https://doi.org/10.1016/j.trc.2007.10.002.

38. Chung, E.-H., and A. Shalaby. A Trip Reconstruction Tool for GPS-Based Personal Travel Surveys. *Transportation Planning and Technology*, Vol. 28, No. 5, 2005, pp. 381–401. https://doi.org/10.1080/03081060500322599.

39. Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, and K. W. Axhausen. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. *Transportation Research Record: Journal of the Transportation Research Board*, 2004: 46–54.

40. Chen, C., H. Gong, C. Lawson, and E. Bialostozky. Estimating Missing Values in Neural Network-Based Diagnostic Systems. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, 1998, pp. 18–28.

41. Usyukov, V. Methodology for Identifying Activities From GPS Data Streams. *Procedia Computer Science*, Vol. 109, 2017, pp. 10–17.

42. Furlatti, B., P. Cintia, C. Renso, and L. Spinsanti. Inferring Human Activities From GPS Tracks. *Proc., 2nd ACM SIGKDD International Workshop on Urban Computing*, Chicago, IL, 2013, pp. 1–8.

43. Feng, T., and H. J. P. Timmermans. Transportation Mode Recognition Using GPS and Accelerometer Data. *Transportation Research Part C: Emerging Technologies*, Vol. 37, 2013, pp. 118–130. https://doi.org/10.1016/j.trc.2013.09.014.

44. Moiseeva, A., J. Jessurun, and H. Timmermans. Semi-automatic Imputation of Activity Travel Diaries: Use of Global Positioning System Traces, Prompted Recall, and Context-Sensitive Learning Algorithms. *Transportation Research Record: Journal of the Transportation Research Board*, 2010: 2183: 60–68.

45. Gopalakrishnan, R., C. A. Guevara, and M. Ben-Akiva. Combining Multiple Imputation and Control Function Methods to Deal With Missing Data and Endogeneity in Discrete-Choice Models. *Transportation Research Part B: Methodological*, Vol. 142, 2020, pp. 45–57. https://doi.org/10.1016/j.trb.2020.10.002.

46. Li, L., J. Zhang, Y. Wang, and B. Ran. Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 8, 2019, pp. 2933–2943. https://doi.org/10.1109/TITS.2018.2869768.

47. Deb, R., and A. W.-C. Liew. Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data. *Information Sciences*, Vol. 339, 2016, pp. 274–289. https://doi.org/10.1016/j.ins.2016.01.018.

48. Rodrigues, F., K. Henrickson, and F. C. Pereira. Multi-Output Gaussian Processes for Crowdsourced Traffic Data Imputation. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 2, 2019, pp. 594–603. https://doi.org/10.1109/TITS.2018.2817879.

49. Data Management Group. Transportation Tomorrow Survey. http://www.dmg.utoronto.ca/.

50. Schouten, R. M., P. Lugtig, and G. Vink. Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure. *Journal of Statistical Computation and Simulation*, Vol. 88, No. 15, 2018, pp. 2909–2930. https://doi.org/10.1080/00949655.2018.1491577.

51. Cranmer, S. J., and J. Gill. We Have to be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data. *British Journal of Political Science*, Vol. 43, No. 2, 2013, pp. 425–449. https://doi.org/10.1017/S0007123412000312.

52. Allison, P. Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology*, Vol. 112, 2003, pp. 545–557. https://doi.org/10.1037/0021-843X.112.4.545.

53. Little, R. J. A., and D. B. Rubin. *Statistical Analysis With Missing Data*. John Wiley & Sons, Ltd, Hoboken, NJ, 2019.

54. Ng, A., and M. Jordan. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naïve Bayes. In *Advances in Neural Information Processing Systems* (Dietterich, T., S. Becker, and Z. Ghahramani, eds.), Vol. 14, 2002.

55. Hearst, M. A. Trends & Controversies: Support Vector Machines. *IEEE Intelligent Systems and Their Applications*, Vol. 13, 1998, pp. 18–28.

56. Mitchell, T. M. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

57. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

58. Duda, R., P. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, Hoboken, NJ, 2000.

59. Sharpe, P. K., and R. J. Solly. Dealing With Missing Values in Neural Network-Based Diagnostic Systems. *Neural Computing & Applications*, Vol. 3, No. 2, 1995, pp. 73–77. https://doi.org/10.1007/BF01421959.

60. Gupta, A., and M. S. Lam. Estimating Missing Values Using Neural Networks. *Journal of the Operational Research Society*, Vol. 47, No. 2, 1996, pp. 229–238. https://doi.org/10.1057/jors.1996.21.

61. Langr, J., and V. Bok. GANs in Action: Deep Learning With Generative Adversarial Networks. Manning Publications, Shelter Island, NY, 2019.

62. Bengio, Y. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, Vol. 2, 2009, pp. 1–155. https://doi.org/10.1561/2200000006.

63. Vincent, P., H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and Composing Robust Features With Denoising Autoencoders. *Proc., 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 1096–1103.

64. Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman.
Missing Value Estimation Methods for DNA Microarrays. 
*Bioinformatics (Oxford, England)*, Vol. 17, No. 6, 2001, pp. 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

65. Enders, C. K. *Applied Missing Data Analysis*. Guilford Press, New York, NY, 2010.

66. Eekhout, I., R. M. de Boer, J. W. R. Twisk, H. C. W. de Vet, and M. W. Heymans. Missing Data: A Systematic Review of How They are Reported and Handled. *Epidemiology (Cambridge, Massachusetts)*, Vol. 23, No. 5, 2012, pp. 729–732. https://doi.org/10.1097/EDE.0b013e3182576cdb.

67. Huisman, M. *Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items*. DSWO Press, Leiden, 1999.