2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI) | 979-8-3503-4273-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTAI59109.2023.00029

DICE: Data Imputation for Cost Estimates from Multiple Sources to Model User Decision-Making

Hailun Wu
*Department of Electrical
and Computer Engineering
New York Institute of Technology*
New York, USA
hwu28@nyit.edu

Ziqian Dong
*Department of Electrical
and Computer Engineering
New York Institute of Technology*
New York, USA
ziqian.dong@nyit.edu

Roberto Rojas-Cessa
*Department of Electrical
and Computer Engineering
New Jersey Institute of Technology*
Newark, USA
rojas@njit.edu

Abstract*—Understanding key factors that affect users' commute mode choice is essential to design policies that promote sustainable transportation. However, the reliance on survey data for these studies often faces incomplete data challenges. One of the regional transportation surveys obtained for the study on commute mode decision-making misses 97% of the parking cost data, an important factor in people's decision-making. To tackle the problem, we propose the data imputation for cost estimates (DICE) scheme to synthesize data from multiple sources to infer the missing data. DICE linearly maps imputed values to missing entries based on the assumption that higher-income users can spend more on their commute. In the absence of ground truth data, we propose to use the accuracy of the regression model trained with the imputed data as a metric to evaluate DICE. We train the regression model with 75% of the imputed data, test it with the remainder, and evaluate it with the complete cases. The prediction accuracy of the test data and the evaluation data are 0.89 and 0.77, respectively. The results indicate that the imputed data and complete cases share similar distributions and the model trained with the imputed data can perform classification. We tested DICE using a 1995 transportation survey and a 2021 housing survey data sets where cost is considered a key feature in decision-making. In both cases, the regression model achieves higher than 0.7 prediction accuracy, which proves the applicability of DICE on different data sets.

***Index Terms*—data imputation, multiple data sources, decision-making, commute mode choice, regression, logistic regression**

I. Introduction

As climate change intensifies and becomes a global emergency, countries and organizations around the world have come together to act on reducing greenhouse gas (GHG) emissions [1]. In the United States (US), the transportation sector stands as the largest contributor to GHG emissions [2]. In 2021, nearly 30% of New York City's (NYC's) GHG emissions were generated by transport, 80% of which was generated by the two million light-duty vehicles registered in NYC [3]. To reach NYC's goal of carbon neutrality by 2050, the city has implemented policies to reduce car usage and improve infrastructure to provide alternatives to driving, such as walking, biking, and using public transportation [4]. Understanding what factors affect people's choice between driving and taking other alternative means of transportation is important for policy-makers to design policies and incentives to reach a city's sustainability goal.

Most studies on understanding the factors that affect people's decision on commute mode use surveys as data sources for their modeling approaches [5], [6]. NYC Department of Transportation (DoT) conducts surveys for residents every year and publishes annual travel results through the NYC Open Data platform [7]. The published data from such surveys have been used to model and predict travel behaviors of NYC commuters [8].

One of the challenges of using survey data is its completeness. Survey participants may not answer some of the questions in a survey thus rendering incomplete data. A traditional way to work with incomplete data is complete case analysis (CCA), also known as the listwise deletion method [9]. This approach discards any observation that has a missing value for any variable and uses only observations that are complete for the study. Table I shows an example of survey entries. Here, CCA excludes the observations with missing values and only uses those with complete information. In this case, the first, second, and fifth observations are discarded and only the third and fourth observations are used in CCA. As a result, a significant loss of information and bias occurs in the modeling approaches and analytic outcomes.

TABLE I
An example of survey data with complete cases and with missing values.

Index	**Variables**		
	Person ID	**Travel date**	**Parking cost**
1580	1901959101	06/07/2019	-Missing-
2781	1903394001	06/04/2019	-Missing-
3092	1903657901	06/05/2019	20.0
7323	1908189701	06/05/2019	50.0
84160	19588106401	06/29/2019	-Missing-

We are interested in studying how dominant factors such as cost [5], income level, and travel time [6] affect people's decisions on commute mode between driving and taking the subway in NYC. After analyzing the data from the most recent NYC transportation survey, the 2019 Citywide Mobility Survey [10], we notice that 97% of the parking costs for private vehicle commute trips are missing. The high missing rate renders CCA unsuitable because most of the records are

979-8-3503-4273-4/23/$31.00 ©2023 IEEE

DOI 10.1109/ICTAI59109.2023.00029

discarded. To tackle the missing data problem, we propose the Data Imputation for Cost Estimates (DICE) method, a data imputation approach that synthesizes information from multiple external data sources. We assume that the parking costs and the users' income follow normal distributions and higher-income users can spend more on parking. With these assumptions, we synthesize the parking cost from other data sources to impute the missing values and map the imputed parking costs to users based on their income level.

Another challenge that needs to be addressed is how to assess the performance of DICE under the lack of ground truth on parking costs in the original data. Here, we evaluate DICE by examining the accuracy of a regression model fitted with the imputed data. A commonly used ratio of training data to test data is $75 : 25$. The regression model is trained with 75% of the imputed data, tested with 25% of the imputed data, and evaluated with the complete cases, which is also referred to as the evaluation set in this study. We compare the test set accuracy with the evaluation set accuracy. The prediction accuracy can reflect how well the model can classify the decision variable. It can also show if the data fed into the model is similar to the training data. Thus, the test set accuracy and evaluation set accuracy can be used together as an indicator of the similarity in the distributions of the imputed data and complete cases.

The remainder of this paper is organized as follows: Section II highlights the related work in data imputation methods. Section III presents the data and its pre-processing, the proposed imputation method, and its performance metric. Section IV shows the results of the imputation method when we test it on multiple data sets, including a regional transportation survey, a national transportation survey, and a national housing survey. Section V concludes the paper.

II. Related Work

In addressing missing data challenges, two primary approaches are commonly employed: the CCA and data imputation, as outlined in Table II. The studies of dominant factors using historical or web-based surveys [5], [6] adopt a CCA approach. However, when the missing rate is high, CCA ends up discarding a significant amount of information, therefore, jeopardizing the model's accuracy. As an alternative, data imputation offers a means to replace missing values with estimated substitutes. This technique proves valuable, particularly when confronted with a high incidence of missing data, as it enables the retention of the available information.

Data imputation can be further categorized into multiple imputation (MI) [12] and single imputation (SI) [17], depending on the number of possible substituted values for one missing record. The final substituted value can be obtained by averaging the M generated values [12]. There are two imputation mechanisms for MI, multivariate normal imputation (MVNI) [13] and fully conditional specification (FCS) [14]. MVNI assumes variables jointly follow a multivariate normal distribution. The imputed values are obtained from the estimated multivariate normal distribution. Lee et al. [13] showed that MVNI can achieve a less biased result compared to CCA. The performance of MVNI implemented using the Markov chain Monte Carlo (MCMC) algorithm is generally similar to the performance of FCS [13]. FCS, on the other hand, is more flexible without the assumption of multivariate normality. To implement FCS, a regression model for the variable with missing values is fitted using the remainder as features. The same process repeats for all other variables with missing data. Lee and Carlin [14] reported that for a data set with a 33% missing rate, the performance of FCS is similar to that of MVNI.

In contrast to MI, SI generates a single value to replace the missing data. Examples of machine-learning-based (ML-based) SIs are regression imputation [15] and k nearest neighbors (kNN) imputation [16]. Due to the nature of ML, where the performance of the model relies heavily on the data quality of the training set, the training set must be sufficiently large and not noisy to achieve unbiased imputation results. Shrive et al. [15] concluded that when the missing rate of the data set is lower than 30%, the performance of regression imputation is similar to that of MI. By comparing the performance of various ML-based imputation methods, Troyanskaya et al. [16] demonstrated that kNN imputation outputs unbiased results when the missing rate is within 1% to 20%.

For a data set with a 97% missing rate, CCA is not feasible due to the high possibility of biased analysis outcomes. Furthermore, the high data missing rate and a small number of available training data also make MI and ML-based SI susceptible to biased imputation results. Here, we propose DICE, an imputation method that synthesizes data from multiple sources to map cost based on income level under high data missing rates.

III. Proposed Data Imputation for Cost Estimates

In this section, we introduce the data set, the 2019 Citywide Mobility Survey [10], the data pre-processing procedure, and the proposed data imputation for cost estimate, DICE. We also introduce an evaluation methodology when there is no sufficient ground truth in the data set.

A. Data Set

The 2019 Citywide Mobility Survey is the latest NYC annual travel survey conducted by NYC DoT. The published survey results include responses to travel-related questionnaires submitted by 3,346 participants, including their daily travel mode, trip purpose, trip distance, and more. The survey data has 85,459 records with trip information and participant demographics.

As indicated in previous studies, cost [5] and travel time [6] are considered top factors that affect peoples' commute mode choice. In this paper, we are interested in how these factors weigh in the NYC travel survey. In the 2019 Citywide Mobility Survey, participants were asked to provide information on their expenses for parking in a garage or on the street. Because parking costs in NYC are higher than other driving costs such as tolls, gas, or electricity combined [18], we use parking costs

TABLE II
EXISTING SOLUTIONS FOR MISSING DATA ON DATA SETS

Category	Imputation mechanism	Maximum missing rate (%)	Performance metric
Complete case analysis [11]	N/A	—	N/A
Multiple imputation [12]	MVNI: multivariate normal distribution [13]	50	Absolute bias and RMSE
	FCS: regression model [14]	33	Bias
Single imputation	Regression model [15]	20	Bias
	kNN model [16]	20	RMSE
Proposed method: DICE	Synthesizing from multiple sources	97	Regression model accuracy

to represent the cost of driving in our model. This survey also provides information on the trip distance, which has no missing values. To study the effect of cost and travel distance on commuters' transportation mode choice, we set the travel mode as the decision variable, which can be either driving a private vehicle or taking the subway.

B. Data Pre-processing

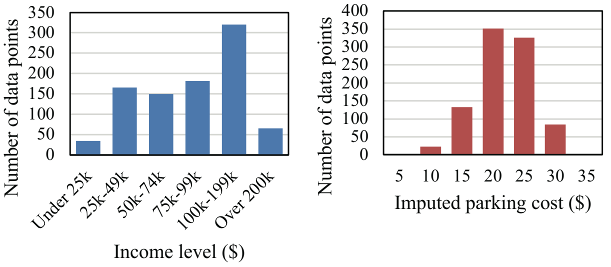
The pre-processing of the data set contains two steps: outlier removal and commute trip selection. For outlier removal, we calculate the z scores for each value x in distance and cost. The z score is defined as $z = \frac{x-\mu}{\sigma}$ where μ is the mean and σ is the standard deviation of the variable. We consider values beyond three times the standard deviation as outliers, or $z > 3$, and they are exclude from these records. In the commute trip selection step, we select commute trips completed with private vehicles or subways by participants aged from 18 to 54 years old. After data pre-processing, 2,200 records remain in the data set, which is referred to as 2019 CMS in this study. Among all those reported in this data set, 1,136 trips are done with private vehicles. However, only 33 out of those have reported parking costs, leading to a 97% missing rate of the cost of private vehicle trips.

C. Proposed Data Imputation Method

To tackle this missing data problem, we propose DICE to synthesize parking cost information from multiple external data sources, including Parking Meters [19], Municipal Parking Facilities [20], and Icon Parking Monthly Parking Deals in Manhattan [21]. These sources contain the parking rate information for street parking, municipal parking, and private garage parking in NYC, respectively.

The cost for 8-hour parking on a weekday in NYC depends on the parking location and whether a monthly pass is used. An 8-hour weekday street parking cost is $37 [19]. For municipal [20] and private parking garages [21], the parking cost for 8 hours ranges between $18 and $30 at the time of writing this paper. A monthly parking pass is available for both municipal and private parking garages. The municipal garage parking monthly rate is $500 in Manhattan [20] while private garages, such as Icon Parking Garages, charge between $400 and $600 per month [21]. The average parking cost for a day is approximately $20.

We assume parking costs for private vehicle commute trips follow a normal distribution, and thus, the imputed cost $\hat{c}$ can be written as $\hat{c} \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = 20$. For the standard deviation of $\hat{c}$, we assume that it remains the same as the 33 complete cases, where $\sigma = 6.87$. The imputed parking costs for private vehicle commute trips can be written as $\hat{c} \sim \mathcal{N}(20, 6.87^2)$.



(a) Income. (b) Imputed Parking Cost.

Fig. 1. The histograms of the users' income and the imputed parking cost.

Algorithm 1 DICE

```
Input: data with missing values, $\mu$, $\sigma$
$n \leftarrow$ the number of rows with missing values in the data
$\hat{c} \leftarrow$ generate $n$ random numbers from $\mathcal{N}(\mu, \sigma^2)$
while $n > 0$ do
    Linearly map larger values from $\hat{c}$ to higher income users that have no parking cost recorded
    $n \leftarrow n - 1$
end while
Output: imputed data
```

We obtain the income levels i of users from the survey. The users are grouped into 6 income levels from under $25,000 to over $200,000. Fig. 1(a) shows the distribution of the income level as a histogram. Because the histogram resembles a bell shape, we assume the users' income levels follow a normal distribution. With the assumption that higher-income users can pay more, the imputation rule for parking costs can be written as $\hat{c} \propto i$. The pseudocode for the cost imputation is shown in Algorithm 1, mapping higher parking costs to higher-income users' trips. The histogram of the imputed parking costs is shown in Fig. 1(b). These two figures illustrate that the generated normally-distributed parking costs are linearly mapped to the users' trips based on the income of the users.

D. Evaluation of the Proposed Imputation Method

Due to the 97% missing rate, most of the actual values of parking costs are not available in the original data. As a result, a commonly used performance metric, RMSE [13], [16], cannot be used to evaluate the imputation method. Therefore, we evaluate DICE with the performance of the regression model trained with the imputed data. The performance of the regression model can reflect the similarity in the distribution between the training data and the test or evaluation data sets. The features of the regression model are trip distance and cost. The decision variable is the travel mode, taking on binary values denoting either driving or using the subway. Therefore, the decision threshold is 0.5. A logistic regression model is trained with 75% of the imputed data. The performance of the regression model is measured by the prediction accuracy when the rest of the imputed data and the complete cases are fed into the model. They are also termed test set accuracy and evaluation set accuracy in this paper. Test set accuracy reveals the model's proficiency in correctly classifying travel modes based on distance and cost. In the realm of binary classification problems, an accuracy of 0.7 or higher is generally considered acceptable [22]. By comparing the accuracy of the evaluation set with that of the test set, we can assess the similarity in the distribution of the imputed data and complete cases. A smaller difference indicates superior imputation performance. The prediction accuracy for binary classification [23] is defined as

$$\text{Accuracy} = \frac{TP + FN}{TP + TN + FP + FN} \tag{1}$$

Here, the definitions of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are described in Table III. For example, when the user drives a vehicle to commute and the model predicts that the user takes the subway, this scenario falls under TN, where the prediction is incorrect. When the test set accuracy and evaluation set accuracy are both larger than 0.7 and close in values, it indicates the distribution of the imputed data is similar to the distribution of complete cases and thus the imputation method has a good performance.

TABLE III
DEFINITION OF VARIABLES FOR ACCURACY

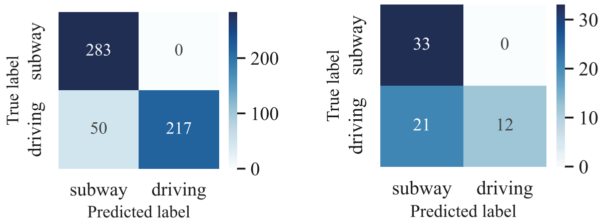
Accuracy Variables	**Definition**		
	True mode	**Predicted mode**	**Correct prediction**
TP	Driving	Driving	Yes
TN	Driving	Subway	No
FP	Subway	Driving	No
FN	Subway	Subway	Yes

IV. Results and Discussions

We test DICE on three data sets, the most recent NYC annual travel survey (2019 CMS), a historical national transportation survey, the 1995 Nationwide Personal Transportation Survey (1995 NPTS) [24], and a recent national housing survey, 2021 American Housing Survey (2021 AHS) [25]. The performance of DICE is then evaluated with the prediction accuracy of the regression model. We also compare the performance of DICE with an imputation method using the same data set (single-source) and kNN imputation, where $k = 2$. The features used for kNN imputation are distance and household income for 2019 CMS and 1995 NPTS. For 2021 AHS, the features are the number of bedrooms and the household income. The test is implemented with the Python *sklearn* library, where a default solver and $L2$ regularization were used for the regression model.

A. 2019 Citywide Mobility Survey

We impute parking costs in 2019 CMS with DICE following Algorithm 1. The performance of DICE is evaluated with the regression model prediction accuracy. Fig. 2 shows the confusion matrix of the commute mode prediction using the logistic regression. The accuracy of the test set and that of the evaluation set (1) are 0.89 and 0.77, respectively, indicating the high similarity of the imputed data and the complete case data. We also compare the performance of DICE with two other imputation methods on 2019 CMS, as shown in Table IV. Single-source imputation uses statistical information of the complete cases to impute the cost. However, the high missing rate biases the imputed data. The result is achieving accuracy below 0.7 when the model is fed with the test set and evaluation set. The accuracy for the model trained with kNN-imputed data has the largest difference as compared to DICE and single-source. This result indicates that with little data, kNN imputation is biased. Comparably, DICE is not susceptible to the high missing rate as using external data sources mitigates the bias caused by only relying on the small amount of complete data.



(a) Test set accuracy.

(b) Evaluation set accuracy.

Fig. 2. The confusion matrices of the regression model trained with DICE imputed data.

B. 1995 Nationwide Personal Transportation Survey

We also test DICE on a national transportation survey, 1995 NPTS [24] where we manually remove 97% of the parking costs for private vehicle trips. 1995 NPTS is a national transportation survey that contains 420,346 records with information on trip distance, parking costs, and more. After removing outliers, we obtain 1,037 commute trips, including those made in private vehicles or public transportation trips. 1995 NPTS includes data collected in multiple states in the US. Therefore,

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT IMPUTATION METHODS ON THREE DIFFERENT DATA SETS

Data sets	Imputation method	Accuracy		RMSE
		Test set	Evaluation set	
2019 CMS	DICE	0.89	0.77	—
	single-source	0.53	0.48	—
	kNN	0.66	0.27	—
1995 NPTS	DICE	0.75	0.75	2.681
	single-source	0.68	0.69	2.689
	kNN	0.70	0.72	2.714
2021 AHS	DICE	0.82	0.81	—
	single-source	0.82	0.80	—
	kNN	0.80	0.81	—

the travel modes are not evenly distributed. Out of 1,037 data points, 7.9% of them are labeled as public transportation trips and the rest of them are labeled as automobile trips. Although 1995 NPTS has fewer data points compared to 2019 CMS, due to travel mode imbalance, the number of data points with parking cost information is 29, which is similar to that of 2019 CMS. The mean and standard deviation of the available parking cost are calculated to be \$1.54 and \$1.26, respectively. Single-source imputation utilizes this information to impute cost.

The proposed imputation method, DICE gets average parking cost information from the 1990 Nationwide Personal Transportation Survey (1990 NPTS) [26]. After removing the outliers, the statistical analysis of the parking cost shows a mean of \$2.10. The imputed costs $\hat{c}$ follow a normal distribution $\mathcal{N}(2.10, 1.26^2)$ and are linearly mapped to users' trips following Algorithm 1.

We also apply DICE on public transportation trips in 1995 NPTS. The average cost for public transportation trips in 1995 was \$2 per day according to the 1995 Transit Fact Book [27]. A small variation of the costs exists in different areas and there is a possible discount of 10% to 30% if users have monthly passes. Using the 68–95–99.7 rule [28], we calculate the standard deviation of the public transportation cost to be \$0.2. The 68–95–99.7 rule states that 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively. The generated costs for public transportation trips also follow a normal distribution, $\mathcal{N}(2, 0.2^2)$ and are linearly mapped to users' public transportation trips following Algorithm 1. We evaluate the performance of the imputation method using two metrics, the accuracy of the regression model and RMSE [13], [16]. Here, the class weight of the logistic regression model built with *sklearn* library is set to *balanced*. In addition to the accuracy of the regression model, we also show RMSE of the imputed costs since the ground truth on parking costs c are available for 1995 NPTS. RMSE measures how well a model fits a data set and is commonly used as the performance metric for imputation methods. RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{\hat{c}_n - c_n}{c_n}\right)^2} \tag{2}$$

where N is the number of data points that have missing values, $\hat{c}_n$ is the imputed value for the n^{th} record and c_n is the actual value for the n^{th} record. Without a fixed threshold limit, the smaller the RMSE, the less biased the imputation result.

The performance comparison of the three imputation methods on 1995 NPTS is shown in Table IV. DICE achieves the lowest RMSE and highest accuracy for test and evaluation sets. It validates the use of regression accuracy to evaluate imputation performance when part of the actual values is not available. The higher test set and evaluation set accuracy for DICE compared to single-source and kNN imputations also show that DICE is less biased when the missing rate is high.

C. 2021 American Housing Survey

To test the imputation method on data sets other than transportation surveys, we perform data imputation on 2021 AHS [25]. 2021 AHS is the latest national housing survey where the responses regarding housing information are collected. In 2021 AHS, housing information, for example, monthly housing cost, household income, market values, the number of bedrooms, and more are available.

We aim to obtain a similar number of data records as 2019 CMS and 1995 NPTS. Thus, we filter single-person houses owned or rented by householders aged between 30 to 35. Then, we remove housing cost, market values, and household income outliers, and obtain 1,153 records. Before imputation, we manually remove 97% of the monthly housing costs in 2021 AHS.

Takaaki et al. [29] showed that housing costs follow a normal distribution in the same period. From Statista [30], we obtain the average monthly housing cost in 2021, which is \$1,114. The computed standard deviation of the available monthly housing cost in 2021 AHS is \$816. Single-source imputation uses the mean and the standard deviation obtained from complete cases in 2021 AHS, which are \$1,442 and \$816, respectively. For DICE and single-source imputation, we assume that monthly housing costs and the users' income both follow normal distributions. We linearly map the imputed monthly housing costs to users with the assumption that higher-income users can pay higher housing costs each month.

We evaluate the performance of the three imputation models by measuring the prediction accuracy of the regression models. The features of the regression models are the imputed monthly housing costs and the number of bedrooms. The labels of the regression models are renting or owning the house. The performance of the imputation methods, which are evaluated in terms of the accuracy of the models, is shown in Table IV.

We observe that DICE achieves the highest accuracy for both test and evaluation sets as compared to the other two imputation methods. By applying DICE to housing survey data, we show that DICE applies to different data sets.

D. Discussion

The performance of DICE is tested on three data sets: 2019 CMS, 1995 NPTS, and 2021 AHS, as shown in Table IV. When the test set accuracy and evaluation set accuracy are both larger than 0.7 and close in values, it indicates that the imputed data share a similar distribution as the complete cases, and the model can classify the travel mode accurately.

The test on 2019 CMS achieves an accuracy of 0.89 and 0.77 for the test and evaluation sets, respectively. These results indicate the similarity in the distributions of the imputed data and the complete cases. DICE is also applied to 1995 NPTS. With the actual costs available, the RMSE is calculated for the imputed data. It shows that when the test set and evaluation set accuracy are the highest, the RMSE of the imputed data reaches the lowest. It validates the use of the accuracy of the regression model as the imputation performance metric. The applicability of DICE on diverse data sets is further validated using 2021 AHS data. The model accuracy of the test and evaluation sets are 0.82 and 0.81, respectively. This result further shows the applicability of DICE on different data sets.

V. Conclusions

We tackle the missing data problem in surveys for human behavior study and propose a data imputation method, DICE, to impute parking cost data with information from multiple data sources. With assumptions that parking costs and the income of users in the 2019 CMS follow normal distributions and higher-income users can pay higher parking costs, DICE imputes parking costs with a linear mapping to users based on their income level. The imputed data is used to train a regression model to predict the user's commute mode. The regression model is then tested with the remainder of the imputed data and evaluated with the complete cases. The prediction accuracy of the regression model is used as a performance measure of DICE. The tests on 2019 CMS, 1995 NPTS, and 2021 AHS show the applicability of DICE as the accuracy of the regression model using test and evaluation sets are all above 0.7. These results indicate the imputed data shares a similar distribution as the records without missing values and the applicability of DICE on various data sets.

Acknowledgment

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 1856032.

References

[1] United Nations, "The paris agreement," 2023. [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement

[2] United States Environmental Protection Agency, "Sources of greenhouse gas emissions," 2023. [Online]. Available: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

[3] "Electrifying new york: an electric vehicle vision plan for new york city," NYC Mayor's Office of Sustainability (MOS), 2021. [Online]. Available: https://www1.nyc.gov/html/dot/downloads/pdf/electrifying-new-york-report.pdf

[4] NYC Department of City Planning, "City of yes for carbon neutrality," 2023. [Online]. Available: https://www.nyc.gov/assets/planning/download/pdf/plans-studies/city-of-yes/carbon-neutrality-hero-series.pdf

[5] D. G. Chatman, "How density and mixed uses at the workplace affect personal commercial travel and commute mode choice," *Transportation Research Record*, vol. 1831, no. 1, pp. 193–201, 2003.

[6] J. Ko, S. Lee, and M. Byun, "Exploring factors associated with commute mode choice: An application of city-level general social survey data," *Transport policy*, vol. 75, pp. 36–46, 2019.

[7] NYC Office of Technology and Innovation (OTI), "Open data for all new yorkers," 2022. [Online]. Available: https://opendata.cityofnewyork.us/

[8] H. Mostofi, "The association between ICT-based mobility services and sustainable mobility behaviors of new yorkers," *Energies*, vol. 14, no. 11, p. 3064, 2021.

[9] M. Jamshidian and M. Mata, "Advances in analysis of mean and covariance structure when data are incomplete," in *Handbook of latent variable and related models*. Elsevier, 2007, pp. 21–44.

[10] NYC DOT, "Citywide mobility survey 2019," 2022. [Online]. Available: https://data.cityofnewyork.us/Transportation/Citywide-Mobility-Survey-Trip-Survey-2019/w9dc-u4ik

[11] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art." *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[12] 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, pp. 1–22, 1977.

[13] J. H. Lee, J. Huber Jr *et al.*, "Multiple imputation with large proportions of missing data: How much is too much?" in *United Kingdom Stata Users' Group Meetings 2011*, no. 23. Stata Users Group, 2011.

[14] K. J. Lee and J. B. Carlin, "Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation," *American journal of epidemiology*, vol. 171, no. 5, pp. 624–632, 2010.

[15] F. M. Shrive, H. Stuart, H. Quan, and W. A. Ghali, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods," *BMC medical research methodology*, vol. 6, pp. 1–10, 2006.

[16] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[17] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[18] Metropolitan Transportation Authority (MTA), "Car toll rates," 2023 (Accessed on 2023-06-23). [Online]. Available: https://new.mta.info/fares-and-tolls/bridges-and-tunnels/tolls-by-vehicle/cars

[19] NYC DOT, "Parking meters - citywide rate zones," 2023 (Accessed on 2022-12-06). [Online]. Available: https://data.cityofnewyork.us/Transportation/Parking-Meters-Citywide-Rate-Zones/da76-p95d

[20] ——, "Municipal parking facilities," 2020 (Accessed on 2022-12-06). [Online]. Available: https://www.nyc.gov/html/dot/html/motorist/parkinglist.shtml

[21] Icon Parking, "Icon parking monthly parking deals," 2023 (Accessed on 2022-12-06). [Online]. Available: https://iconparkingsystems.com/monthly-parking/midtown-west/monthly

[22] J. E. Fischer, L. M. Bachmann, and R. Jaeschke, "A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis," *Intensive care medicine*, vol. 29, pp. 1043–1051, 2003.

[23] C. E. Metz, "Basic principles of roc analysis," in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.

[24] United States Department of Transportation and Federal Highway Administration, "Nationwide personal transportation survey, 1995," 2006. [Online]. Available: https://doi.org/10.3886/ICPSR03595.v1

[25] U.S. Census Bureau, "2021 american housing survey," 2021. [Online]. Available: https://www.census.gov/programs-surveys/ahs.html

[26] United States Department of Transportation and Federal Highway Administration, "Nationwide personal transportation survey, 1990," 2006. [Online]. Available: https://doi.org/10.3886/ICPSR09816.v1

[27] American Public Transit Association, "Transit fact book," 1995. [Online]. Available: https://www.apta.com/wp-content/uploads/Resources/resources/statistics/Documents/FactBook/APTA-1994-1995-Transit-Fact-Book.pdf

[28] L. J. Kazmier, M. K. Staton, D. L. Fulks *et al.*, "Business statistics: based on schaums outline of theory and problems of business statistics," McGraw-Hill, Tech. Rep., 2003.

[29] O. Takaaki, M. Takayuki, S. Chihiro, W. Tsutomu *et al.*, "The evolution of house price distribution," Research Institute of Economy, Trade and Industry (RIETI), Tech. Rep., 2011.

[30] Statista Research Department, "Average Monthly Apartment Rent in the United States," 2023 (Accessed on 2023-04-13). [Online]. Available: https://www.statista.com/statistics/1063502/average-monthly-apartment-rent-usa/#statisticContainer