mathematics

MDPI

Article

Machine Learning Data Imputation and Prediction of Foraging Group Size in a Kleptoparasitic Spider

Yong-Chao Su [1], Cheng-Yu Wu [1], Cheng-Hong Yang [2,3,4,*], Bo-Sheng Li [2], Sin-Hua Moi [5] and Yu-Da Lin [6,*]

1 Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 80708, Taiwan; ycsu527@kmu.edu.tw (Y.-C.S.); u109551003@kmu.edu.tw (C.-Y.W.)
2 Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 80778, Taiwan; f108152169@nkust.edu.tw
3 Ph. D. Program in Biomedical Engineering, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
4 Drug Development and Value Creation Research Center, Kaohsiung Medical University, Kaohsiung 80708, Taiwan
5 Center of Cancer Program Development, E-Da Cancer Hospital, I-Shou University, Kaohsiung 82445, Taiwan; ed113177@edah.org.tw
6 Department of Computer Science and Information Engineering, National Penghu University of Science and Technology, Penghu 880011, Taiwan
* Correspondence: chyang@cc.kuas.edu.tw (C.-H.Y.); yudalinemail@gms.npu.edu.tw (Y.-D.L.)

Abstract: Cost–benefit analysis is widely used to elucidate the association between foraging group size and resource size. Despite advances in the development of theoretical frameworks, however, the empirical systems used for testing are hindered by the vagaries of field surveys and incomplete data. This study developed the three approaches to data imputation based on machine learning (ML) algorithms with the aim of rescuing valuable field data. Using 163 host spider webs (132 complete data and 31 incomplete data), our results indicated that the data imputation based on random forest algorithm outperformed classification and regression trees, the *k*-nearest neighbor, and other conventional approaches (Wilcoxon signed-rank test and correlation difference have *p*-value from < 0.001–0.030). We then used rescued data based on a natural system involving kleptoparasitic spiders from Taiwan and Vietnam (*Argyrodes miniaceus*, Theridiidae) to test the occurrence and group size of kleptoparasites in natural populations. Our partial least-squares path modelling (PLS-PM) results demonstrated that the size of the host web ($T = 6.890$, $p = 0.000$) is a significant feature affecting group size. The resource size ($T = 2.590$, $p = 0.010$) and the microclimate ($T = 3.230$, $p = 0.001$) are significant features affecting the presence of kleptoparasites. The test of conformation of group size distribution to the ideal free distribution (IFD) model revealed that predictions pertaining to per-capita resource size were underestimated (bootstrap resampling mean slopes <IFD predicted slopes, $p < 0.001$). These findings highlight the importance of applying appropriate ML methods to the handling of missing field data.

Keywords: machine learning; data imputation; group foraging; PLS-PM; ideal free distribution; kleptoparasitism; resource allocation

check for updates

Citation: Su, Y.-C.; Wu, C.-Y.; Yang, C.-H.; Li, B.-S.; Moi, S.-H.; Lin, Y.-D. Machine Learning Data Imputation and Prediction of Foraging Group Size in a Kleptoparasitic Spider. *Mathematics* **2021**, *9*, 415. https://doi.org/10.3390/math9040415

Academic Editor: Mikhail Kolev

Received: 16 January 2021
Accepted: 17 February 2021
Published: 20 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In natural populations, it is common to see multiple conspecific individuals foraging in a single resource patch. The simplest explanation for this phenomenon is the gathering of individuals in the vicinity of resources that are patchily distributed [1]. Theoretical models of foraging behavior have been developed to facilitate the prediction of the foraging group size [2]. Field ecologists frequently conduct ecological surveys of resource utility strategies in natural populations; however, the data they bring back are often incomplete due to instrument failure, human error, or weather conditions [3]. Researchers require a large number of samples to reveal features that could be used to predict the number of individuals in a resource patch and to assess the fitness costs and benefits of remaining

within a group of conspecifics. The data used for this cost–benefit analysis must first undergo pre-processing to compensate for missing values and eliminate noise [4].

Although conspecific individuals are potential competitors for limited resources, group foraging often increases the per-capita fitness of individuals. Overall, the allocation of resources to individuals in a population, which is referred to as dispersion economics [2], is governed by the carrying capacity of the resource patch. In natural populations, the resource size is a strong indicator of the foraging group size. The aggregation of individual animals often serves social functions that benefit all group members, as seen in parent–offspring aggregations [5], aggregations for mating [6], group hunting [7], the guarding of resources, and defence against predators [8]. The group-living spider *Argyrodes miniaceus* (Theridiidae), which conducts kleptoparasitism in host webs, is a convenient natural system to test hypotheses pertaining to group-size prediction [9].

The features that predict the group size have been studied in nine of the twenty known group-living Argyrodinae species [10]. Previous findings provide a solid background by which to test theoretical frameworks. Among group-living spiders, the primary feature related to group size is the size of the host web, which is positively correlated with the availability of prey, and Agnarsson [11] suggested that among group-living species, the distribution of kleptoparasites in the host web follows an ideal free distribution (IFD) [12]. IFD theory states that if individuals can distinguish the quality of habitats and migrate freely among them, then the individuals in that population will disperse according to the sizes of resource patches to maximize individual intake. Argyrodinae populations are convenient systems by which to model foraging group size. Note, however, that features must be collected from hundreds of host webs to make reliable predictions [13]. In studies of this scale, it is common to end up with incomplete data matrices due to irregularities occurring in the field. The complete deletion approach can be used to generate a clean and complete data matrix; however, this results in data loss, which reduces the sample size and corresponding statistical power. Note that the collection of field data from different field sites, often from different countries, is time-consuming and expensive. Furthermore, the missing data are not necessarily distributed randomly in the matrix, such that the complete deletion of missing data systematically leads to biases of analysis toward complete samples [4].

A variety of statistical methods have been developed for the imputation of ecological data. The arithmetic zero, mean, median, and linear regression methods [14,15] can be used to fill in missing values. Machine learning (ML) is widely used in analyzing large-scale datasets [16–20]. Recent advances in ML have led to the development of sophisticated imputation strategies [21–23]. In a study by Biessmann et al. [24], deep neural networks (DNNs) outperformed 11 baseline methods in the imputation of missing values. Tsai and Chang [25] proposed four imputation methods based on *k*-nearest neighbor (KNN) clustering for numeric data and mixed data (29 datasets). Their results revealed that instance selection and imputation could improve the classification of time-series data with missing values resulting from sensing interruptions. Li et al. [26], proposed a mixed model of long short-term memory (LSTM) and support vector regression (SVR) to impute missing values in a dataset with a high loss rate. Penone et al. [4] implemented ML in conjunction with phylogenetic information for the imputation of missing values in a life-history trait dataset.

In this study, we developed a novel protocol for the imputation of field data pertaining to *A. miniaceus* populations collected in Taiwan and Vietnam. The proposed system is referred to as the ecological machine-learning imputation tool (Eco-MIT) (https://gitlab.com/yudalinemail/imputation-data, accessed on 1 December 2020). We evaluated this model using a case study of group-living kleptoparasitic spiders [9]. We performed statistical analysis on group-size predictions derived using six types of data treatment: raw data (hereafter RAW, for complete deletion approach), ZERO, MEAN, KNN [3], classification and regression trees (CART) [27], and random forest (RF) [28]. The results were used to test three hypotheses: (1) the group size of kleptoparasites among host

webs follows the IFD model; (2) the presence/absence of kleptoparasites and the group size are determined primarily by features related to resource availability; (3) the densities of hosts and kleptoparasites and the physical conditions of microhabitats are key features in predicting the occurrence of kleptoparasites.

2. Materials and Methods

2.1. Field Site Selection and Surveys

This study surveyed three populations of *A. miniaceus* in Taiwan, including Huoyenshan (June 2008), Green Island (June 2019; November 2018), Orchid Island (November 2018), as well as Hanoi in Vietnam (September 2018). Using the transect line method, we surveyed 163 host webs of *Nephila pilipes* (Nephilidae), which are linearly distributed along light gaps (i.e., trails) in forests. The length and width of each transect were based on local populations of hosts within an area measuring 1.0×0.1 km^2. We searched exhaustively to find all potential host webs in the target area (with and without kleptoparasites), determined the coordinates of each web using a hand-held GPS device (Garmin eTrex Summit HC, Taipei, Taiwan), and counted the number of kleptoparasites present in each web. We also measured features related to resource size, microclimate, and resource density at each web site (see below).

2.2. Kleptoparasite Foraging Group Size and Occurrence

We examined features that could potentially be used to predict the size of kleptoparasites groups in a given web (Klepto NUM) and the occurrence of kleptoparasites (Klepto Y/N) for > 40 webs/transect. The putative features fell into three categories related to resource size, microclimate, and resource density. Resource size was determined by the body length of host (BLH), vertical length of web (VLW), horizontal length of web (HLW), and web area (WA). Microclimate was characterized by instant wind speed (IWS), height of the web from ground (HWG), humidity (H%), luminance (Lux), and temperature (Temp). Resource density estimates were based on the minimum distance to the next host (Min D Host) and minimum distance to the next kleptoparasitic foraging group (Min D Klepto) on a different web. VLW and HLW were used to calculate the web area using the formula for ovals ($\frac{\mathrm{VLW}}{2} \times \frac{\mathrm{HLW}}{2} \times \pi$). Note that only web area data were used for analysis. All data were obtained on the same clear day within a single, four-hour time interval. Distances to the nearest host web and distances to the nearest host web with kleptoparasites were estimated using GPS data. A hand-held mini-environmental meter (LM-8000, Lutron, Taipei, Taiwan) was used to measure the environmental features. We counted the numbers of female, male, and juvenile kleptoparasites in each web, the three of which were summed to give the group size.

2.3. Data Imputation

Data with missing values were rescued via substitution using MEAN and ZERO values, as well as three ML-based data imputation processes [29], namely KNN [3] (https://github.com/scikit-learn/scikit-learn/tree/master/sklearn/neighbors, accessed on 1 December 2020), CART [27] (https://github.com/scikit-learn/scikit-learn/tree/master/sklearn/tree, accessed on 1 December 2020), and RF [28] (https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/ensemble/_forest.py, accessed on 1 December 2020). All parameters of KNN, CART, and RF algorithms used the default values of the scikit-learn tool (Table 1). Figure 1 shows the flowchart of the ML-based data imputation processes for imputing missing data, including the following steps: (1) the columns of numerical data were normalized to the interval (0,1) to avoid imbalances in the weighting of any features when training the models, such as the bias toward larger values when using KNN. (2) We divided the original dataset (N = 163 samples) into a set with missing values (N = 31 samples) and a set without missing values (N = 132 samples). The latter dataset was termed the RAW dataset. (3) The mean absolute error (MAE) and root mean squared error (RMSE) were estimated using the leave-one-out cross-validation (LOOCV) method, which involves

sequentially selecting one sample from the RAW dataset. The trained ML-based models were used for imputing missing data. The final results were compared with a RAW dataset (i.e., the total dataset minus the missing data) (Supplementary Materials: Table S1). MAE was calculated as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|(y_i - \hat{y}_i)| \quad (1)$$

where m indicates the number of missing values, y_i is a real value from the RAW dataset, and $\hat{y}_i$ is the imputed value. RMSE was calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (2)$$

where m indicates the number of missing values, y_i is a real value from the RAW dataset, and $\hat{y}_i$ is the imputed value.

Figure 2 presents the workflow used to evaluate the effectiveness of the data imputation methods. One to three features (HWG, HLW, VLW, BLH, Lux, H%, Temp, and IWS) are then sequentially deleted from the sample in question. The deleted data points are used as a testing set, and the remaining (i.e., non-deleted) data points are used as a training set. These data imputation methods were ranked in terms of MAE and RMSE.

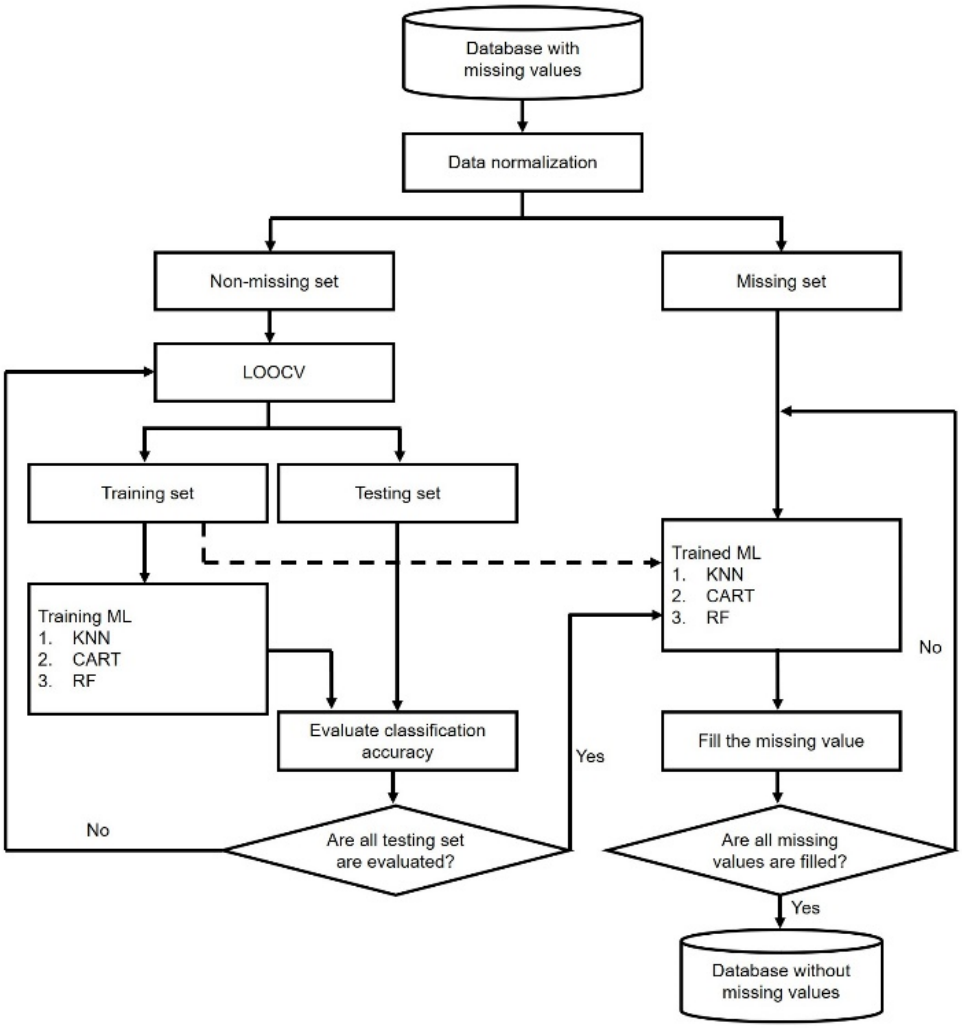


Figure 1. Flowchart of the machine learning methods for imputing missing data.

Table 1. The parameters of MEAN, ZERO, *k*-nearest neighbor (KNN), classification and regression trees (CART), and random forest (RF).

Algorithms	Parameters	Values
MEAN	No parameter	-
ZERO	No parameter	-
KNN	*k*	5
CART	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
RF	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	Number of trees	100

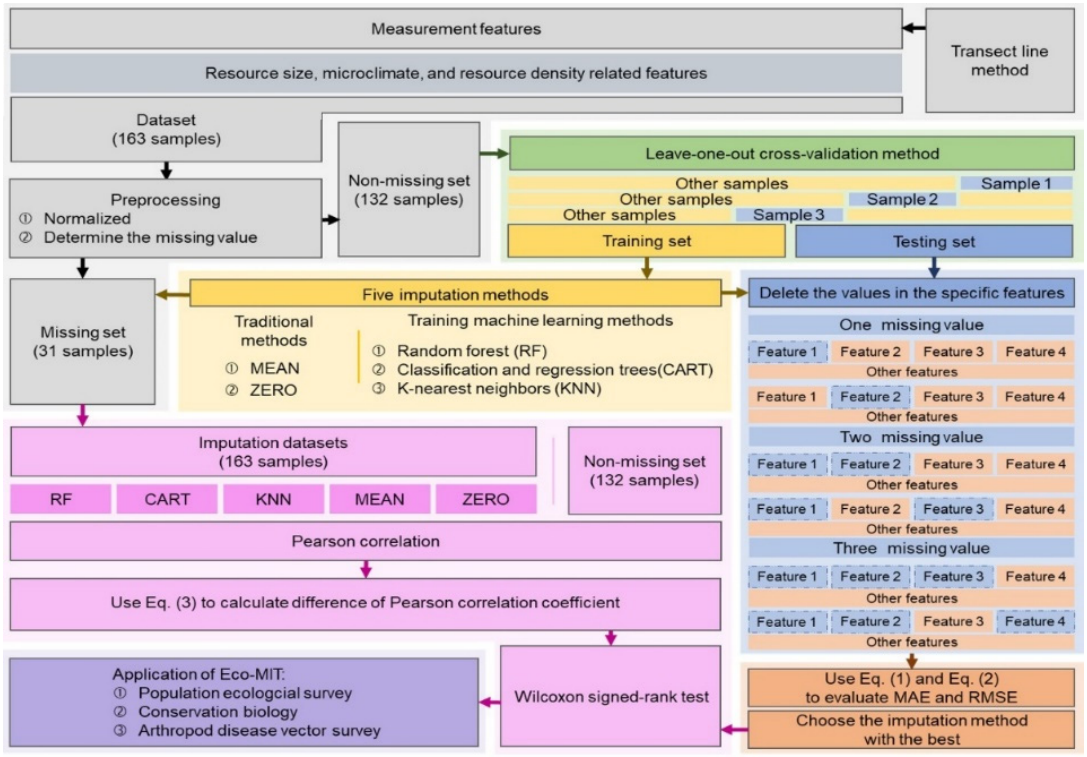


Figure 2. The procedures of simulation experiment design and the process of validation of different data imputation methods. The area shaded in gray is the process of numerical data standardization. We validated the qualities of imputation results using the leave-one-out cross-validation (LOOCV) method using test samples by evaluating the deviation of mean absolute error (MAE) and root mean squared error (RMSE) (shaded in green). For the missing set, the new sub-sets (pink blocks) are generated after the imputation by different methods. After integrating the new sub-set with the non-missing dataset, the Pearson correlation and Wilcoxon signed-rank test are used to evaluate the different methods deviation of the correlation. The positive ranks indicate the lower deviation, whereas the negative ranks indicate the larger deviation. The lower MAE and RMSE are the optimal imputation results. The application of the ecological machine-learning imputation tool (Eco-MIT) method is higlighted in purple.

We determined the values for data points in the missing set using a variety of methods and then substituted those values into the RAW dataset to create an imputed dataset. Finally, we estimated the degree of correlation between each feature from the imputation datasets using the Pearson correlation coefficient (Supplementary Materials: Table S2). We calculated the differences between the correlation coefficients of feature pairs in the

RAW dataset for each data imputation method (Supplementary Materials: Table S3). The differences between correlation coefficients were calculated as follows:

$$D = |I - N| \quad (3)$$

where I is the correlation coefficient of one feature pair from the imputation dataset, and N is from the RAW dataset. A smaller value means that there is less deviation between the imputation method and the RAW dataset, thereby indicating a superior imputation method. Calculations were performed using the Wilcoxon signed-rank test in SPSS version 22.

2.4. Statistical Significance of Features Related to Group Size

2.4.1. Partial Least-Squares Path Modelling (PLS-PM) of Population Size Features

Partial least-squares path modelling (PLS-PM) was used to evaluate the significance of latent features and sup-features relative to response features. This analysis was applied to all of the datasets. The results were then compared with the RAW dataset.

PLS-PM is one approach to structural equation modelling aimed at evaluating the direct and indirect regression relationships between features. Before conducting PLS-PM, we conducted pairwise correlation analysis between features using the R package *GGally* [30]. We then removed one of the collected features in cases where the correlation coefficient was >0.7 (Supplementary Materials: Figure S1). The retained features were used to construct the PLS-PM model using the R package PLSPM [31]. Two sub-models were used for PLS-PM: an inner model (path correlations between response features and latent features) and an outer model (path correlations between latent features and its block of sub-features). In the inner model, we defined two response features: Klepto NUM and Klepto Y/N of *A. miniaceus* foraging groups [31], and set resource size, microclimate, and resource density as latent features. Each latent feature was clustered according to its reflective/formative sup-features, called the outer model: (1) resource density was clustered in terms of reflective sup-features, as Min D Host and Min D Klepto. (2) Resource size was clustered in terms of formative sup-features, as WA and BLH. (3) We defined all environmental features as formative sup-features of microclimate. We used the loading (for reflective model) and weight (for formative model) $> |0.7|$ to determine the significance of a feature in the outer models. The importance of the latent features was evaluated using R^2 values. All statistical analysis was conducted in the R environment [32].

The Python package RFPIMP [4] (a Scikit-learn random forest approach, [4] RFPIMP: https://github.com/parrt/random-forest-importances, accessed on 1 December 2020) was used to evaluate the importance ranking of features to response features (Klepto NUM and Klepto Y/N). Using out-of-bag data, we bootstrapped the number of trees = 10,000 and set the minimum sample size at a leaf node of 1 with the other settings at their default values. For Klepto Y/N (discrete feature), we used the loss of mean accuracy. For Klepto NUM (continuous feature), we used the change in R^2. Evaluations were performed by removing the data of one feature and measuring the resulting statistical changes. The features were ranked in terms of their contribution to variations in the response features.

2.4.2. IFD Test: Group Size and Web Area

Tests were conducted to determine whether the theory of IFD held within our sample population. In this context, web area represents resource size. We calculated the sum of all web areas (WA)/total population size (N) (i.e., the total number of kleptoparasites in all samples) from the RAW and best-imputation datasets. This sum represents the per-capita web area (WA/N), or slope, under the assumption of IFD. We bootstrapped 50% of the data 10 times and 100 times to generate the distributions of the subsets. We repeated this process 100 times to generate a probability distribution for the slope mean (i.e., linear regression coefficients). The one-sample Mann Whitney U-Test was implemented in *SciPy* ver1.3.1 [33] to test the significance of the difference between theoretical values (from the RAW and best-imputation datasets) and the bootstrapped value in the Python package *pandas* ver. 0.25.1 (Pandas development team, 2020).

3. Results

In the original data (163 host spider webs, N = 132 complete data, and N = 31 incomplete data), the number of missing samples ranged from 2.45% to 17.79% of the values pertaining to each feature (mean = 4.17 ± 5.09%). Among the 163 samples, 43 had no kleptoparasites (26.71%). Among the 118 webs with kleptoparasites, the mean group size was 4.84 ± 8.19 individuals per web and the largest group included 44 individuals.

3.1. Comparison of Data Imputation Strategies

As shown in Figure 3a,b, the ZERO method presented the largest error in terms of MAE (0.757 ± 0.242–0.757 ± 0.155) and RMSE (0.757 ± 0.242–0.783 ± 0.135), which indicates that simply substituting zeros for the missing values greatly skewed the results. The MEAN method produced the second-highest MAE (0.226 ± 0.213–0.231 ± 0.122) and RMSE (0.226 ± 0.213–0.280 ± 0.134). Among the methods based on ML, RF achieved the smallest MAE (0.073 ± 0.100–0.076 ± 0.062) and RMSE (0.073 ± 0.100–0.098 ± 0.082).

We also used the Wilcoxon signed-rank test and the Pearson correlation coefficient to determine whether the RF method statistically outperformed the other methods. We compared the differences in the correlation coefficients between the RAW dataset and the datasets established using the various imputation methods (RAW-RF, RAW-ZERO, RAW-MEAN, RAW-KNN, and RAW-CART). Overall, RAW-RF significantly outperformed all of the other methods (Table 2).

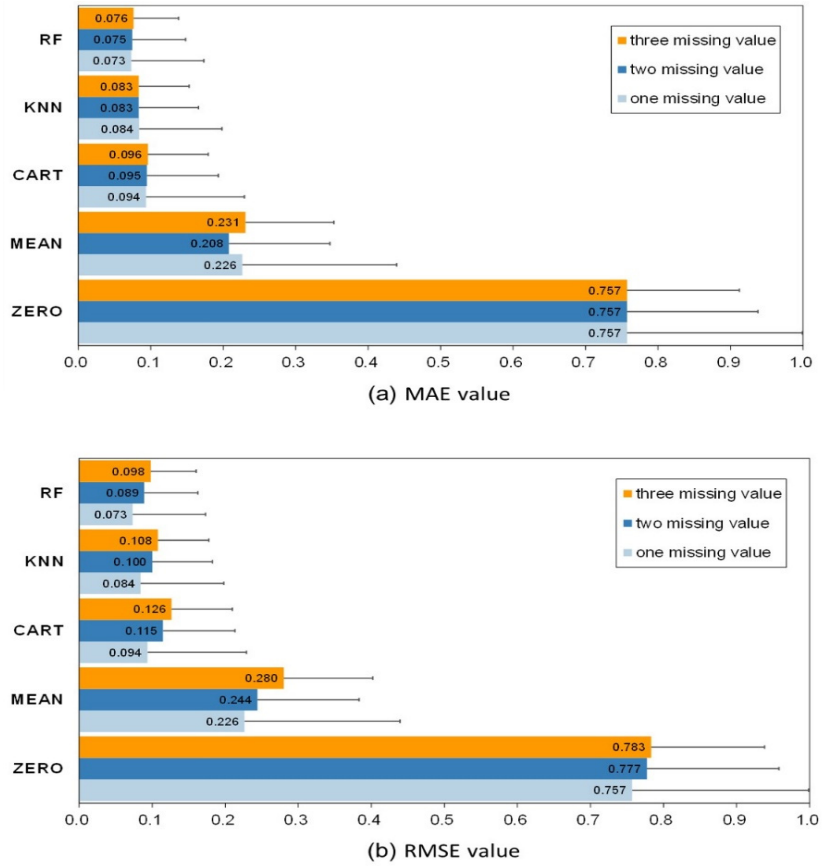


Figure 3. MAE value and RMSE value of each method with one, two, and three missing values. The RF method shows the best performance among all methods.

Table 2. Wilcoxon signed-rank test results of RAW-RF and other methods. To evaluate the effect of each method on the addition of missing datasets, we calculated the Pearson correlation difference between the imputation dataset and RAW. The Wilcoxon signed-rank test evaluated the result differences between RAW-RF and other methods. By comparing the difference of 66 Pearson correlations, results show that RAW-RF has more R^+ than other methods, which indicates that the Pearson correlation deviation of the dataset is smaller after the RF method supplement. RF is stable for the imputation of the dataset. The correlation difference between RAW-RF and another method was evaluated using a two-tailed significance test, in which the *p*-values were less than 0.05, indicating a significant improvement.

Method	Rank	Number	Mean Rank	Sum Rank	*p*-Value
RAW-ZERO	R^+	51	34.25	1746.5	<0.001
	R^-	9	9.28	83.5	
	$R^=$	6			
RAW-MEAN	R^+	34	29.19	992.5	0.030
	R^-	20	24.63	492.5	
	$R^=$	12			
RAW-KNN	R^+	32	27.8	889.5	0.014
	R^-	18	21.42	385.5	
	$R^=$	16			
RAW-CART	R^+	33	25.18	831	0.012
	R^-	15	23	345	
	$R^=$	18			

R^-, negative ranks; R^+, positive ranks, $R^=$ ties.

A comparison of the PLS-PM outcomes with the imputation dataset and the RAW dataset (Table 3) made it possible to determine whether our imputation method affected statistical interpretation. The effects of the two inner models were as follows: (1) the path correlations between Klepto NUM and latent features revealed that the trends produced by all of the methods matched the trend of the RAW dataset; i.e., resource size was significantly positively correlated with group size. The outer model of resource size produced larger weights for web area (WA weight = 0.775 to 0.886) than for the body length of hosts (BLH weight = 0.195 to 0.338), regardless of the imputation method. Unlike the RAW datasets where WA weight = 0.527 and BLH weight = 0.562, all of the imputation results identified web area as a more important factor of group size. The PLS-PM outcomes of KNN, CART, and RF data conformed to the RAW dataset in terms of resource density and microclimate. However, the PLS-PM outcomes of MEAN and ZERO data presented a trend opposite to that of the RAW dataset in terms of temperature weights. (2) The path correlations between Klepto Y/N and latent features presented the same trend for the imputed datasets and RAW dataset. For the outer models of each latent feature, resource size (with the feature WA weight = 0.621 to 1.137 from all datasets) and microclimate (with the feature HWG weights = 0.684 to 0.759) presented significantly positive regression to Klepto Y/N. Nonetheless, the sup-feature BLH from ZERO data presented a trend opposite to that of the other datasets.

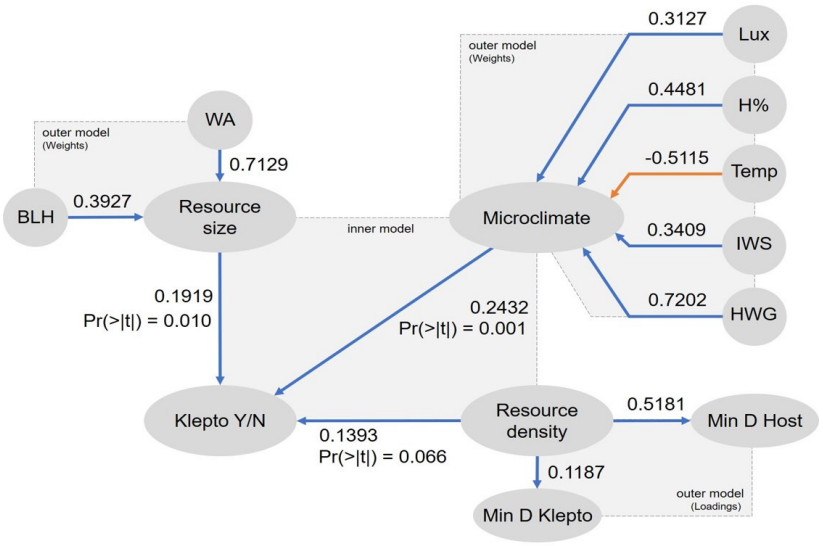
Table 3. Comparison of models constructed using the partial least square path model. (See details: Supplementary Materials: Figures S2–S6).

Klepto NUM													
Inner Model	**RAW**	**ZERO**	**MEAN**	**KNN**	**CART**	**RF**	**Outer Model**	**RAW**	**ZERO**	**MEAN**	**KNN**	**CART**	**RF**
Resource size							**Formative sup-features (weights)**						
Path coefficients	0.494	0.485	0.499	0.483	0.479	0.500	BLH	0.562	0.195	0.304	0.338	0.293	0.319
Pr(>|t|)	0.000	0.000	0.000	0.000	0.000	0.000	WA	0.527	0.886	0.788	0.775	0.804	0.775
Resource density							**Reflective sup-features (loadings)**						
Path coefficients	−0.104	−0.049	−0.076	−0.085	−0.084	−0.093	Min D Klepto	0.997	1.000	1.000	1.000	1.000	1.000
Pr(>|t|)	0.176	0.465	0.149	0.216	0.216	0.171	Min D Host	0.870	0.907	0.907	0.907	0.907	0.907
Microclimate							**Formative sup-features (weights)**						
Path coefficients	0.044	0.134	0.106	0.093	0.128	0.096	Lux	0.742	0.858	0.875	0.885	0.919	0.901
Pr(>|t|)	0.642	0.062	0.257	0.221	0.080	0.196	H%	0.407	0.108	0.189	0.141	0.137	0.143
							Temp	0.162	−0.060	0.094	0.083	0.070	0.069
							IWS	0.016	−0.016	−0.004	0.011	0.003	0.003
							HWG	0.529	0.361	0.322	0.305	0.254	0.278
Klepto Y/N													
Inner model	**RAW**	**ZERO**	**MEAN**	**KNN**	**CART**	**RF**	**Outer model**	**RAW**	**ZERO**	**MEAN**	**KNN**	**CART**	**RF**
Resource size							**Formative sup-features (weights)**						
Path coefficients	0.209	0.176	0.170	0.202	0.192	0.192	BLH	0.169	−0.361	0.293	0.514	0.431	0.393
Pr(>|t|)	0.013	0.019	0.024	0.007	0.011	0.010	WA	0.876	1.137	0.797	0.621	0.690	0.713
Resource density							**Reflective sup-features (loadings)**						
Path coefficients	0.069	0.132	0.128	0.135	0.137	0.139	Min D Klepto	0.842	0.119	0.119	0.119	0.119	0.119
Pr(>|t|)	0.414	0.083	0.092	0.076	0.072	0.066	Min D Host	0.991	0.518	0.518	0.518	0.518	0.518
Microclimate							**Formative sup-features (weights)**						
Path coefficients	0.276	0.255	0.257	0.234	0.237	0.243	Lux	0.250	0.309	0.304	0.275	0.371	0.313
Pr(>|t|)	0.002	0.001	0.001	0.002	0.002	0.001	H%	0.596	0.805	0.533	0.481	0.414	0.448
							Temp	−0.322	−0.671	−0.416	−0.454	−0.509	−0.512
							IWS	0.372	0.302	0.312	0.361	0.354	0.341
							HWG	0.741	0.684	0.738	0.759	0.718	0.720

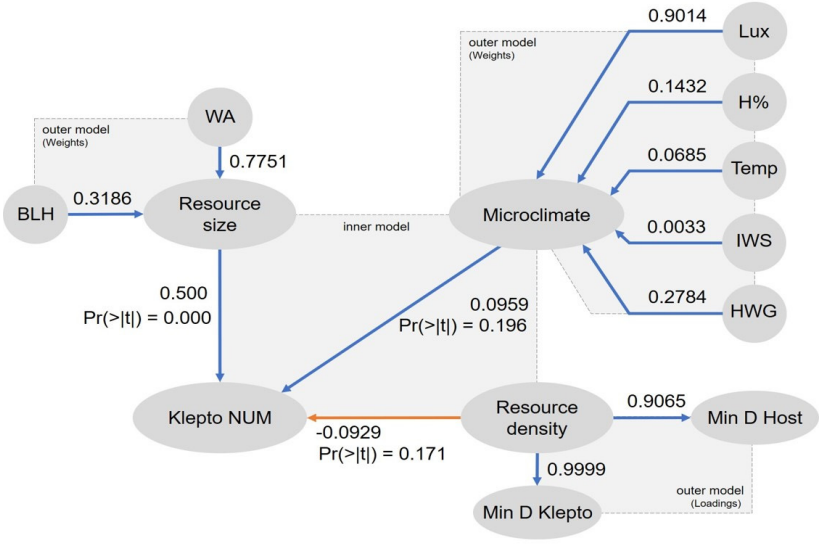
3.2. Biological Significance of Features Related to Group Size

All 163 samples from the RF dataset (Figure 4) were subjected to PLS-PM analysis to identify features capable of explaining the variation in the two response features (Klepto NUM and Klepto Y/N). In the inner model, (1) resource size ($T = 2.590$, $p = 0.010$) and microclimate ($T = 3.230$, $p = 0.001$) were both significantly correlated with Klepto Y/N (Figure 4a). For the outer model of each latent feature, the weight of WA was > 0.7 (weight = 0.713) with resource size, and the weight of WHG was > 0.7 (weight = 0.720) with microclimate (Figure 4a). (2) Resource size ($T = 6.890$, $p = 0.000$) was significantly correlated with Klepto NUM. The weight of WA was > 0.7 (weight = 0.775) with resource size (Figure 4b). The PLS-PM results revealed that resource size was significantly correlated with Klepto NUM

and Klepto Y/N. Furthermore, the microclimate of the web was significantly correlated with Klepto Y/N.



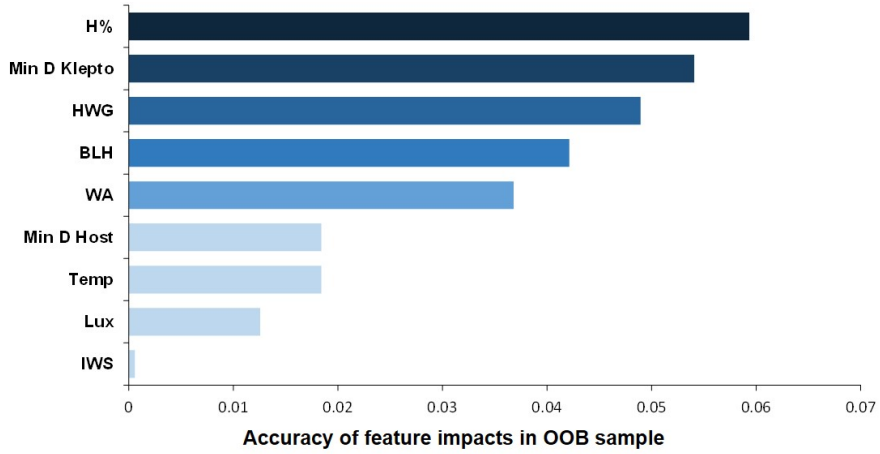
(a) PLS-PM results for Klepto Y/N



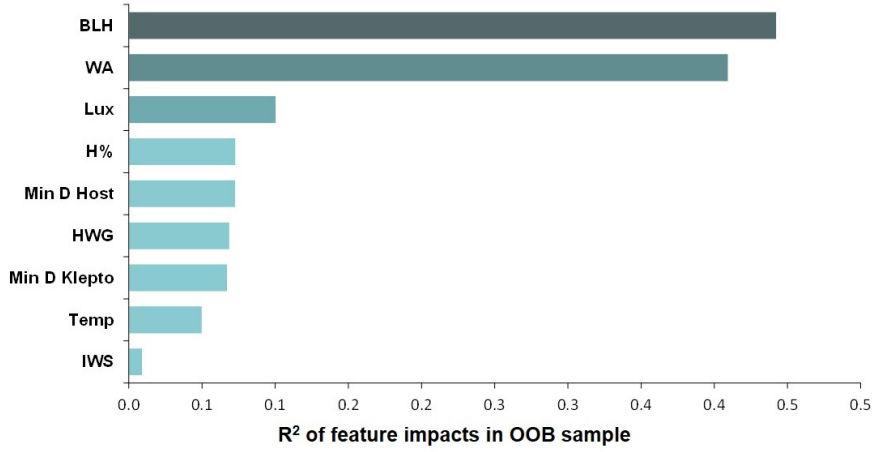
(b) PLS-PM results for Klepto NUM

Figure 4. The results of PLS-PM for the occurrence and group size of kleptoparasites from RF dataset. (**a**) The resource size and microclimate show significant contribution to the prediction of occurrence of kleptoparasites in which Area and height to the ground of a web show >0.70 of the weight. (**b**) Only the resource size (i.e., area, weight = 0.7751) shows a significant contribution to the group size.

The results of RF ranking matched those obtained using PLS-PM (Figure 5). In the predictive results for Klepto Y/N (Figure 5a), Min D Klepto, WHG, and BLH led to a > 20% increase in MSE the original data were randomized. In the predictive results of Klepto NUM (Figure 5b), WA and BLH led to a > 20% increase in MSE the original data were randomized. The results of RF ranking revealed that the microclimate (height of the web), resource size (area size and host size), and density of competitors (minimum distance to the kleptoparasites in another web) may also affect the occurrence and group size of kleptoparasites.



(a) Classification Importance for Klepto Y/N



(b) Regression Importance for Klepto NUM

Figure 5. Random forest importance ranking using (**a**) loss of mean accuracy for the occurrence and (**b**) using the change of R^2 for group size. The H%, Min D Klepto, height of the web from ground (HWG), body length of host (BLH), and web area (WA) have a higher contribution in predicting occurrence. Only the resource size associated features BLH and WA have a higher contribution in predicting group size.

The IFD test showed that the mean slopes of the bootstrapped per-capita web area (cm^2, area/individual), which represents the resource area that an individual can occupy, are less than the prediction of IFD. This pattern was observed in the RAW dataset (Figure 6a) as well as the RF imputed dataset (Figure 6b), regardless of bootstrapping. The mean slopes obtained from the bootstrapped data were significantly smaller ($p < 0.001$ for most of the resampling) than the slopes predicted under the IFD theory.

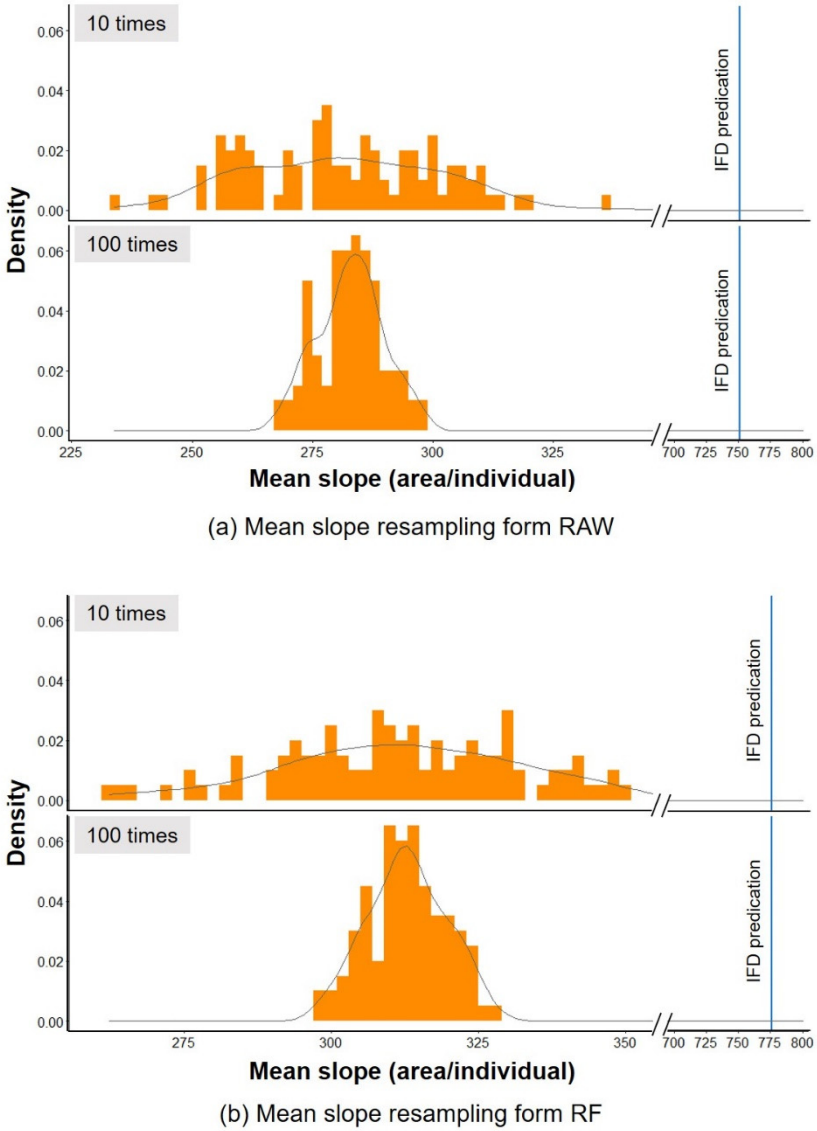


Figure 6. Bootstrapped resampling of mean slopes using RAW and random forest data imputation data. (**a**) The mean slope distribution of the bootstrapped 10 times and 100 times results showed strong under-matching of the per capita resource predicted under IFD (blue lines). (**b**) The results using the RF dataset show the same pattern as the results of RAW.

4. Discussion

The prediction and modelling of foraging group size is an important aspect of the theories related to foraging [34]. In this study, the feature with the most pronounced effect on foraging group sizes was resource size (Figure 5). The dispersion of kleptoparasites did not conform to the predictions of the IFD theory, in which overloading is expected in some resource patches (Figure 6). Note that the presence of kleptoparasites depended on resource size, as well as the density of neighboring competitors, and environmental conditions (humidity and elevation of the web above the ground). Moreover, the proposed data imputation method using the RF approach could be of considerable benefit to field ecologists examining the relationship between population dispersion and environmental factors.

Our results are in line with those of previous studies, indicating that the host area is crucial to the size of kleptoparasite groups [11,13]. The positive correlation between host areas and group sizes has previously been cited as evidence that the dispersion of kleptoparasites follows IFD [11,13]; however, our results revealed a smaller per-capita

resource occupation scenario. The case study of kleptoparasitic spiders conforms to a continuous-input IFD model [35,36] because the host webs have a constant influx of prey, which is rapidly consumed by kleptoparasites or the host. It has been proposed that the group size in a resource patch can be predicted by the size of the resource (or "dispersion economy") [2]. The deviation from IFD in our case study could be attributed to free access to other resource patches. This indicates that access to other resources could be used as an additional criterion for IFD [35,36]. In this study, the overloading of some resource patches (smaller per-capita resources than expected) showed evidence of impeded dispersal among webs or lower within-group competition. Impeded dispersal would lead to the depletion of resources as the group size increases; the reduction in competition could be attributed to the high conspecific tolerance of *A. miniaceus* [9].

In our PLS-PM results, the predictive power of resource size for group size was greater than those of resource density and microclimate (Figure 4a). However, microclimate significantly affects the occurrence of kleptoparasites (Figures 4b and 5b), and the density of kleptoparasites (Min D Klepto) outweighed resource size (Figure 5b) in our RF ranking. These results demonstrate the importance of accounting for resource patch conditions and the density of neighboring competitors when modelling foraging group size, especially under the IFD theory. Our results empirically demonstrate that in addition to resource size, dispersal probability, degree of competition, and interspecies competition [37], it is important to model various aspects of the ambient community.

The ZERO and MEAN algorithms are commonly used to fill in missing values; however, our results indicate that ZERO and MEAN have large errors in terms of MAE and RMSE. The KNN, CART, and RF are the classification algorithms that demonstrated more accuracy for data imputation. The differences in the correlation coefficients between the RAW dataset and the imputed dataset using imputation algorithms established that KNN, CART, and RF outperformed ZERO and MEAN. Classification algorithms can train the model and detect samples that are similar to a sample with missing values. Therefore, a sample with the missing values can refer to similar samples to determine values to fill in missing values. KNN is widely discussed and applied to pattern recognition, however, when the sample size is not large enough, KNN is no longer optimal, especially when compared to the inherent dimensionality of the feature space. CART can handle highly skewed or multi-modal numerical data, as well as classification predictors with ordinal or non-ordinal structure. However, if the regression tree in CART does not show significantly different segments, i.e., homogeneous, CART cannot perform well. RF is a special type of simple regression tree set, which is based on the majority vote (in the case of classification) or average (in the case of ensemble) to predict each tree in the set using some input data. RF is based on regression trees and allows for non-linear links and instability of variable influence across different segments, and it does not require a detailed model description. RF is provided by considering the differences in the set of missing value determinants in a column. The prediction of the imputed dataset is in the same range as the prediction of the RAW, which can prevent the trend opposite due to excessive deviation. Consequently, RF can be effectively used in data imputation. Table 4 shows the time and space complexities of MEAN, ZERO, KNN, CART, and RF. Although the time and space complexities of RF is higher when compared to the rest of the classifiers, however, RF can implement within an acceptable time and obtain the superior data imputation than MEAN, ZERO, KNN, and CART.

Mixed-type data and outlier data are common in situations that rely on self-collection and recording. These issues can be caused by clerical mistakes or data categorization. Numerous methods based on integrated learning have been developed to deal with this issue [38]. The selection of an appropriate learner depends on the type of data, the size of the dataset, and the research topic. In this study, the random forest method outperformed all other imputation methods in dealing with continuous and discrete data as well as noise. This is hardly surprising considering the sensitivity of the arithmetic means to extreme values, which can lead to large errors in data imputation. Note that this approach also

eliminates the need for data pre-processing when dealing with mixed-type and outlier data. The degree of variance is inversely proportional to the number of trees, which means that increasing the number of trees can reduce the likelihood of overfitting, albeit with an increase in computational complexity. The parameters of the random forest are easily adjusted, and no additional verification set is required. This is particularly beneficial in situations where datasets are difficult to obtain [39].

Table 4. Time and space complexities of MEAN, ZERO, KNN, CART, and RF.

Algorithms	Time Complexity	Space Complexity
MEAN	O(n)	O(n)
ZERO	O(1)	O(1)
KNN	O(nm)	O(nm)
CART	O($m \cdot n\log n$)	O(p)
RF	O($n\log n \cdot mt$)	O(pt)

n represents number of datapoints, m represents the number of features that determine, t is the number of trees, p is number of nodes in tree.

5. Conclusions

In this study, we proposed ML-based data imputation processes for imputing missing data. We analyzed the performance of MEAN, ZERO, KNN, CART, and RF methods in terms of performance of data imputation, time, and space complexities. After data imputation using the proposed ML-based data imputation processes, we determined that the RF method is superior to other imputation methods in dealing with continuous and discrete data, noise, and outlier data. Furthermore, the results of RF ranking were indicated to match those obtained using PLS-PM. Our results demonstrated that the size of the host web and the ambient environment are significant features of group size in natural populations of *A. miniaceus* in Taiwan and Vietnam. We argued that among group-living kleptoparasites, social interactions provide benefits that favor remaining in groups [39], which could be the reason underlying the deviation of our results from IFD theory. In the future, the proposed method could be applied to the ecological surveys of the endangered populations, or the surveys of the disease vector populations, where missing data might have a strong impact on the accuracy of the ecological inferences. This method, ECO-MIT, could also be improved when updated machine learning algorithms are available.

Supplementary Materials: The following are available online at https://www.mdpi.com/2227-7390/9/4/415/s1, Figure S1. Pearson correlation between features pair from RF dataset. Figure S2. The results of PLS-PM for (a) the occurrence and (b) group size of kleptoparasites from RAW dataset. Figure S3. The results of PLS-PM for (a) the occurrence and (b) group size of kleptoparasites from ZERO dataset. Figure S4. The results of PLS-PM for (a) the occurrence and (b) group size of kleptoparasites from MEAN dataset. Figure S5. The results of PLS-PM for (a) the occurrence and (b) group size of kleptoparasites from KNN dataset. Figure S6. The results of PLS-PM for (a) the occurrence and (b) group size of kleptoparasites from CART dataset. Table S1. All data matrices used in this article. Table S2. The data matrix of the Pearson correlation coefficient between each feature from the imputation datasets. Table S3. The data matrix of the difference between the correlation coefficient of feature pairs in the non-missing set and each data imputation method.

Author Contributions: Y.-C.S. leads this research, designs the experiments, and composes the manuscript for biological interpretation of the results. C.-Y.W. and S.-H.M. performed statistical analyses and collected the field data. B.-S.L. implements the algorithm for data imputation. C.-H.Y. coordinated and oversaw this study. Y.-D.L. participates in the design of the algorithm and writes the data imputation parts in the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The funding sources are the Ministry of Science and Technology, Taiwan (107-2311-B-037-004-MY3), and partially from the Ministry of Science and Technology, R.O.C. (under Grant no. 107-2811-E-992-500-, 107-2221-E-214-013-, 107-2811-E-992-500-, 108-2221-E-992 -031 -MY3 and 109-2811-E-992 -502 -MY2). Part of this research is from YCS dissertation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in supplementary material.

Acknowledgments: We thank the assistance of fieldwork from Academic of Sciences, Vietnam; We thank S. H. Su and I. V. Chiu for the assistance of fieldwork.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Begon, M.; Harper, J.L.; Townsend, C.R. *Ecology. Individuals, Populations and Communities*; Blackwell Scientific Publications: Oxford, UK, 1986.
2. Giraldeau, L.-A.; Caraco, T. *Social Foraging Theory*; Princeton University Press: Princeton, NJ, USA, 2018; Volume 73.
3. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
4. Penone, C.; Davidson, A.D.; Shoemaker, K.T.; Di Marco, M.; Rondinini, C.; Brooks, T.M.; Young, B.E.; Graham, C.H.; Costa, G.C. Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods Ecol. Evol.* **2014**, *5*, 961–970. [CrossRef]
5. Yi, Y.; Kim, Y.; Hikmat, A.; Choe, J.C. Information transfer through food from parents to offspring in wild Javan gibbons. *Sci. Rep.* **2020**, *10*, 1–9. [CrossRef] [PubMed]
6. Roth, T.S.; Rianti, P.; Fredriksson, G.M.; Wich, S.A.; Nowak, M.G. Grouping behavior of Sumatran orangutans (*Pongo abelii*) and Tapanuli orangutans (*Pongo tapanuliensis*) living in forest with low fruit abundance. *Am. J. Primatol.* **2020**, *82*, e23123. [CrossRef] [PubMed]
7. Steinegger, M.; Sarhan, H.; Bshary, R. Laboratory experiments reveal effects of group size on hunting performance in yellow saddle goatfish, *Parupeneus cyclostomus*. *Anim. Behav.* **2020**, *168*, 159–167. [CrossRef]
8. Teunissen, N.; Kingma, S.A.; Peters, A. Nest defence and offspring provisioning in a cooperative bird: Individual subordinates vary in total contribution, but no division of tasks among breeders and subordinates. *Behav. Ecol. Sociobiol.* **2020**, *74*, 1–9. [CrossRef]
9. Su, Y.-C.; Peng, P.; Elgar, M.A.; Smith, D.R. Dual pathways in social evolution: Population genetic structure of group-living and solitary species of kleptoparasitic spiders (Argyrodinae: Theridiidae). *PLoS ONE* **2018**, *13*, e0208123. [CrossRef]
10. Whitehouse, M. *Kleptoparasitic Spiders of the Subfamily Argyrodinae: A Special Case of Behavioural Plasticity*; Cambridge University Press: Cambridge, UK, 2011.
11. Agnarsson, I. Habitat patch size and isolation as predictors of occupancy and number of argyrodine spider kleptoparasites in Nephila webs. *Naturwissenschaften* **2011**, *98*, 163–167. [CrossRef]
12. Cardoso, J.C.F.; Gonzaga, M.O. Spiders follow an ideal free distribution based on traits of the plant community. *Ecol. Entomol.* **2020**. Early view. Available online: https://onlinelibrary.wiley.com/doi/10.1111/een.12951 (accessed on 1 December 2020). [CrossRef]
13. Agnarsson, I. Spider webs as habitat patches—the distribution of kleptoparasites (*Argyrodes*, Theridiidae) among host webs (*Nephila*, Tetragnathidae). *J. Arachnol.* **2003**, *31*, 344–349. [CrossRef]
14. Pigott, T.D. A review of methods for missing data. *Educ. Res. Eval.* **2001**, *7*, 353–383. [CrossRef]
15. Engels, J.M.; Diehr, P. Imputation of missing longitudinal data: A comparison of methods. *J. Clin. Epidemiol.* **2003**, *56*, 968–976. [CrossRef]
16. Soleymani, F.; Masnavi, H.; Shateyi, S. Classifying a lending portfolio of loans with dynamic updates via a machine learning Technique. *Mathematics* **2021**, *9*, 17. [CrossRef]
17. Jukic, S.; Saracevic, M.; Subasi, A.; Kevric, J. Comparison of ensemble machine learning methods for automated classification of focal and non-focal epileptic EEG signals. *Mathematics* **2020**, *8*, 1481. [CrossRef]
18. Nosratabadi, S.; Mosavi, A.; Duan, P.; Ghamisi, P.; Filip, F.; Band, S.S.; Reuter, U.; Gama, J.; Gandomi, A.H. Data science in economics: Comprehensive review of advanced machine learning and deep learning methods. *Mathematics* **2020**, *8*, 1799. [CrossRef]
19. Chen, J.-B.; Lee, W.-C.; Cheng, B.-C.; Moi, S.-H.; Yang, C.-H.; Lin, Y.-D. Impact of risk factors on functional status in maintenance hemodialysis patients. *Eur. J. Med Res.* **2017**, *22*, 1–8. [CrossRef]
20. Çatak, F.Ö. Classification with boosting of extreme learning machine over arbitrarily partitioned data. *Soft Comput.* **2017**, *21*, 2269–2281.
21. Raja, P.; Thangavel, K. Missing value imputation using unsupervised machine learning techniques. *Soft Comput.* **2020**, *24*, 4361–4392. [CrossRef]
22. Rafsunjani, S.; Safa, R.S.; Al Imran, A.; Rahim, M.S.; Nandi, D. An empirical comparison of missing value imputation techniques on APS failure prediction. *I. J. Inf. Technol. Comput. Sci.* **2019**, *2*, 21–29. [CrossRef]
23. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* **2018**, *8*, 1–10. [CrossRef]

24. Biessmann, F.; Salinas, D.; Schelter, S.; Schmidt, P.; Lange, D. "Deep" Learning for missing value imputationin tables with non-numerical data. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 22–26 October 2018; pp. 2017–2025.
25. Tsai, C.-F.; Chang, F.-Y. Combining instance selection for better missing value imputation. *J. Syst. Softw.* **2016**, *122*, 63–71. [CrossRef]
26. Li, L.; Zhang, J.; Wang, Y.; Ran, B. Missing value imputation for traffic-related time series data based on a multi-view learning method. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 2933–2943. [CrossRef]
27. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Schloerke, B.; Crowley, J.; Cook, D.; Briatte, F.; Marbach, M.; Thoen, E.; Elberg, A.; Larmarange, J. GGally: Extension to 'ggplot2'(R Package Version 1.3. 1). Date 2016-11-13. [Electronic Resource]. Available online: https://cran.r-project.org/web/packages/GGally/index.html (accessed on 1 December 2020).
31. Sanchez, G.; Trinchera, L.; Russolillo, G. plspm: Tools for Partial Least Squares Path Modeling (PLS-PM). *R Package Version 0.4* **2013**, *1*. Available online: https://cran.microsoft.com/snapshot/2014-11-23/web/packages/plspm/index.html (accessed on 1 December 2020).
32. R_Core_Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria; Available online: https://www.R-project.org/ (accessed on 1 December 2020).
33. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
34. Grinsted, L.; Deutsch, E.K.; Jimenez-Tenorio, M.; Lubin, Y. Evolutionary drivers of group foraging: A new framework for investigating variance in food intake and reproduction. *Evolution* **2019**, *73*, 2106–2121. [CrossRef]
35. Fretwell, S.D. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheor.* **1969**, *19*, 45–52. [CrossRef]
36. Tregenza, T. Common misconceptions in applying the ideal free distribution. *Anim. Behav.* **1994**, *47*, 485–487. [CrossRef]
37. Křivan, V.; Cressman, R.; Schneider, C. The ideal free distribution: A review and synthesis of the game-theoretic perspective. *Theor. Popul. Biol.* **2008**, *73*, 403–425. [CrossRef]
38. Krawczyk, B.; Minku, L.L.; Gama, J.; Stefanowski, J.; Woźniak, M. Ensemble learning for data stream analysis: A survey. *Inf. Fusion* **2017**, *37*, 132–156. [CrossRef]
39. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]