The Journal of Supercomputing (2022) 78:1759–1782
https://doi.org/10.1007/s11227-021-03934-x

Visual analysis for panel data imputation with Bayesian network

Hanbyul Yeon[1] · Seongbum Seo[1] · Hyesook Son[1] · Yun Jang[1]

Accepted: 4 June 2021 / Published online: 21 June 2021

Abstract

Bayesian network is derived from conditional probability and is useful in inferring the next state of the currently observed variables. If data are missed or corrupted during data collection or transfer, the characteristics of the original data may be distorted and biased. Therefore, predicted values from the Bayesian network designed with missing data are not reliable. Various techniques have been studied to resolve the imperfection in data using statistical techniques or machine learning, but since the complete data are unknown, there is no optimal way to impute missing values. In this paper, we present a visual analysis system that supports decision-making to impute missing values occurring in panel data. The visual analysis system allows data analysts to explore the cause of missing data in panel datasets. The system also enables us to compare the performance of suitable imputation models with the Bayesian network accuracy and the Kolmogorov–Smirnov test. We evaluate how the visual analysis system supports the decision-making process for the data imputation with datasets in different domains.

Keywords Visual analysis · Imputation · Missing data · Bayesian network

✉ Yun Jang
jangy@sejong.edu

Hanbyul Yeon
hbyeon109@gmail.com

Seongbum Seo
sbum.seo@gmail.com

Hyesook Son
atieer@naver.com

[1] Sejong University, Seoul, South Korea

1 Introduction

The Bayesian network model is a relative data-driven approach that explores similar data and weaves them to create approximate inference margin. The Bayesian network overcomes the characteristics of incomplete data such as irregular interval, measurement times, uncertain overall trend, and lack of data. However, if the data are missing, the predictions from the Bayesian network are difficult to trust. Missing data indicate that values are not recorded in the data for some reason, and missing data can occur in any analytical domain where data are utilized. If the data scientists exclude missing values in the analysis, the missing data lead to weak statistical power due to the reduction in the number of samples [39]. Besides, missing data may not be representative of the population, which may result in a bias in the analysis results [42]. For example, in a survey, item nonresponses do not occur randomly, and only those with specific characteristics refuse to respond to particular questionnaires. In this case, the sample is biased and cannot represent the population. To minimize problems caused by missing data, the data scientists must substitute missing values with appropriate values. Various techniques have been studied to estimate missing values, such as averaging, regression, probability-based single imputation models, and multiple imputations.

However, it is difficult to choose an optimal imputation model since the missing mechanism, ratio, and distribution in missing data determine an imputation model [9]. Therefore, depending on the knowledge level of the data scientists, a complete dataset is estimated, and the quality of the data analysis is determined [8]. From this perspective, the visualization has the potential to present complicated information by revealing missing patterns. Also, visual analysis is an effective way to discover an optimal solution for estimating missing data with interactive visual interfaces. In the visualization community, recently, many researchers have emphasized the importance of visualization approaches for missing data analysis [19]. Nevertheless, it is not easy to locate studies on visual analysis tools that support decision-making on how to deal with missing data [19, 41].

This paper presents a visual analysis system for estimating missing values in a prediction model with Bayesian networks. The system locates the missing patterns in the Bayesian network and then estimates missing values by employing imputation models, such as multiple imputation, weighted moving average, mean, and EM-spline. Besides, we compare the imputation performances with the accuracy of the Bayesian network and the Kolmogorov–Smirnov test. Our visual analysis system provides visual clues for the user to estimate the missing values correctly. The visual cues include identifying missing points, analyzing missing patterns, and validating imputation models. We demonstrate the usefulness of our proposed visual analysis system through use cases including physical growth prediction, regional movement prediction, and PM (particulate matter) prediction. The main contributions of this paper are as follows.

– We propose a visual analysis system that supports the decision-making for estimating missing values in panel datasets.

– We compare the performance of imputation models that estimate missing values to improve the accuracy of the Bayesian network.
– Our visual analysis system visually supports the process of selecting a suitable imputation model, which is difficult to achieve in the statistical approach.
– We validate our approach and system through use cases with physical growth, air quality, and regional movement panel datasets.

2 Related work

In this section, we review related studies, including statistical and visualization techniques for handling missing data.

2.1 Statistical techniques for managing missing data

Various imputation models have been studied to estimate the missing values appropriately. The data scientist must explore the causes of missing data before determining the appropriate imputation model since an applicable imputation model varies depending on the cause of missing data [8]. However, unless the data scientist is involved in the data collection process, it is challenging to discover the cause of the missing. Little et al. classify the missing characteristics into 3 types by using the parameter, R, indicating whether it is missing, the observed value, V_{obs}, and the missing value, V_{miss} [30], where R is a vector consisting of 1s for the observed values and 0s for the missing values. The missing types are divided into missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR indicates that observed values and missing values are independent and occur completely randomly, which means that the missing occurrence, R, is independent of the observed value V_{obs} and the missing value V_{miss}. MAR means that the missing occurrence R is interpreted only by the observed value V_{obs} and independent of V_{miss}. If the missing mechanism is MAR, the data scientist can expect the missing value from the observed value. Many imputation models assume that the missing mechanism is MAR [18]. However, since V_{miss} is not observed, it is not possible to verify that the missing mechanism satisfies MAR conditions [17]. NMAR implies that the missing occurrence R is related to both the missing value V_{miss} and the observed value V_{obs}. Since V_{miss} is not observable in the actual data, the relationship with R is unknown. Therefore, we can distinguish MCAR from other mechanisms NMAR and MAR depending on whether there is a relation between R and V_{obs}, but it is difficult to distinguish between MAR and NMAR [32].

Techniques for handling missing values include deletion, single imputation, and multiple imputations. The deletion can be applied only when the missing mechanism is MCAR, and the deletion includes Listwise deletion and Pairwise deletion depending on how data are deleted [2]. Listwise deletion is a way to delete the case containing the missing value completely. Pairwise deletion uses a case where statistics can be calculated from the data. In other words, even though some of the parameters in the case are missing, we do not delete the case if we can calculate the

statistics for other parameters. If the missing ratio is less than 5% of the data regardless of the missing mechanism, the deletion does not degrade the statistical power of the sample or cause a bias [21]. On the other hand, when the missing ratio is more than 5%, the deletion may cause a bias in the estimates of the sample, if the missing mechanism is either MAR or NMAR [8].

Single the imputation is a method of substituting a missing value with a single value and applicable when the missing mechanism is MCAR. The single imputation techniques include hot deck, mean, regression, and stochastic regression. Hot deck is a way of substituting a missing value with the most similar value from the observed values [3]. The approach to select one of the observed values includes random selection and nearest neighbor. Mean imputation is a way of substituting a missing value with the average of the observed values. The mean imputation is a better approach than the deletion in terms of reducing the loss of information. However, the mean imputation may cause a problem of reducing the variance of the estimated parameters and lowering the correlation with other parameters [37]. Regression imputation is a technique of determining a missing value with a regression estimated by setting the observed value of a missing parameter as a dependent parameter and the remaining parameters as explanatory parameters [17]. The regression imputation can be applied, assuming that the missing value has a strong correlation with the observed value. Stochastic regression adds the residual to the value that substitutes the missing value with the regression model. The advantage of adding the residuals is that we can restore the variability of the estimated parameters. Probabilistic regression has the advantage of reducing bias, but it can result in underestimating the variance of missing parameters [17].

The single imputation techniques mentioned above substitute each missing value with a single value. The advantage is that we can configure a complete dataset through the imputation. However, the standard error is underestimated since the imputation replaces the missing value with a single fixed value [13] without any alternatives. On the other hand, multiple imputation is a way of estimating a missing value with values obtained by performing an imputation multiple times differently from a single imputation. The multiple imputation can be used assuming that the missing mechanism of incomplete data is MAR, and it consists of three phases including imputation phase, analysis phase, and pooling phase [12]. The first step is to create n copies of datasets in which each missing value is imputed differently with a specific algorithm. It then analyzes each of the n complete datasets and calculates the parameter estimates and standard errors from the analysis results of each dataset. Finally, we combine the results of each dataset using Rubin's rule [39]. The multiple imputation has the advantage that it minimizes the bias of the estimated sample, unlike single substitution [24].

2.2 Visualization techniques for missing data

The fact that a value is missing in the data can convey valuable information independent of the observed value [7]. In the statistics domain, basic visualizations have been adopted to recognize missing values as valuable information and identify them.

MANET is the earliest interactive statistical software that visually depicts missing values [4, 22]. The MANET represents missing data and observed data in bar charts, box plots, scatter plots, mosaic plots, and histograms to help the data scientist recognize data incompleteness. Also, there is a VIM r-package that provides the visualization needed to search for missingness structures [44, 45]. VIM provides a way to display missing values using Parallel coordinates, Marginplot, and Matrix plots. In the parallel coordinates of the VIM package, the missing values are plotted as 0 and highlighted in a different color. However, since the missing value does not mean a zero value, it can confuse the data analyst. To compensate for this, Schulz et al. have proposed a new parallel coordinates plot by adding missing factors to the axes [40]. Besides, Interactive Statistical Graphics has been developed to support searching for missingness structures [14, 43].

There are studies of visualization techniques for exploring missing data and the impact of missing data visualization on the data analysis. Eaton et al. [16] have analyzed how users interpret graphs with missing data. They have found that users might misinterpret data depending on how the missing data are represented. Song et al. [41] have analyzed the influence of representing missing values in incomplete time series data on data recognition. They have proved that linearly interpolating missing values is effective in understanding the data. Fernstad [19, 20] have defined three missingness patterns to understand the missing mechanism. They have compared visualization techniques that represent missing values and proposed combinations of visualizations that enable to identify missing patterns.

Visual analytics has been in collaboration with various fields requiring decision-making such as predictive model design [15, 29], model verification [34], and data quality management [5] to prove the need of reasoning with analytical visual interfaces. Only a few studies have been published on visual analytics that supports decision-making to deal with missing data.

Alemzadeh et al. [1] present VIVID, a framework for the visual analytics of missing values in cohort study data. Yeon et al. [46] propose a visual analytics system to improve a Bayesian network model built from missing panel data. The visual analytics system focuses on analyzing the influence of missing data on predictive models.

Missing data are very likely to mislead the data scientist. Before analyzing data, the data scientist needs to find ways to recognize and resolve missing values proactively. Statistical scientists have developed various algorithms that can impute missing values. Since no one knows the optimal solution for estimating missing values, the data scientist chooses an appropriate imputation model based on various hypotheses. Therefore, it is necessary to examine a visual analytics approach to assist the selection decision of imputation models based on data characteristics.

3 Panel datasets with missing values

In this section, we introduce panel datasets containing missing values in three different domains, such as physical growth, air quality, and regional movement. We also present data characteristics of the three datasets.

3.1 Physical growth data

Physical growth data consist of student information and 20 body components for each student. The student information consists of an ID, gender, age, and date. The body components consist of the abdomen, arm, BMI, BMR, body fat, body water, chest, fat-free, height, hip ratio, mineral, neck, osseous, percentage of body fat, percentage of waist–hip, protein mass, skeletal muscle, soft lean, thigh, and weight. The data include 387,037 pieces of physical data collected for several years from 59,126 teenagers aged 7–18. We have received the physical growth data from a healthcare company. The student physical growth data have been collected using a body composition measurement device. However, due to the limited data collection conditions, data are missing. For this reason, the number of data and intervals are irregular for each student, as illustrated in Fig. 1a. The growth records for 7 to 12 years old are about 73% of the total data. The number of records for 13 years old and above is scarce, as presented in (b). The data are collected 1–5 times per year (average three times a year) for 1–7 years (average four years) for each student in (c). Moreover, the amount of data is unbalanced in specific ages, as seen in (d). Since the physical growth data are not entirely homogeneous, the data are missing overall, and it is not easy to predict the future values from the missing physical growth panel data.

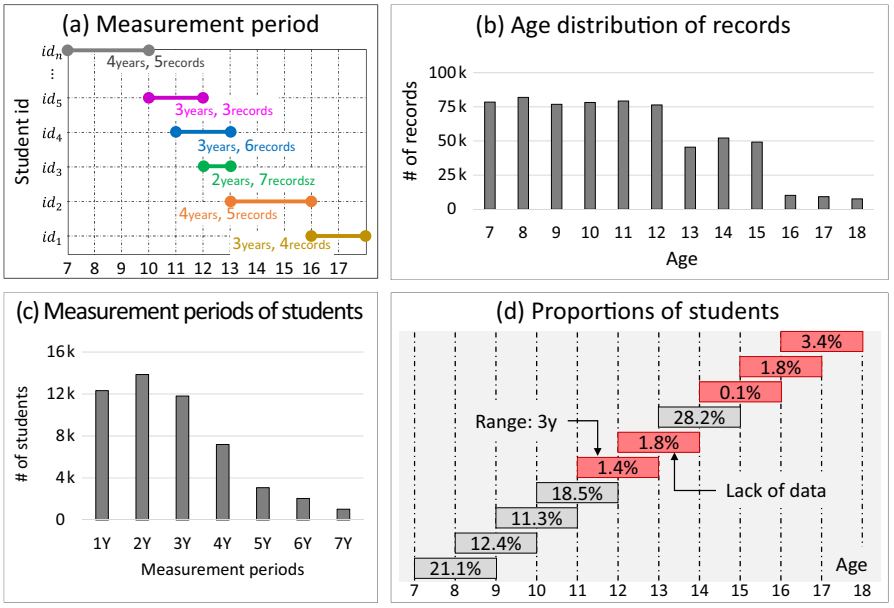


Fig. 1 **a** Each student has a different measurement period, time, and interval. **b** The number of measurement records for students ages 13–18 is relatively small. **c** The number of students measured for more than five years is relatively small. **d** The proportions of data are unbalanced in specific age ranges

3.2 Air quality panel data

Air quality panel data were collected from 413 discrete stations in Seoul, South Korea. We utilize the data measured for 75 days from September 5, 2019, to November 18, 2019. The air quality panel data consist of station id, temperature, humidity, $PM_{2.5}$, and PM_{10}. PM (particulate matter) is a particle generated naturally or artificially and contained in the air as an aerosol. PM is divided into PM_{10} whose diameter is 10 micrometers or less, and $PM_{2.5}$ whose diameter is 2.5 micrometers or less. PM is measured at the station, generally using an optical sensor. However, the optical sensor may operate erroneously due to physical limitations, such as humidity, which causes missing values in the data. Besides, the data collected from the stations are transferred across the networks and stored on the data collection server. During this process, about 18% of the total data were missing due to network delays. Therefore, it is not straightforward to predict future PM trends from the air quality panel data with missing values.

3.3 Regional movement data

Regional movement panel data represent the vehicle trajectories within urban regions over time. The regional movement panel data include region id, time, vehicle volume, vehicle speed, vehicle direction, and vehicle id. To analyze the taxi trajectory data in a more structured way, we create the regional movement data from Beijing taxi data, as presented in Fig. 2, for examining the macroscopic taxi movements within the entire city. The Beijing taxi GPS data consist of taxi id, time, and GPS coordinates, as illustrated in (a). We map the GPS coordinates onto 20 regions in Beijing, as seen in (b). We extract the trajectory sequence among the regions where the taxis visit, as observed in (c). We calculate the vehicle speed and vehicle volume using the approach presented by Pi et al. [38]. The Beijing taxi GPS data contain duplicated time stamps and unrealistic coordinates. We remove GPS coordinates with duplicate time stamps so that each taxi has only one GPS coordinate

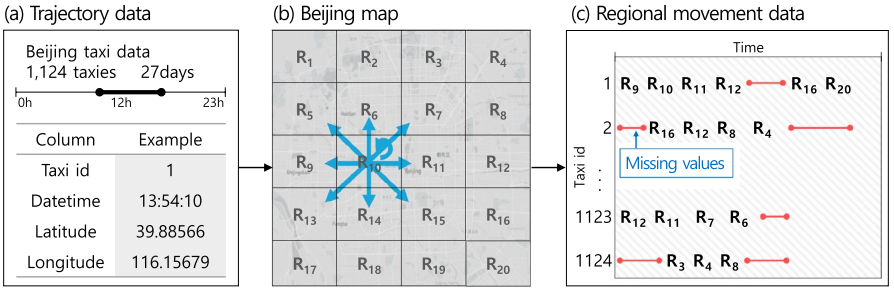


Fig. 2 Regional movement data with missing values from the Beijing taxi trajectory data. **a** The Beijing taxi trajectory data contain id, DateTime, and GPS coordinates. **b** We divide the Beijing area into 20 regions. Each region has nine movement directions. **c** We map the taxi trajectories onto the 20 regions for the regional movement data including missing movements. Note that the missing movements are marked as the missing values in the figure

at a time stamp. The unrealistic coordinates refer to GPS coordinates off the roads and GPS coordinates of the taxis that are stationary for more than 5 minutes. These coordinates are outliers in the regional movement data. The elimination of outliers leads to data missings. Besides, the GPS coordinates for some taxis are missing due to network problems. Therefore, it is impossible to calculate the speed and direction of a taxi at a specific time because the GPS coordinates are missing. For this reason, the regional movement data are missing for all variables except time. Some of the taxi routes are not known due to the missing data, as shown in Figure (c). Consequently, it is not easy to predict future regional transitions from the missing regional movement data.

4 Bayesian network

In this section, we present the Bayesian network model, which enables to predict future values from missing panel data. The Bayesian network is a probabilistic graphical model that uses Bayesian theory for the probability computations. The Bayesian network represents a set of variables and their conditional dependencies via a directed acyclic graph as illustrated in Fig. 3. The Bayesian theory is a statistical method, where the posterior probability as a consequence of two antecedents, a prior probability and a likelihood function, is derived for the observed data. The Bayesian model begins with the conditional probability theory. The probability that an event x and an event y occur simultaneously can be derived as $P(X \bigcap Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)$. This equation can be further derived as $P(X|Y) = (P(X) \cdot P(Y|X))/P(Y) \rightarrow P(H|E) = (P(H) \cdot P(E|H))/P(E)$ by summarizing the probability of the event x occurrence when the event y happens. Note that H is the hypothesis and E is the evidence. $P(H|E)$ is the posterior probability and $P(H)$ is the prior probability. $P(E|H)$ is the likelihood and $P(E)$ is the evidence probability. In order to obtain the posterior probability of the inferred hypothesis, we multiply the probability calculated by the observed data against the prior probability of the hypothesis predicted by subjective knowledge in advance and the probability of the evidence occurrence.

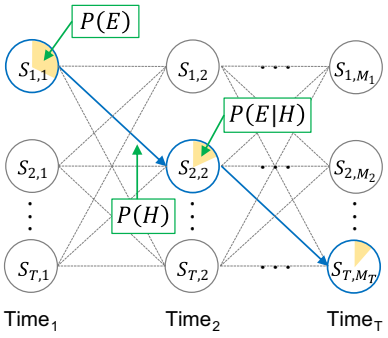


Fig. 3 Illustration of the Bayesian network model

We apply the Bayesian model to predict future values from missing data. Figure 3 illustrates how we build the Bayesian model from missing data. In the figure, S indicates a state with the probability of data included in the group, where $\sum_{k=1}^{M_t}(S_{t,k})$ is 1. Note that M_t is the number of states classified as similar groups in $Time_t$. The probability that data are classified as $S_{t,k}$ at $Time_t$ is set to $P(E) = P(S_{t,k})$. The hypothesis probability, $P(H)$, is set to the transition probability from $S_{t,k}$ to $S_{t+1,k}$. $P(E|H)$ is set to the probability that data belonging to $S_{t+1,k}$ have been included in $S_{t,k}$ in the past. The Bayesian model is automatically generated within our proposed system once the user sets model variables.

As mentioned in Sect. 3, We utilize three missing panel data, including physical growth data, air quality data, and regional movement data. In the physical growth data case, we build the Bayesian network model to predict body growth as follows. $P(E) = P(S_{t,k})$ indicates a state with the probability that a student is included in the growth group. M_t is the number of states classified as similar growth patterns in age_t. $P(H)$ is the probability that a student body growth trend transitions from S_{t,k_1} to S_{t+1,k_2}. $P(E|H)$ represents the probability that a student who is currently in S_{t+1,k_2} was included in S_{t,k_1} in the past. In the air quality data case, $P(E) = P(S_{t,k})$ indicates a state with the probability that stations are included in the air quality group. M_t is the number of states classified as similar air quality patterns in $time_t$. $P(H)$ is the probability that a air quality trend transitions from S_{t,k_1} to S_{t+1,k_2}. S_{t+1,k_2}. $P(E|H)$ represents the probability that a station currently in S_{t+1,k_2} was included in S_{t,k_1} in the past. In the regional movement data case, $P(E) = P(S_{t,k})$ is a probability that a taxis is included in a specific region k at certain time t. M_t is the number of regions in $time_t$. $P(H)$ is the probability that a taxi moves from S_{t,k_1} to S_{t+1,k_2}. $P(E|H)$ represents the probability that a taxis in S_{t+1,k_2} was included in S_{t,k_1} in the past.

5 Data imputation

In this section, we introduce four imputation models adopted in our system, including EM-spline, weighted moving average, mean, and multiple imputation. We also present an evaluation of imputing missing data.

5.1 Imputation models

EM-spline is an algorithm that determines the estimated value of a variable by repeatedly applying the expected value (E) and the maximization (M) [26]. To impute a missing value using EM-spline, we estimate the variables of the missing value by calculating the expected value of the log likelihood in step (E). Then, in step (M), we calculate the variable estimates that maximize the expected value. The value calculated in the maximize step is used as an estimate for the next expected step. The algorithm iteratively estimates the missing value through the step (E) and the step (M) until it reaches a convergence criterion.

The weighted moving average is applied to estimate a missing value in the panel data. When estimating the missing value using the observed values, the

weighted moving average technique assigns higher weights to the observed values near the missing while assigning lower weights to the observed values far from the missing. We use the weighted moving average function provided by imputeTS r-package [33].

Mean imputation is a technique in which the mean of the observed values replaces the missing value on a certain variable. The mean imputation is uncomplicated to use and maintains the size of the samples. Furthermore, the advantage of the mean imputation is that it does not change the sample mean of the variables. We employ the mean imputation function provided by imputeTS r-package [33].

Multiple imputation generates multiple complete datasets using multiple single imputations. The multiple imputation technique uses the mean and variance of the estimated values from multiple complete datasets, or estimates a missing value using Rubin rule [39]. We utilize the multiple imputation function provided by Amelia r-package [25].

We compare the means and standard deviations by four imputation models using the dataset introduced in Sect. 3.1. Table 1 presents the means and standard deviations of the height, weight, skeletal muscle, and protein for male students. Group 1 indicates 859 male students who are ten years old. We create Group 2 by eliminating about 15% of the records in Group 1 under MNAR condition. We observe that the mean height, weight, skeletal muscle, and protein of Group 2 are lower than Group 1. Also, the standard deviations of the other variables except for height in Group 2 are lower than ones in Group 1. We impute the missing values in Group 2 with four imputation models. In Table 1, the numbers in bold indicate the values close to the mean or standard deviation of Group 1. Overall, it appears that the missing values with multiple imputation and EM-spline are adequately imputed. However, multiple imputation and EM-spline produce high performance only for the data introduced in Sect. 3.1. Note that multiple imputation and EM-spline are not always the best for all of the missing datasets compared to other models.

Table 1 Comparison of the imputation models for the missing values

Group	Record	Height		Weight		Skeletal muscle		Protein	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Group 1	859	141.81	6.33	39.94	9.48	15.34	2.52	5.75	0.84
Group 2	731	139.68	6.79	38.28	9.06	14.71	2.43	5.54	0.80
EM-spline	859	141.92	5.98	40.24	**9.31**	**15.36**	2.42	**5.76**	0.80
Weighted moving average	859	141.56	5.83	39.91	8.88	15.26	2.33	5.72	0.77
Mean	859	141.41	5.76	15.22	2.21	15.22	2.21	5.70	0.73
Multiple imputation	859	**141.74**	**6.23**	**39.96**	9.29	15.30	**2.44**	**5.74**	**0.81**

Group 1 is the reference data, and Group 2 is the data after removing 15% of Group 1. Four imputation models are compared, including EM-spline, weighted moving average, mean, and multiple imputation

5.2 Evaluation of imputing missing data

Since the missing data cannot be observed, the best imputation model is not known. Nevertheless, probability distributions and variances can be used to verify the suitability of the imputation model roughly. If the variables follow a specific distribution, we can compare whether the probability distributions after the imputation are similar to the specific distribution. If the variables do not follow the specific distribution, we can choose another imputation model with a probability distribution that is similar to the probability distribution of the observed data. One way to compare probability distribution models is the Kolmogorov–Smirnov test [6].

The Kolmogorov–Smirnov test (K–S test) is a nonparametric test that compares the identity of two continuous probability distributions. If the missing type is missing completely at random (MCAR) mentioned in Sect. 2.1, the observed value V_{obs} is likely to follow the probability distribution of the variable. Moreover, since the data are completely randomly missing, the complete data $V_{obs} \cup V_{imp}$ generated by the imputation model are similar in the probability distribution to V_{obs} [35]. Therefore, the imputation model can be indirectly evaluated with the K–S test [36]. If the missing type is missing not at random (MNAR) or missing at random (MAR), we compare $V_{obs} \cup V_{imp}$ with a specific probability distribution model assuming that the variables follow a specific distribution and evaluate the performance of the imputation model.

6 Visual imputation system

In this section, we present the design principles necessary to impute incomplete panel data with the Bayesian network. We also introduce visualization and interaction techniques based on the design principles.

6.1 Design rationale

As mentioned in Sect. 1, the optimal solution for estimating missing values from incomplete data is not known. Nevertheless, since the loss of information due to missing values can affect the reliability of the analysis results, the data scientist needs to find a way to impute the missing values appropriately. Various imputation models have been studied, and the imputation impact on analysis results may be positive or negative depending on the imputation models. Therefore, to find the answer to the question "Did we impute the missing value with the appropriate value?," we set the following goals for the processes that impute missing values. (G1) Data scientists must be able to construct a Bayesian network to predict future trends in data. While modeling the Bayesian network, the data scientists need to select the parameters they want to predict and choose appropriate parameters to construct the states of the Bayesian network. (G2) The data scientist must be able to recognize the uncertainty due to the missing values and identify the missing parameters, ratios,

and missing mechanisms. Furthermore, the data scientist must be able to distinguish whether the missing data are related to a particular pattern within the data. (G3) The data scientist must be able to select a suitable imputation model to estimate the missing values. Unfortunately, since the missing values are not known, the optimal imputation solution is unknown. Nonetheless, the data scientist should compare the Bayesian network accuracy, Kolmogorov–Smirnov test, and probability distribution between the observed dataset and the estimated dataset to determine whether the missing values have been imputed appropriately.

6.2 System design

Figure 4 shows the analytical procedures of our proposed visual analysis system for the data imputation. The procedures involve (a) building a Bayesian network, (b) analyzing missing pattern in the Bayesian network, and (c) analyzing imputation models and Bayesian network prediction, based on the goals as mentioned in Sect. 6.1. The interactions between visualization modules supporting the procedures allow us to analyze missing patterns and choose an appropriate imputation model. Our system is a web-based application developed under the APM (Apache, PHP, Mysql) framework, and visual analysis modules are implemented using D3.js [10] and highcharts [23]. In the back end, the imputation models and the statistics calculations are performed with R.

Figure 5 presents the proposed visual analysis system for imputing missing data. The first step for imputing missings in the incomplete panel data is to build a Bayesian network.

The module in Fig. 5a provides datasets, parameters, and the number of states for the Bayesian network.

When we select a parameter for its future trend prediction, the system computes the correlation between the selected parameter and other parameters in the table, as shown in (a). The system visualizes the Pearson correlation coefficients in the bar chart by converting them into absolute values. Once picking

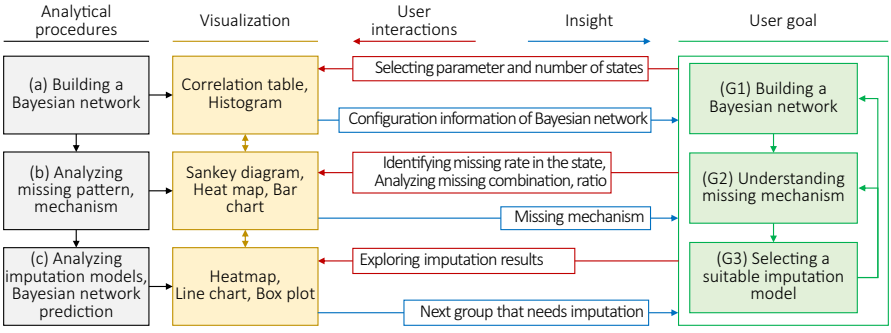


Fig. 4 Visual imputation system design. There are three analytical procedures with visualization techniques. For the imputation, the system provides various user interactions based on the analytical procedures. We can identify observed missings in the data while modeling a Bayesian network in the system, understand the missing mechanism, and then choose an appropriate imputation model

Fig. 5 Our visual analysis system for imputing missing values in incomplete panel data. The system consists of various visualization modules, including **a** Bayesian network modeling, **b** missing ratio identification within the Bayesian network, **c** information of the selected state in the Bayesian network, **d** missing pattern analysis, **e** error rates of imputation models, **f** K–S test matrix, **g** probability distribution, **h** imputation history, **i** precision chart, **j** confusion matrix, and **k** prediction results in the Bayesian network

the predictor parameters for the state configurations in the Bayesian network, we determine the number of states. Then, our system automatically creates the Bayesian network model with the selected parameters and the number of states.

The second step is to analyze the missing patterns in the Bayesian network. We visualize the Bayesian network using the Sankey diagram, as shown in Fig. 5b. Note that the x-axis represents the time in the Sankey diagram, and each node indicates a state. The size of each node is encoded with the number of data records contained in the state. Each state is assigned a unique identifier. The color of each state implies the missing rate in the data. The line thickness represents the transition probability between states. In this module, we can recognize the transition probability between states and discover which state has more missing data. When we select a state, the system provides the data statistics of the Bayesian network state in Fig. 5c and the missing patterns in (d). The table in (c) presents the selected state information, including the number of records the ratio of the observed records and the missing and imputed records. The aggregation plot proposed by Kowarik et al. [28] in (d) presents the missing patterns of the selected

state in (b). The purpose of the aggregation plot is to identify the missing variable combinations used to configure the state in the next step.

The missing combination is an identifier checking whether there are missings from the variable point of view. The missing combination extracts records only with missing values from the data and then counts the number of missing values for each variable. The aggregation plot consists of a heat map table in (d1), and a bar chart in (d2). The heat map table in (d1) visualizes the variables corresponding to missing combinations and gives us the relative missing ratio for each combination. The relative missing rate is calculated using the min–max normalization of the missing rate for each combination. The missing percentage for each combination is computed by dividing the number of records with missings by the total number of records. In the heat map table, the variables without any missing are colored in gray, and the variables with missings are in blue. The variables containing imputed data are in yellow. The bar chart in (d2) shows the relative missing and imputed ratios of the variables. Note that we calculate the relative ratios as dividing the number of combinations and variables.

The third analysis procedure involves selecting an imputation model, which allows us to estimate missing data. In this procedure, the system utilizes the error rates of imputation models in Fig. 5e, Kolmogorov–Smirnov test in (f), probability distributions in (g), imputation model history in (h), precision chart in (i), confusion matrix in (j), and prediction comparison of the Bayesian network in (k). The system automatically estimates the missing data with the number of iterations. To find a suitable imputation model, we need to evaluate the imputation models and precision of the Bayesian network The box plot in (e) shows the error rates of the imputation models. We cannot choose an optimal imputation model because no one knows the actual record of the missing data. Therefore, we calculate the indirect performance of each imputation model using the following procedure. The system generates missing data under the MCAR condition from the observed dataset as many as the number of iterations. With the missing data, the system estimates the missing data utilizing the four imputation models presented in Sect. 5.1. The system then calculates the performance of each imputation model using MAPE (mean absolute percentage error) between the estimated and observed data. Figure 5f presents the K–S tests introduced in Sect. 5.2 as the heat map table with the observed data V_{obs} and complete data $V_{obs} \cup V_{imp}$. The x-axis in the heat map table represents the variables, and the y-axis represents the imputation models. The color of each cell indicates whether the two distributions are similar. If p value > 0.05, the two distributions are similar to each other, and the cell is displayed in dark grey. On the other hand, if p value ≤ 0.05, the two distributions are different, and the cell is represented in white. (g) renders the probability distributions of the observed data and imputed data in the cell selected in (f). Figure 5h shows the history of imputation models and states examined within the system. This module provides the history of the imputation models applied to the state.

To select a suitable imputation model, we need to analyze the precision of the Bayesian network with the performance of the imputation model. The precision means the probability that the Bayesian network designed with the imputed data classifies the data records into correct states. Figure 5i shows the precision of the

Bayesian network for each imputation model. (j) is the confusion matrix of the imputation model selected in (i). The confusion matrix in (j) presents the Bayesian network performance on the state, as displayed in (c). The number of correct and incorrect predictions are summarized. The gray cell in (j) indicates the number of correctly predicted data records, and the white cell means the number of incorrectly predicted data records in the state. The precision of the imputation model in the confusion matrix is calculated as follows. Precision = (correctly predicted states) / (number of data records in the state).

The line chart in (k) presents the prediction results of the Bayesian network on the state, as shown in (c). In the line chart, the x-axis denotes the time, and the y-axis means the variable value that the Bayesian network model predicts. The lines represent the predictions with the observed data and the imputed data. The line chart allows us to inspect how the imputation model changes the prediction results of the Bayesian network.

7 Case studies

In this section, we present case studies that impute missing values while modeling the Bayesian networks using our visual analysis system. We also show the predictions of missing panel data through the Bayesian networks.

7.1 Visual imputation system for Physical growth data

We introduce the process of imputing missing values while designing a predictive model based on the Bayesian networks with the physical growth data. As shown in Fig. 5b, the missing ratio is highest in the State A8S3. The State A8S3 contains the records for 177 students with 32% missing ratio displayed in (c). Note that the 32% missing ratio indicates the sum of *Missing ratio*, 20%, and *Imputed ratio*, 12%. The State A8S3 has eight missing combinations, as seen in (d). The bar charts in (d1) indicate the missing ratios of the eight combinations. The bar charts in (d2) show the missing ratios of the selected variables in (a). We apply the imputation model, starting from the combination with the lowest missing ratio. The low missing ratio indicates that there are fewer data records with missings. The fewer data records to be imputed at one time, the less likely there is a bias in imputing the missing values. The combination C6, C7, and C8 marked in yellow indicate that the imputations are completed. The imputation models applied to the combination C6, C7, and C8 are presented in (h).

We select one of the imputation models for Combination C5 through the modules in Fig. 5e–k. Our system automatically imputes missing values for Combination C5 with four imputation models. For the performance comparison of the imputation models, the system provides MAPE (mean absolute percentage error) and K–S test. Figure 5e shows MAPEs of the imputation models. In the four imputation models, the means of the MAPEs are similar. However, the multiple imputation (MI) and weighted moving average (WMA) models have less variation

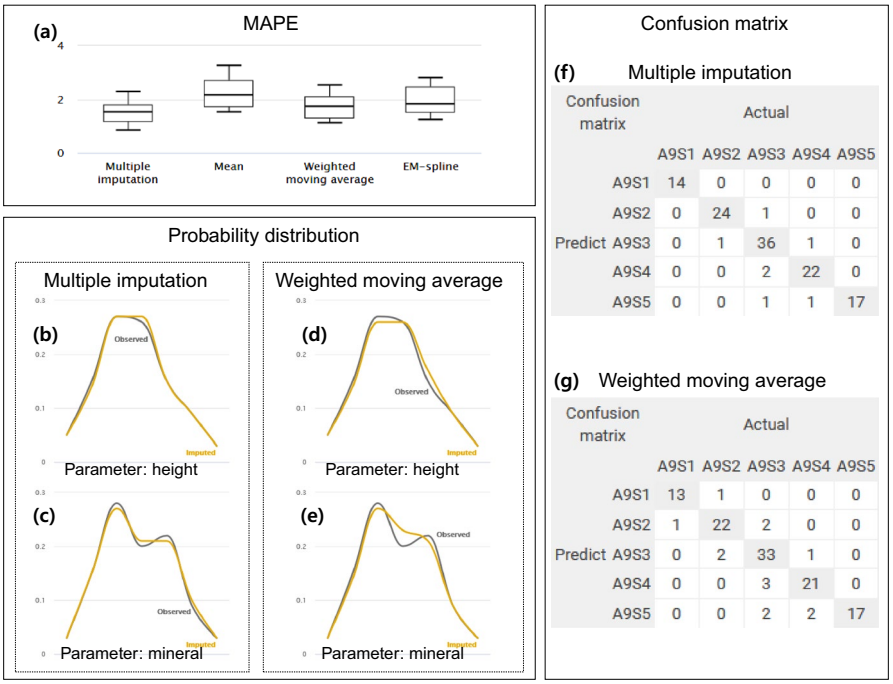
in the expected MAPE values than the other two models. Therefore, from the box plots, we realize that the MI and WMA are suitable imputation models for Combination C5. Figure 5f presents the K–S tests according to the variables and imputation models. Figure 5g represents the probability distribution of the variable in the cell that we select in (f). In Figure 5f, the white cells indicate that the probability distributions of the complete dataset are different from the observed dataset. The different probability distributions mean that the characteristics of data have been changed due to the imputation model. In this case, the datasets imputed by the WMA and mean imputation have one variable, *weight*, whose distribution is different from the observed data. This indicates that the imputed *weight* may not have the same characteristics as found in the observed dataset.

The data characteristic change means that the imputation model introduces a bias while imputing the missing values. Therefore, the WMA and mean imputation models in (f) are not suitable for Combination C5. The K–S test tells that the MI and EM-spline models are suitable for Combination C5. However, the model with the low MAPE and no change of the data characteristics is MI, which is found in (e) and (f), respectively.

We evaluate the performance of the model indirectly through the difference between the estimated and actual data by the imputation model in (e)–(g). We compare the imputation model in terms of the prediction with the Bayesian network through (i)–(k). The system provides the precision in (i), confusion matrix in (j), and prediction results on the Bayesian network in (k). We can see that most of the data are accurately predicted.

Figure 5k shows the *height* prediction of the Bayesian network built on a complete dataset by the MI model. We can see that the prediction of the Bayesian network has changed after the MI model imputes the missing values. Therefore, it is desirable to choose the MI as an imputation model suitable for Combination C5.

Following the imputation for Combination C5, we impute the missing values in Combination C4. As shown in Fig. 5(d1), Combination C4 of State A8S3 has missing values in *height* and *mineral*. Figure 6 shows the performance of imputation models for Combination C4. In (a), we can recognize that the MAPE means are similar for the multiple imputation (MI), weighted moving average (WMA), and EM-spline. However, the interquartile range of MAPE for EM-spline is greater than that of the MI and WMA. Therefore, the candidate imputation models suitable for Combination C4 are the MI and WMA. Figures (b)–(e) show the probability distributions of imputed and observed variables. (b) and (c) present the probability distributions of the variable *height* and *mineral* imputed by the MI. (d) and (e) display the probability distributions of the variables *height* and *mineral* imputed by the WMA. In (b), (c), and (d), we observe that the two probability distributions are similar before and after the imputation. However, we recognize that the two probability distributions for *mineral* are different in (e). Hence, we infer that the MI is better than the WMA. We analyze the performance of the Bayesian network designed on the complete dataset imputed by the imputation model in (f) and (g). (f) and (g) present the confusion matrices for the MI and the WMA, respectively. We see that most states are correctly predicted in (f), while there are many mispredicted states in (g). Consequently, we conclude that



(f) Multiple imputation

Confusion matrix	Actual				
	A9S1	A9S2	A9S3	A9S4	A9S5
A9S1	14	0	0	0	0
A9S2	0	24	1	0	0
Predict A9S3	0	1	36	1	0
A9S4	0	0	2	22	0
A9S5	0	0	1	1	17

(g) Weighted moving average

Confusion matrix	Actual				
	A9S1	A9S2	A9S3	A9S4	A9S5
A9S1	13	1	0	0	0
A9S2	1	22	2	0	0
Predict A9S3	0	2	33	1	0
A9S4	0	0	3	21	0
A9S5	0	0	2	2	17

Fig. 6 **a** The error rates of imputation models applicable to Combination C4 are presented. **b**–**e** The probability density distributions before and after the imputations are compared. **b** and **c** utilize the multiple imputation (MI), and **d** and **e** use the weighted moving average (WMA). **f** and **g** show the confusion matrices of the Bayesian network predictions with the imputed datasets

the prediction of the Bayesian network using the data imputed by the MI becomes more reliable.

7.2 Visual imputation system for Air quality data

We design a Bayesian network that can predict the future trend of *PM10* from 9:00 to 14:00 on an hourly basis. In Fig. 7a, each state in the Bayesian network is composed of variables including *PM2.5, PM10, humidity,* and *temperature*, and there are eight states every hour. We see many missings occurring in State T9S6, T11S4, T13S2, and T13S8. Among them, we want to impute the missings in State T11S4. As shown in (c), State T11S4 contains 54 data records, which are the measure stations. From the heat map table in (d1), we realize that State T11S4 has eight missing combinations. Many missings are found in *temperature* among the variables constituting the state as shown in (d2).

Moreover, most missings are observed in Combination C1 consisting of *humidity* and *temperature*. The fewer data record to impute at one time, the less likely there is a bias in estimating the missing values. We, therefore, start to apply the imputation models to Combination C5–C8, where the missing ratios are relatively low among

Fig. 7 Visual imputation system for the air quality data. The system includes **a** Bayesian network modeling, **b** missing ratio identification within the Bayesian network, **c** information of the selected state in the Bayesian network, **d** missing pattern analysis, **e** error rates of imputation models, **f** K–S test matrix, **g** probability distribution, **h** imputation history, **i** precision chart, **j** confusion matrix, and **k** prediction results in the Bayesian network

eight combinations. Then, we impute the missings in Combination C4 consisting of *PM2.5, PM10,* and *humidity*.

As presented in Fig. 7e, the MAPE means of the four imputation models are similar. For the multiple imputation (MI), the interquartile range is narrow. On the other hand, for the other three models, the interquartile ranges are wide, which implies that the MI imputes the missings within a relatively constant margin of error than other models. Therefore, from the box plots, the MI is most likely the best model. However, from the K–S test, the probability distribution of the variable *humidity* changes after the imputation with the MI since the cell is colored white in (f). The probability distributions before and after the imputation are observed in (g). Moreover, the mean and EM-spline models alter the probability distributions for *humidity* and *PM2.5*, respectively, after the imputations, which is observed in (f) as the cells are colored white. The weighted moving average (WMA) has a slightly higher MAPE than the MI, but the WMA does not alter the probability distribution shape during the imputation. The precision of WMA is slightly lower than that of EM-spline, as seen in (i). Nevertheless, the confusion matrix in (j) tells that the WMA more accurately predicts data records. We also recognize that the prediction for State T11S4 is modified after the WMA imputes the missing values in (k). Therefore, we presume that the prediction of the Bayesian network using the data imputed by the WMA becomes more reliable.

7.3 Visual imputation system for regional movement data

In the physical growth data and air quality data, the records of some variables, not all, are missing. On the other hand, in the Beijing taxi data, the records of all variables are missing in a specific time range due to the network problem. Therefore, the regional movement data generated from the Beijing taxi data are missing some movement sequences, as shown in Fig. 2. The four imputation models introduce in Sect. 5 and impute the values of the missing variable by referring to the variable observed in the record where the missings occur. If all variables are missing at a specific time, the missing values can be imputed through the Bayesian network. When a region sequence of a taxi is missing at time t, we search for a state including the taxi at time $t-1$. Then, we impute the missing sequence of the taxi according to the biggest transition probability toward the next state.

Figure 8 shows our system for imputing the missing regions in the regional movement data. Note that Fig. 8 is a part of our system focusing on the all-variable missing problem. As introduced in Sect. 4, we model a Bayesian network using the observed regional movement data. The Bayesian network in (a) consists of 20 states at 10-min intervals for an hour. Note that T stands for time, and S stands for State. For example, *T3S11* indicates State 11 at time T3. The color of each state represents the ratio of taxis with missing region sequences. In (a), we select State T3S11, and the system imputes the state after State T3S11 in (b). We pick the state with the

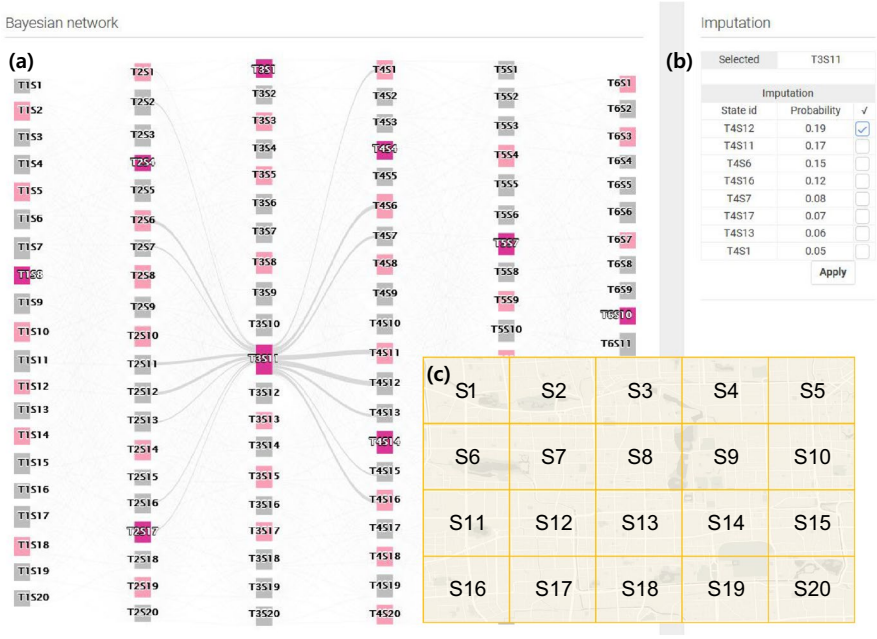


Fig. 8 Visual imputation system for the regional movement data. **a** Bayesian network in which traffic flows of 20 regions are modeled at 10-min intervals for 1 h. **b** Bayesian network-based data imputation module. **c** Regional map in Beijing

biggest transition probability of State T3S11 in (b). The system imputes the region that is missing after State T3S11 with the state picked in (b). In this way, we impute all missing regions of the data when all variables are missing

8 Evaluation

In this section, we evaluate the accuracy of the Bayesian network that our system automatically build. We utilize RMSE (root mean square error) and MAPE (mean absolute percentage error) as follows. RMSE $= \sqrt{\frac{\sum_{j=1}^{n}(f_j - o_j)^2)}{n}}$ and MAPE $= \frac{100}{n}\sum_{j=1}^{n}|\frac{o_j - f_j}{o_j}|$, where f_j is the predicted value and o_j is the actual observation value. Because the RMSE is a measure of the difference between the predicted value and observed value, it is suitable to evaluate the sensitivity of the prediction model. However, since this is sensitive to the outliers, another error measurement must be evaluated to compensate this disadvantage, which is the MAPE in this work. The MAPE provides the percentage of the state error for the actual value, which is suitable for the relative accuracy of the multiple prediction models.

Table 2 shows the performance of the imputation models according to the missing ratios for three different datasets. We have generated missing data with the missing completely at random (MCAR) condition from the datasets to evaluate the imputation model performance. We have estimated the missing values using four imputation models and then calculated the MAPE and RMSE. As seen in Table 2, in all four imputation models, the higher the missing ratio, the higher the error rate. In all datasets, we can see that the multiple imputation has the lowest error rate. The physical growth data and regional movement data are estimated accurately with the imputation models provided in the system. On the other hand, the errors with all imputation models are relatively higher in the air quality data.

Table 2 Accuracy comparisons of imputation models according to the missing rates with three datasets

Imputation model	Metrics	Missing rate								
		Physical growth data			Regional movement data			Air quality data		
		10%	20%	30%	10%	20%	30%	10%	20%	30%
Multiple imputation	MAPE	0.43	0.93	1.32	1.07	1.79	2.99	1.85	4.17	6.10
	RMSE	2.42	3.82	4.41	1.43	2.37	3.21	3.43	5.38	6.68
Weighted MA	MAPE	0.87	1.71	2.5	1.18	2.04	2.88	1.94	4.04	6.35
	RMSE	4.80	6.46	7.92	1.82	2.81	3.49	3.45	5.61	6.98
Mean	MAPE	0.92	1.71	2.76	1.63	2.65	3.53	3.31	5.84	8.37
	RMSE	4.89	6.52	8.46	2.29	2.92	3.63	6.12	7.18	9.20
EM-splinen	MAPE	0.46	1.11	1.67	0.96	2.16	3.76	2.69	4.29	7.63
	RMSE	2.80	4.48	5.91	1.54	2.63	4.04	5.29	6.43	8.13

Table 3 shows the prediction performance of the Bayesian network model for three datasets. We have created missing data corresponding to 3 missing types using MICE R package [11] and modeled the Bayesian network by dividing the datasets into two cases. In Case 1, we have discarded the missing values from the data. In Case 2, we have used the complete data including imputed values by the multiple imputation. If the data are missing, the error rate in the prediction after imputing the missing values is lower than that in the prediction after removing the missing values. Overall we observe that the error rate is high in the MCAR missing type. In MCAR and MNAR, we notice that the MAPE increases linearly as the—missing ratio increases. At 10% and 20% missing ratios, the MAR has the lowest MAPE, but when the missing ratio is 30%, it is observed that the MAPE of MAR increases sharply. The MAR is the missing type in which the missing of a particular variable relates to other variables within the data. Therefore, the higher the missing ratio of the MAR, the more likely it is that the probability distribution of variables alters after imputing the data [27, 31]. Hence, when the data missing pattern is the MAR, we need to analyze not only the K–S test but also the precision of the Bayesian network when selecting a suitable imputation model.

9 Discussion

When designing a predictive model with incomplete panel data, analysts need to minimize the uncertainty of missing data on the predicted result. A common approach is to understand the data characteristics and then construct a complete dataset using a suitable imputation model. The predictive model then estimates future trends based on the complete dataset. If the predictive model performs poorly, analysts need to redesign the predictive model or re-estimate missing data. In such cases, it is difficult to determine whether the cause of the predictive model performance degradation is incomplete data or a poorly designed predictive model. Our proposed system visualizes the transition probabilities of states over time, the missing ratio, and the

Table 3 Accuracy comparison of Bayesian network models according to the missing ratios for three datasets. In the table, Case 1 denotes the Bayesian network modeled with the data, excluding the missing values. Case 2 indicates the Bayesian network modeled with the data after the imputation using the multiple imputation

Missing type	Evaluation group	Physical growth data			Air quality data			Regional movement data		
		10%	20%	30%	10%	20%	30%	10%	20%	30%
MCAR	Case 1	2.17	3.48	4.00	2.42	3.91	4.93	2.15	3.46	4.55
	Case 2	1.09	2.79	2.91	1.33	1.66	3.12	1.75	2.34	2.86
MNAR	Case 1	2.65	2.94	4.08	2.55	2.91	3.95	2.15	2.81	4.15
	Case 2	1.24	2.04	3.74	1.7	2.03	2.39	1.78	1.93	2.47
MAR	Case 1	1.68	3.17	6.19	1.17	2.34	6.04	1.67	2.12	5.89
	Case 2	0.64	0.96	3.71	0.94	1.97	5.49	1.16	2.02	4.90

missing pattern in Bayesian networks. Also, the system visualizes changes in statistical characteristics of the data and changes in values predicted by the model during the missing value imputation. The proposed system allows analysts to improve predictive model performance on various incomplete panel data. Besides, the system provides multiple visualizations and visual interactions to help them examine the relationship between incomplete data and the predicted results of the model. We believe that we minimize the uncertainty in the prediction result caused by missing values through a feedback process such as Fig. 4. Nevertheless, there are some limitations in our proposed analysis process and visual analysis system. The proposed system improves predictive model performance by recursively repeating processes such as identifying missing patterns, selecting imputation models, and comparing prediction results. In this process, the system does not tell definitely how long the analysis process should be repeated. The system considers the process completion condition as the case where the missing values are accurately estimated, and the prediction error of the Bayesian network model is minimum. Since no one knows the actual value of the missing value, the proposed system estimates the missing values under the assumption that it follows the statistical characteristics of the observed sample. Suppose missings occur in a panel that does not follow the statistical characteristics of the sample. In that case, since the panel is not recognized as a new pattern, the panel is treated as the average pattern of the observed sample. As more panels correspond to the sample mean, the Bayesian network model is more likely to overfit. We will investigate how the visual analytics system can handle the overfitting of the model while imputing missing values to improve the efficiency of the analysis process.

We have applied the Bayesian network as a model to predict future trends in missing panel data. The Bayesian network predicts which groups a panel will be classified into at a specific time. The system is designed to predict future trends for a particular panel from a classification point of view. Therefore, the system can apply other models such as deep neural network (DNN) beside of Bayesian network. However, it is not easy to apply time series prediction models such as recurrent neural network (RNN) or long short-term memory (LSTM) to our current system. We plan to examine how to handle missing values in time series prediction models in the future.

10 Conclusion

In this paper, we have introduced a visual analysis system that enables us to estimate missing data from incomplete panel data in the Bayesian network. The system supports the visualizations to identify the missing ratio of each state in the Bayesian network and analyze the missing patterns of parameters. The system has also provided imputation algorithms for missing data estimation based on missing patterns. The system also provides accuracy and Kolmogorov–Smirnov tests for a performance comparison of imputation algorithms. Our visual analysis system supports analysis procedures such as missing detection, exploration of missing patterns, imputation

model, and its outcome comparison. We have presented the effectiveness of the proposed visual analysis system through use cases.

Our proposed system allows us to identify states in the Bayesian network that require the imputation. Our system has the advantage of intuitively identifying missing patterns, but we did not perform the impact analysis of missing data on the Bayesian network. Missing data can increase uncertainty in certain parts of the Bayesian network. The missing data imputation might affect and change the Bayesian networks, whereas certain parts of the Bayesian network may not be affected by missing data. We will work on techniques to quantify the impact of missing data on the Bayesian network in the future. We will also investigate machine learning techniques for the missing data imputation and an automatic imputation of various missing patterns in the future.

Acknowledgements This work was supported in part by the Basic Research Program through the National Research Foundation of Korea (NRF) funded by the MSIT under Grant 2019R1A4A1021702, and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00242, Development of a Big Data Augmented Analysis Profiling Platform for Maximizing Reliability and Utilization of Big Data).

References

1. Alemzadeh S, Niemann U, Ittermann T, Völzke H, Schneider D, Spiliopoulou M, Bühler K, Preim B (2020) Visual analysis of missing values in longitudinal cohort study data. In: Computer Graphics Forum. Wiley Online Library, vol 39, pp 63–75
2. Allison PD (2001) Missing data: Sage university papers series on quantitative applications in the social sciences (07–136). Thousand Oaks, CA
3. Andridge RR, Little RJ (2010) A review of hot deck imputation for survey non-response. Int Stat Rev 78(1):40–64
4. Antony U, George H, Heike H, Bernd S (1996) Interactive graphics for data sets with missing values-manet. J Comput Graph Stat 5(2):113–122
5. Arbesser C, Spechtenhauser F, Mühlbacher T, Piringer H (2017) Visplause: visual data quality assessment of many time series using plausibility checks. IEEE Trans Visual Comput Graph 23(1):641–650
6. Arnold TB, Emerson JW (2011) Nonparametric goodness-of-fit tests for discrete null distributions. R J 3(2)
7. Babad YM, Hoffer JA (1984) Even no data has a value. Commun ACM 27(8):748–756
8. Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48(1):5–37
9. Berglund P, Heeringa SG (2014) Multiple imputation of missing data using SAS. SAS Institute, Cary
10. Bostock M, Ogievetsky V, Heer J (2011) D^3 data-driven documents. IEEE Trans Visual Comput Graph 17(12):2301–2309
11. Buuren Sv, Groothuis-Oudshoorn K (2010) mice: Multivariate imputation by chained equations in r. J Stat Software 1–68
12. Yy C (2010) Multiple imputation for missing data: Concepts and new development (version 9.0). SAS Institute Inc, Rockville, MD 49:1–11
13. Carpenter J, Kenward M (2012) Multiple imputation and its application. Wiley, Boca Raton
14. Cheng X, Cook D, Hofmann H et al (2015) Visually exploring missing values in multivariable data using a graphical user interface. J Stat Softw 68(1):1–23
15. Dingen D, van't Veer M, Houthuizen P, Mestrom EH, Korsten EH, Bouwman AR, Van Wijk J (2019) Regressionexplorer: interactive exploration of logistic regression models with subgroup analysis. IEEE Trans Visual Comput Graph 25(1):246–255
16. Eaton C, Plaisant C, Drizd T (2005) Visualizing missing data: Graph interpretation user study. In: IFIP Conference on Human–Computer Interaction. Springer, pp 861–872
17. Enders CK (2010) Applied missing data analysis. Guilford press

18. Enders CK, Gottschall AC (2011) Multiple imputation strategies for multiple group structural equation models. Struct Equ Model 18(1):35–54
19. Fernstad SJ (2018) To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. Inform Visual. 1473871618785387
20. Fernstad SJ, Glen RC (2014) Visual analysis of missing data–to see what isn't there. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp 249–250. IEEE
21. Graham JW (2009) Missing data analysis: making it work in the real world. Annu Rev Psychol 60:549–576
22. Heike MT, Hofmann H, Siegl B, Unwin A (1997) Manet extensions to interactive statistical graphics for missing values. In: In New Techniques and Technologies for Statistics II. Citeseer
23. Highcharts: highcharts. https://www.highcharts.com/. Accessed 10 June 2019
24. Honaker J, King G (2010) What to do about missing values in time-series cross-section data. Am J Polit Sci 54(2):561–581
25. Honaker J, King G, Blackwell M (2011) Amelia II: A program for missing data. J Stat Software 45(7):1–47
26. Johnson R, Wichern D (2002) Applied multivariate statistical analysis, 5th edn. Prentice Hall, Upper Saddle River
27. Kang H (2013) The prevention and handling of the missing data. Korean J Anesthesiol 64(5):402
28. Kowarik A, Templ M (2016) Imputation with the r package vim. J Stat Softw 74(7):1–16
29. Krause J, Perer A, Bertini E (2014) Infuse: interactive feature selection for predictive modeling of high dimensional data. IEEE Trans Visual Comput Graph 20(12):1614–1623
30. Little RJ, Rubin DB (2014) Statistical analysis with missing data, vol 333. Wiley, Boca Raton
31. Marshall A, Altman DG, Royston P, Holder RL (2010) Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol 10(1):7
32. McKnight PE, McKnight KM, Sidani S, Figueredo AJ (2007) Missing data: a gentle introduction. Guilford Press
33. Moritz S, Bartz-Beielstein T (2017) imputets: time series missing value imputation in r. R J 9(1):207–218
34. Mühlbacher T, Piringer H (2013) A partition-based framework for building and validating regression models. IEEE Trans Visual Comput Graph 19(12):1962–1971
35. Nguyen CD, Carlin JB, Lee KJ (2013) Diagnosing problems with imputation models using the Kolmogorov–Smirnov test: a simulation study. BMC Med Res Methodol 13(1):144
36. Nguyen CD, Carlin JB, Lee KJ (2017) Model checking in multiple imputation: an overview and case study. Emerg Themes Epidemiol 14(1):8
37. Osborne JW (2013) Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data. Sage
38. Pi M, Yeon H, Son H, Jang Y (2019) Visual cause analytics for traffic congestion. IEEE Trans Visual Comput Graph
39. Rubin DB (2004) Multiple imputation for nonresponse in surveys, vol 81. Wiley, Boca Raton
40. Schulz HJ, Nocke T, Heitzler M, Schumann H (2017) A systematic view on data descriptors for the visual analysis of tabular data. Inform Visual 16(3):232–256
41. Song H, Szafir DA (2018) Where's my data? evaluating visualizations with missing data. IEEE Trans Visual Comput Graph
42. Stuart EA, Azur M, Frangakis C, Leaf P (2009) Multiple imputation with large data sets: a case study of the children's mental health initiative. Am J Epidemiol 169(9):1133–1139
43. Swayne DF, Buja A (1998) Missing data in interactive high-dimensional data visualization. Comput Stat 13(1):15–26
44. Templ M, Alfons A, Filzmoser P (2012) Exploring incomplete data using visualization techniques. Adv Data Anal Classif 6(1):29–47
45. Templ M, Filzmoser P (2008) Visualization of missing values using the r-package vim. Reserach report cs-2008-1,Department of Statistics and Probability Therory,Vienna University of Technology
46. Yeon H, Son H, Jang Y (2021) Visual performance improvement analytics of predictive model for unbalanced panel data. J Visual 1–14