Accuracy Comparison of Data Imputation Estimation Methods Between Partial Least Squares of Structural Equation Modeling and K-Nearest Neighbor

Narong Phothi [1] Somchai Prakancharoen [2]
[1]Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangsue, Bangkok, Thailand 10800, E-mail: narong@sueksa.go.th
[2]King Mongkut's University of Technology North Bangkok
Bangsue, Bangkok, Thailand 10800, E-mail: spk@kmutnb.ac.th

Abstract— **This study aimed to the accuracy comparison of data imputation estimation methods between partial least squares of structural equation modeling (PLS of SEM) and k-nearest neighbor (K-NN). The measurement accuracy of the model based on the mean magnitude of relative error (MMRE). Model development using data from the online database UCI data set waveform database generator. Indicators 21 (1,200 sets) methods were as follows: 1) Data set was divided into 2 groups (experimental group of 1,000 sets and test group of 200 sets). 2) The experimental group was analyzed by three main factors. 3) PLS of SEM method: Created a SEM with three main factors, then the remaining factors to created new the relationships with PLS method and created new SEM. The test data was substituted in the equation to find the MMRE which was 36.90% (accuracy was 63.10%). 4) K-NN method: Selected the main factor was the relationship of the missing data. Measure the euclidean distance between test group and experimental group and selected 5 (K=5) of data sets were nearest to the missing data for estimate by mean. The MMRE which was 61.52% (accuracy was 38.48%). Thus, comparing estimates of missing data showed that using the PLS of SEM method were more accuracy about 24.62% and MMRE declined than K-NN method.**

Keywords— **Data Imputation Estimation, Partial Least Squares of Structural Equation Modeling, K-Nearest Neighbor**

I. INTRODUCTION

A. Background

General research information in this area is required to complete the analysis in order to achieve the most accurate and precise results. However, some data may be missing or incomplete. Therefore, in order to bring a data set that is complete and ready to use, some data will be missing. This would result in the records becoming redundant or obsolete, thus analysis and forecast of data is needed. If some of the data set were missing in large amounts, data that is needed should avoid deviation which would lead to error of the results and need to be obtained through processing. The estimation of missing data will help in preparing data to replace the missing research data sets. From the research on estimation of missind data, such as researcher Prakancharoen [1] using structural equation modeling to estimate the time to develop application software oriented network, also researchers Phothi and Prakancharoen [2] using structural equation modeling between with discriminant analysis and without discriminant analysis for accuracy comparison of imputation methods, also researcher Rufus [3] using solutions for missing data in structural equation modeling for new data all use similar approaches to solve this issue, also researchers Meesad and Hengpraprohm [4] using combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data and Thomas [5] using K-NN algorithm for prediction and classification data.

In this study, researchers compared the accuracy of data imputation estimation methods between partial least squares of structural equation modeling and K-nearest neighbor. The research information is taken from the online database, UCI machine learning repository is a collection of waveform database generator data sets (1,200 sets). The measurement accuracy of the estimated missing data from the mean magnitude of relative error was found to be highly accurate.

B. The purpose of the research.

- To estimate missing data by using partial least squares of structural equation modeling.
- To estimate missing data by using k-nearest neighbor.
- To compare the accuracy of estimates the missing data between partial least squares of structural equation modeling and k-nearest neighbor.

C. Scope of research

- The data used in this operation was a waveform database generator data set from the online database, and UCI machine learning repository as a data type with 1,200 sets which were divided into 2 groups: experimental group (1000 sets) and test group (200 sets)
- The data set has 21 indicators, namely, V1-V21 and C1 classes for the description of each indicator V. Researchers Leo and Etal [6] which can be viewed at http://archive.ics.uci.edu/ml/datasets.html determined

that the fifth indicators (V5) in the equation of the test group was missing valuable data used to compare the accuracy of the estimation method. Missing value due to a measure of this needs to find the best relationship associated with other indicators in a waveform database generator data set and K is set equal to 5.

II. THEORY AND METHODOLOGY

A. Factor Analysis

Factor analysis (FA) [7] is a technique used to extract the factors (component) from a group of indicators that are related to each factor. This will be used instead of a group of indicators that have the same group. This is a technique that reduces the number of dimensions or manifest variable and considers the suitability of the extracted factors. By checking the statistics Kaiser-Meyer-Olkin: KMO (KMO > 0.60) factors obtained will only validate the considered values. Able to explain the variability of all the factors together (total variance explained) with the inverse of each variable with no apparent extraction factor would greatly benefit this approach. If the value of a high percentage (cumulative explained variance) showed that the factors can represent a good indicator, this can be formulated as follows

$$F_j = w_{j1}x_1 + w_{j2}x_2 + \ldots + w_{jp}x_p + e \quad (1)$$

where F =factor, w =coefficient of variable x, x =manifest variable and e =margin of error.

B. Structural Equation Modeling

Structural equation modeling (SEM) [1], [8] is a technique used to analyze the relationship of factors from the survey (exploratory) with a key and then extract a model of the relationship of various factors which is the main theory or hypothesis of this study. From the statistics of 1) Chi-square (χ^2) should be a non-significance (P>0.05) 2) Goodness of Fit Index (GFI>0.90) 3) Root Mean Square Error of Approximation (RMSEA<0.06) and 4) Hoelter's N, the value (Hoelter's. N>75) is used to check the adequacy and appropriateness of sample size (case) in structural equation model.

C. Partial Least Squares of Structural Equation Modeling

Partial least squares of structural equation modeling (PLS of SEM) [9], [10] is a technique used to estimate the stability of the equation appears in the relationship between variables. The equation is made up of indicators that are related. To creates [11],[12] a new factor X_iZ_j by bringing a measure of the factor X_i by match multiplied with all indicators of factor Z_j and repeated until completed, formulated as follows

$$X_iZ_j = \lambda_i\lambda_j\xi_1\xi_2 + \lambda_i\xi_1\delta_j + \lambda_j\xi_2\delta_i + \delta_i\delta_j \quad (2)$$

where X_i= predictor of variable, Z_j= moderator of variable, ξ= factor, λ= factor loading and δ= margin of error.

D. K-Nearest Neighbor

K-nearest neighbor (K-NN) [4],[13] is a technique for classification data and estimation missing data. Using a indicators of all experimental groups that are associated with a set of missing data in the test group as a model. The researcher must be define value for use in the K to the nearest whole number, which must be positive. The step of K-NN imputation are as follows:

Step 1: Define value for use in the K that are most similar to the missing data. (should be always an odd number).

Step 2: Measure the distance by Euclidean distance between test data and experimental data according to the formula.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \quad (3)$$

where $d(x_i, x_j)$ = Euclidean distance between x_i and x_j, n = number of all indicator or features of experimental group, $x_{i,k}$ = value of indicator or features x_i in order k of test group, $x_{j,k}$ = value of indicator or features x_j in order k of experimental group.

Step 3: Sort distance based on nearest to K of experimental group.

Step 4: To estimate missing data from the mean of the data sets to the nearest number of K defined by formula.

$$\hat{x}_{i,j} = \frac{\sum_{k=1}^{k} x_k}{K} \quad (4)$$

where $\hat{x}_{i,j}$ = value of estimate missing data between x_i and x_j, x_k = values that match the test indicator is missing, in the test group ranging from 1, 2, ..., k

E. Accuracy Evaluation Criterion

Accuracy evaluation criterion [1] of a new data set which must be precisely compatible (model best fit) by applying a set of new data (predicted data) derived from the estimation of missing data to verify the real data set (actual data) and then calculate the magnitude of relative error (MRE) according to the formula

$$MRE = \frac{|Actual - Predicted|}{Actual} \quad (5)$$

The missing data (i = 1, 2, ..., n) must be used for calculating the mean magnitude of relative error (MMRE). If it is found that the results of MMRE have small values, the results should be precise or very close to the real data as formulated below

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Actual_i - Predicted_i|}{Actual_i} x100 \quad (6)$$

$$Accuracy = 100 - MMRE \quad (7)$$

III. EXPERIMENTS

A. *Classification of data sets for the research*

Classification or divided data set of waveform database generator 1,200 sets into 2 groups: the experimental group was 1,000 data sets and the test group was 200 data sets.

B. *The factor analysis of experimental group*

The experimental group focused on the factor analysis method by principle component analysis to provide a measure that is relevant to the factors in the same way as rotation varimax to reduce the number of points. This should measure the weight of each factor to as low as possible. Results from the analysis of new factors with KMO were 0.961, and new factors from extraction consist of three main factors F1, F2 and F3 are shown in Table I.

TABLE I. Results of main factors and indicators

Factor	Indicator of Factor
F1	V17 , V9 , V10 , V16 , V15 , V18 , V8 , V19 , V20
F2	V5 , V13 , V12 , V6 , V7 , V4 , V14 , V11 , V3 , V2
F3	V21, V1

C. *Partial Least Squares of Structural Equation Modeling*

The main factors F1, F2 and F3 of building a structural equation model are shown in Fig. 1. The model appropriate to review the statistics of the compatibility of the model to goodness of fit: RMSEA, GFI and Hoelter's N which is the adequacy of sample case. The results in Table II and the new structural equation model are shown in Fig. 2.

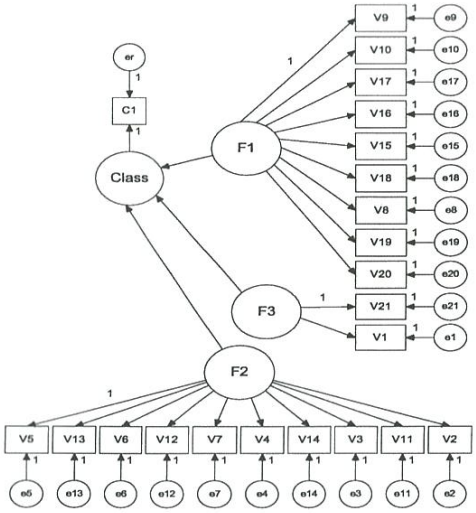

Fig. 1. Prototype of structural equation model

TABLE II. The statistics compatibility structural equation model

Model	χ^2	P	GFI	RMSEA	Hoelter's N
Default	27.7	0.956	0.995	0.000	2094/2385

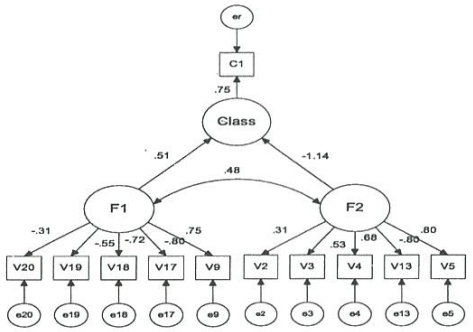

Fig. 2. Structural equation model standardized type

The only measure left over from the results of the structural equation model according to Fig. 2 is F1 = {V9, V17, V18, V19, V20} and F2 = {V2, V3, V4, V5, V13} to create new factors related. By bringing a measure of the factor F1 one by one to match multiplied with a measure of factor F2 all and repeat until all indicators of the factors F1. Result is F1F2 = {V17V2, V17V3, V17V4, V17V5, V17V13, V18V2, V18V3, V18V4, V18V5, V18V13, V19V2, V19V3, V19V4, V19V5, V19V13, V20V2, V20V3, V20V4, V20V5, V20V13}, and then create new structural equation model as Fig. 3 have the statistics of compatibility in Table III and the new structural equation model is depicted in Fig. 4 with the equation estimated by equation 8-11.

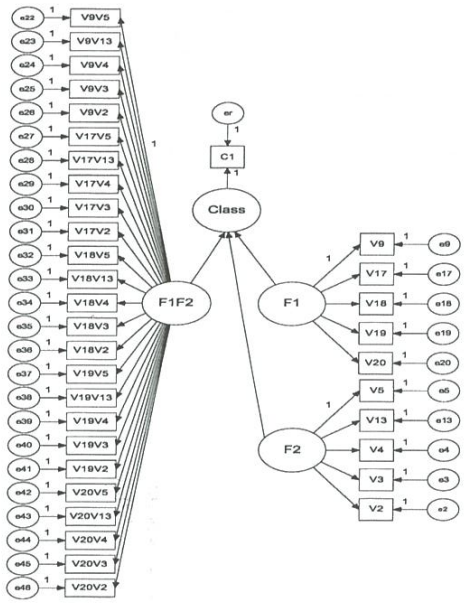

Fig. 3. Prototype PLS of SEM

TABLE III. The statistics compatibility PLS of SEM

Model	χ^2	P	GFI	RMSEA	Hoelter's N
Default	26.907	0.107	0.993	0.020	1120/1344

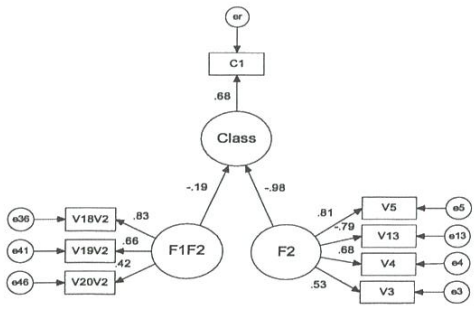

Fig. 4. PLS of SEM standardized type

$$Class = (0.68*C1)+er \quad (8)$$

$$F2 = (Class+(0.19*F1F2))/(-0.98) \quad (9)$$

$$F1F2 = (0.83*V18V2)+(0.66*V19V2)+(0.42*V20V2) \quad (10)$$

$$V5 = (F2-((0.53*V3)+(0.68*V4)-(0.79*V13)))/(0.81) \quad (11)$$

D. K-Nearest Neighbor

The F2 key factors of the experimental group. (which is a measure V5 defined as the missing data included in this factor) to calculate the distance by Euclidean distance to the test data with missing value by equation 3. Selected 5 (K=5) of data sets were nearest to the missing data for estimate of the new indicator V5 by equation 4.

IV. RESULTS

The test group of 200 sets were assigned to find missing V5 and estimate the replacement value of missing data as follows: 1) the data imputation estimation methods using partial least squares of structural equation modeling as equation 8-11, the result of MMRE was 36.90% (accuracy was 63.10%) and 2) the data imputation estimation methods using K-nearest neighbor, The result of MMRE was 61.52% (accuracy was 38.48%) shown in Table IV.

TABLE IV. Comparison of estimates of missing data

Model	PLS of SEM	K-NN
MMRE	36.90	61.52
Accuracy	63.10	38.48

Thus, comparing estimates of missing data showed that using the partial least squares of structural equation modeling and related indicators had high accuracy about 24.62%, while MMRE declined using the K-nearest neighbor.

V. DISCUSSION AND CONCLUSIONS

Data imputation estimation methods using partial least squares of structural equation model with a data set from the waveform database generator. Numeric indicators 21 of 1,200 sets of nonlinear type showed that the grouping of data sets or analysis of main factors for the indicators are related to factors in the same area. When estimating missing data, the results of MMRE errors were reduced. Making a new data from the missing estimation method is more accurate then the new values.

Suggestions about the data imputation estimation methods using Product Indicator Approaches. The related indicators are used in the case of latent factors outside the relationship between the two directions only. If no such event, this method will not be able to be used.

REFERENCES

[1] Prakancharoen S. "The estimated time to develop application software oriented network Using structural equation modeling". *Information Technology Journal*. Year 4 Vol. 7. Bangkok: King Mongkut's University of Technology North Bangkok, 2008.

[2] Phothi N. and Prakancharoen S. "Accuracy Comparison of Imputation Methods Using Structural Equation Modeling Between With Discriminant Analysis and Without Discriminant Analysis". *Conference on Science and Technology No. 8*. Pathum Thani: Thammasat University Rangsit Campus, 2010.

[3] Rufus L. C. *Solutions for Missing Data in Structural Equation Modeling*. Research & Practice in Assessment Vol. 1, Issue 1 March 2006.

[4] Meesad P.and Hengpraprohm K. "Combination of KNN-Based Feature Selection and KNN-Based Missing-Value Imputation of Microarray

Data". *The 3rd Intetnational Conference on Innovative Computing Information and Control (ICICIC'08)*. IEEE computer society.

[5] Thomas B. *K-Nearest Neighbors Algorithm: Prediction and Classification*. Department of Economics Southern Methodist University Dallas, TX 75275 February 2008.

[6] Leo B., Jerome H. F., Adam O., Jonathan S. *Classification and Regression Trees*. Wadsworth International Group. California, 1984.

[7] Vanitbancha K. *Multivariate Data Analysis*. Vol. 2. Bangkok: Chulalongkorn University Book Center, 2007.

[8] Garson G. D. *Data Imputation for Missing Values*. North Carolina State University, USA, 2005.

[9] Wynne W. C., Barbara L. M., Peter R. N. "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice Mail Emotion/Adoption Study". *Proceedings of the Seventeenth Internatinal Conference on Information Systems*. Cleveland, Ohio, December 16-18, 1996.

[10] Karin S., Christina W., Helfried M. "Nonlinear Structural Equation Modeling: Is Partial Least Squares an Alternative?". *Meeting of the Working Group Structural Equation Modeling*. Berlin, Germany, February 26-27, 2009.

[11] Jöreskog, K. G., & Yang, F. *Nonlinear structural equation models: The Kenny-Judd model with interaction effects*. In G. Marcoulides & R. Schumacker (Eds.), Advanced structural equation modeling (pp. 57-87). Mahwah, NJ: Lawrence Erlbaum Associates. 1996.

[12] Marsh, H. W., Wen, Z., & Hau, K. T. *Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction*. Psychological Methods, 9, 275–300. 2004.

[13] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. B. "Missing values estimation methods for DNA microarrays". *Bioinformatics*, 2001, vol. 17, pp. 520-525.