The Estimation of Missing Values in Air Quality Data from Nortern Thailand Air Quality Monitoring Stations

Phatarapon Vorapracha[1*], Rungruang Musiri[2], Airiya Pongpittaya[3] and Suparada Khanaruksombat[4]

[1, 2] Division of Information Technology Management, Faculty of Industrial Technology, Phranakhon Rajabhat University.
[3, 4] Division of Information Technology, Faculty of Science and Technology, University of Bangkok Thonburi.
[*]**Corresponding author**, **E-mail:** phatarapon@yahoo.com

Abstract

The Carbon dioxide (CO_2), Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), Ozone (O_3) and Particulate measure (PM_{10}) that stored from air monitoring station show the missing data due to various reasons i.e. human data entry errors or measurements data entry errors. The issue of the missing values can affect the performance of management and analysis for data integrity. The important factor for selected the method of missing data or processes and data mechanisms. There are various ways to deal with missing values. Some of the methods are available to analyze their own to handle the missing data with reduced the data set to solve by putting up in missing data. This paper presents a simple method for inserted the missing data via using most common data, data mining algorithms association rules or regression interpolation method

Keywords: Missing values, Interpolation

Introduction

The report of Thailand air quality from the Pollution Control Department showed that air pollutants are the most problematic, especially particulate measure smaller than 10 microns (PM_{10}) because there is a high concentration of dust levels that may cause health problems.The health of the population both short-term and chronic cause by respiratory particulate matter are harmful to health. It is very important for planning the prevention and control of air pollution in the future. The Pollution Control Department was monitoring the air quality carried out of the country by measuring the air quality monitoring stations but permanent air quality monitoring stations have restrictions on high investment. The storage of air quality monitoring stations are stored Carbon dioxide (CO_2), Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), Ozone (O_3) and Particulate measure (PM_{10}) by storing hour by hour in every station. Nowadays, each measurement stations do not have adequate tools and some tools were damaged that make it impossible to store data at certain times. Some of the stations cannot collect such data especially in the north of the country consists of nine provinces including Chiang Mai, Lamphun, Lampang, Mae Hong Son, Nan, Phayao, Phrae and Uttaradit. The North of Thailand has swampy pan terrain, surrounded by mountains and the flow of air is relatively stagnant especially during the winter on December to January. The measurements air quality station in Northern Thailand has a total of 14 stations: central Hall Chiang Mai station, Yuparat station, Chiang Rai Station, Mae Sai station, Lampang Meteorological Station, Pillar shine station, Sobpard station, Tasri station, Papar station, Mae Hong Son station, Nan Station, Lamphun station, Pare station and Payaow station. This paper presents the missing values in data estimation.

I. Characteristics of Missing Data Mechanisms

The missing data mechanism is usually classified as missing completely at random, missing at random or not missing at random (K. Lakshminarayan, A. Harp, & T. Samad, 1999) Unfortunately identification of missing data mechanism is not always easy.

A. *Missing Completely at Random (MCAR)*

The missing data mechanism is considered as missing completely at random (MCAR), when the probability of a record having a missing value for an attribute does not depend on either the observed data or the missing data. MCAR is sometimes called uniform non-response. An example of a MCAR mechanism would be that a laboratory sample is dropped, so the resulting observation is missing. Data which is missing due to structural reasons cannot be regarded as MCAR. (London School of Hygiene and Tropical Medicine, 2013)

B. *Missing at Random (MAR)*

The missing data mechanism is considered as missing at random (MAR), when the probability of a record having a missing value for an attribute could depend on the observed data, but not on the value of the missing data itself. Data which is incomplete only due to structural reasons are MAR. (N. Horton & K. P. Kleinman, 2007) describe MAR mechanism as states that the missingness depends only on observed quantities, which may include outcomes and predictors (in which case the missingness is sometimes labeled covariate dependent missingness (CDM))". A special case of MAR is uniform non-response within classes. For example (London School of Hygiene and Tropical Medicine, 2013) , suppose we seek to collect data on income and property tax band. Typically, those with higher incomes may be less willing to reveal them. If we have everyone's property tax band and given property tax band non-response to the income question is random, then the income data is missing at random. The reason (or mechanism) for it being missing depends on property band. Given property band missingness does not depend on income itself.

C. Not Missing at Random (NMAR)

The missing data mechanism is considered as not missing at random (NMAR), when the probability of a record having a missing value for an attribute could depend on the value of the attribute. Missing data mechanism that is considered as NMAR is non-ignorable. This can be solved by going back to the source of data and obtaining more information about the mechanism or obtain complete data set. Unfortunately it is very rare to know the appropriate model for the missing data mechanism.

Examples NMAR missing data mechanism (K. Lakshminarayan, A. Harp, & T. Samad, 1999) include a sensor not detecting temperatures below a certain threshold, people not filling in yearly income in surveys if the income exceeds a certain value.

Some sources, e.g. (N. Horton and K. P. Kleinman, 2007), called this mechanism missing not at random (MNAR).

D. Strategies for Dealing with Missing

There are three main strategies for dealing with missing data. The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all samples with missing values (M. Kantardzic, 2003). Another solution is to treat missing values as special values. Finally missing values problem can be handled by various missing values imputation methods. Unfortunately missing values imputation methods are suitable only for missing values caused by missing completely at random (MCAR) and some of them for missing at random (MAR) mechanism. If missing values are caused by not missing at random mechanism (NMAR) it must be handled by going back to the source of data and obtaining more information or the appropriate model for the missing data mechanism have to be taken into account.

1) Using Missing Values Policy of Used Analytical Method

There is no need to use a special method for dealing missing values if method that is used for data analysis has its own policy for handling missing values. Decision rules extraction methods may consider attributes with missing values as irrelevant. (J. Luengo, 2013) Association rules extraction methods may ignore rows with missing values (conservative approach) or handle missing values as they are supporting the rule (optimistic approach) or are in contrary with the rule (secured approach) as described by Berka.

2) *Reducing the Data*

The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all missing values. This can be done by elimination of samples (rows) with missing values [4] or elimination of attributes (columns) with missing values. (K. Lakshminarayan, A. Harp, & T. Samad, 1999) Both approaches can be combined. Elimination of all samples is also known as complete case analysis.

Elimination of all samples is possible only when large data sets are available, and missing values occur only in a small percentage of samples and when analysis of the complete examples will not lead to serious bias during the inference. Elimination of attributes with missing values during analysis is not possible solution if we are interested in making inferences about these attributes. Both approaches are wasteful procedures since they usually decrease the information content of the data.

3) *Treating Missing Attribute Values as Special Values*

This method deals with the unknown attribute values using a totally different approach. Rather than trying to find some known attribute value as its value, we treat missing value itself as a new value for the attributes that contain missing values and treat it in the same way as other values [6]. Instead of storing value of the attribute we store the information that the value is missing. This approach assumes that we handle these values as they don't influence future analyses.

4) *Replace Missing Value with Mean*

This method replaces each missing value with mean of the attribute [4]. The mean is calculated based on all known values of the attribute. This method is usable only for numeric attributes and is usually combined with replacing missing values with most common attribute value for symbolic attributes.

5) *Replace Missing Value with Mean for the Given Class*

This method is similar to the previous one. The difference is in that the mean is not calculated from all known values of the attribute, but only attributes values belonging to given class are used. This approach is possible only for classification problems where samples are classified in advance (M. Kantardzic, 2003) or there is a possibility to create the classes.

6) *Replace Missing Value with Median for the Given Class*

Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing value belongs. (E. Acuna and C. Rodriguez, 2004) This method is usable only for numeric attributes and requires existence of classes or possibility to create classes as previous method.

7) *Replace Missing Value with Most Common Attribute Value*

This method simply uses most common attribute value for missing value imputation. (J. W. Grzymala-Busse & M. Hu, 2001) The most common value of all values of the attribute is used. This method is usable only for symbolic attributes and is usually combined with replacing missing values with missing values imputation using mean for numeric attributes.

8) *Concept Most Common Attribute Value*

This method is similar to the previous one. The concept most common attribute value method is a restriction of the previous method to the concept. (J. W. Grzymala-Busse & M. Hu, 2001) This method uses most common value of the attribute but uses cases belonging to the given class or concept instead of using global most common value. This method is usable for symbolic attributes and requires existence of classes or possibility to create classes.

II. Missing Values Using Association Rules

An association rule is a simple probabilistic statement about the co-occurrence of certain events. For binary variables association rule takes the following form (D. J. Hand, H. Manilla & P. Smyth, 2001)

$$\text{IF } A=1 \text{ AND } B=1 \text{ THEN } C=1$$

where A, B and C are variables and

$$p=p(C=1|A=1,B=1)$$

the conditional probability that C=1 given that A=1 and B=1. The conditional probability p is referred to as the “confidence” of the rule, and p (A=1, B=1, C=1) is referred to as the “support”. The support can be used as a constraint of minimum count of cases supporting association rule. The “If part of the rule is often called antecedent and the “then” part is often called consequent

The input for missing values imputation is incomplete data set. Algorithms for association rules generation are usually unable to handle missing values. Some association rules extraction methods may ignore rows with missing values (conservative approach) or handle missing values as they are supporting the rule (optimistic approach) or are in contrary with the rule (secured approach) as described by Berka.

Another possibility is getting complete data set for association rules generation. First way to get complete data set for association rules generation is reducing the data set.by elimination of cases and/or attributes as described in one of previous chapters. Another way to get complete data set for association rules generation is to handle missing values as special values. Using special values to obtain complete data set for association rules can be recommended because association rules with special values can be easily ignored. Association rules cannot be often generated directly from numeric attributes. Missing values imputation may be directly used only for symbolic attributes.

Algorithms for association rules generation usually generates rules based on setting of minimum support, minimum confidence and maximum number of rules parameters. Most analyses, that use association rules methods, use only few association rules. For missing values imputation usually many more rules are required because not all association rules are suitable for missing values imputation.

Useable association rule for missing values imputation has the consequent containing value of attribute whose value is being searched and the antecedent correspond to values of other given case attributes. If complete data set for association rules generation was obtained by replacing missing values by special values association rules with these special values in consequent must be omitted. It is possible to use only association rules with consequent length equal to 1 and impute missing values one by one in cases with more than one missing values. Another possibility is using association rules with consequent length equal to the count of missing values in given case.

Support and confidence of the rule are two criteria that should be maximized during making decision about using the association rule for missing value imputation. For missing value imputation can be used the rule with maximum confidence along with required support. It IS also possible to ignore support of the rule and use only confidence. Setting required min. support of the rule highly depends on given data set.

There is a possibility to combine association rules approach with most common attribute value method. At first the most common attribute value method can be used if it is not possible to find suitable association rule with required support for missing value imputation. At second the most common attribute value method can be used if there is no suitable association rule with required support and confidence not lower then relative frequency of occurrence of most common attribute value. It is also possible to use only few association rules to improve missing values imputation accuracy using the most common attribute value.

III. Regression Interpolation Method

Interpolation is a method that adjusts missing value to reduce the estimated deviation. The basic idea of regression interpolation is using the linear relationship of the auxiliary variables X_k = (k = 1, 2, ...) and the target variable Y to establish the regression model and using the known information of auxiliary variables to estimate the missing values of the target variables. So the estimated value of missing value can be expressed as:

$$Z_i = \beta_0 + \sum_{k=1}^{k} \beta_k + X_{ki} + e_i$$

Where β is the regression coefficient. If the auxiliary variables are qualitative variables, we could use the approach of dummy variable. If the target variable Y is a qualitative variable, we may consider the Log it transforms for Logistic linear Regression. The regression model above could have different forms of evolution. In view of the characteristics of time series, this paper uses linear regression, multiple linear regressions, and iterative regression to interpolate missing value [8].

A. *One Variable Linear Regression*

A reasonable regression equation is given by construction one variable regression model between induced variable and a certain correlative dependent variable, and then the missing value can be estimated. Assuming that the element j in the data sample is missing value, that is

$$X_i = (X_{i1}, X_{i2}, ..., X_{ij-1}, X_{ij+1}, ..., X_{im})$$

If we may find the regression equation of incomplete variable and a certain related variable:

$$v_j = u_0 + \mu v_i$$

The values of missing data can be obtained simply and efficiently. The task of one variable linear regression is to find the best linear fit. It is usually obtained by using the least square method.

B. *Multivariate Regression Model*

Albrecht studied the application of multivariate regression model in dealing with missing value in 1992. (G. H. Albrecht, 1992) In the processing of the obviously relative variable data set, the effect is usually better and more direct than other statistical method. It constructs the regression model between independent variable and induced variable through regression analysis and provides the reasonable regression equation, and then estimates the missing value. Assuming that the element j of data sample is missing data, that is

$$X_i = (X_{i1}, X_{i2}, ..., X_{ij-1}, X_{ij+1}, ..., X_{im})$$

If we can find the regression equation of incomplete variable and other variables:

$$v_j = u_0 + \sum_{\substack{l:l=1 \\ l \neq j}}^{m} \mu_l v_l + \varepsilon$$

It is easy to predict the values of missing data. We can establish a reasonable regression equation, or establish a regression equation of a transformed variable (transfer the nonlinear model into linear model) to use regression model to predict the missing value.

For the example of missing information from Carbon dioxide (CO_2), Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), Ozone (O_3) and Particulate measure (PM_{10}).

Table 1 Missing information from Carbon dioxide (CO_2), Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), Ozone (O_3) and Particulate measure (PM_{10})

Date	Hour	CO	NO_2	SO_2	O_3	PM_{10}
		at 3 m (ppm)	at 3 m (ppb)	at 3 m (ppb)	at 3 m (ppb)	at 3 m ($\mu g/m^3$)
120101	100	1	22	2	1	67
120101	200	0.7	18	2	1	47
120101	300	0.8	18	1	1	45
120101	400	0.7	15	1	1	33
120101	500	0.5	-	1	1	35
120101	600	0.4	7	1	3	27
120101	700	0.3	8	-	3	28
120101	800	0.5	13	1	2	33
120101	900	0.9	21	1	5	55
120101	1000	0.6	17	1	16	42
120101	1100	-	-	-	-	40
120101	1200	0.3	11	1	42	30

IV.CONCLUSION AND FUTURE WORK

The insertion of missing values based on data set structure of attributes and missing data. The important factor in deciding to choose the right approach is by view completely at random, missing at random or not missing at random. For the further work, missing values from 14 stations were tested. There are central Hall Chiang Mai station, Yuparat station, Chiang Rai Station, Mae Sai station, Lampang Meteorological Station, Pillar shine station, Sobpard station, Tasri station, Papar station, Mae Hong Son station, Nan Station, Lamphun station, Pare station and Payaow station. Missing Values Using Association Rules and Regression Interpolation Method were used by the data in 2012-2015.

Acknowledgment

This study was supported byDivision of Information TechnologyManagement, Faculty of Industrial Technology, Phranakhon Rajabhat University, Thailand.

References

D. J. Hand, H. Manilla & P. Smyth. (2001). *Principles of Data Mining* (1st ed). (pp.150-151). Cambridge, Massachusetts London England : n.p.

E. Acuna & C. Rodriguez. (2004). *The Treatment of Missing Values and its Effect on Classifier Accuracy, Classification, Clustering, and Data Mining Applications* (pp.639-647). N.P. : Springer.

G. H. Albrecht. (1992). *Multivariate morphometrics with missing data: techniques for canonical varieties and generalized distances, Am. J. phys. Anthropol. Suppl* (pp.14-42). N.P. : Springer.

J. Luengo. (2 September 2013). *Missing Values in Data Mining*. Retrieved form http://sci2s.uqr.es /MVDIWindex.php

J. W. Grzymala-Busse & M. Hu. (2001). A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In *the Second International Conference on Rough Sets and Current Trends in Computing* (pp. 378-385). N.P. : Springer.

K. Lakshminarayan, A. Harp & T. Samad. (1999). Imputation of Missing Data in Industrial Databases. *Applied Intelligence*, *11*(12), 259-275.

London School of Hygiene and Tropical Medicine. (2 September 2013). *Missingness mechanisms*. Retrieved form http://missinqdata.lshtm.ac.uk/index.php?view=cateqorv&id=40%3Amissinqness- echanisms&option=com content&ltemid=96

M. Kantardzic. (2003). Data Mining - Concepts, Models, Methods, and Algorithms. *IEEE*, 65-176.

N. Horton & K. P. Kleinman. (2007). Much Ado About Nothing: A Comparison of Missing DataMethods and Software to Fit Incomplete Data Regression Models. *The American Statistician*, *61*(1), 79-90.