ISSN 1512-1232

POWER CONSTANT BASED METHODS FOR DEALING WITH MISSING VALUES IN KNOWLEDGE DISCOVERY

Shukla Diwakar[1], Rahul Singhai[2], Narendra Singh Thakur[3]

[1]Deptt. of Mathematics and Statistics Dr. H.S. Gour University, Sagar(M.P.) India
email: diwakarshukla@rediffmail.com
[2]IIPS, Devi Ahilya University, Indore (M.P.) India
email : singhai_rahul@hotmail.com,
[3]B.T. Institute of Research and Technology Sironja, Sagar (M.P.) India
email : nst_stats@yahoo.co.in

Abstract

One relevant problem in data quality is the presence of missing data. In cases where missing data are abundant, effective ways to deal with these absences could improve the performance of Data Mining. Missing data can be treated using imputation. Imputation methods replace the missing data by values estimated from the available data. Missing data imputation is an actual and challenging issue in data mining. This is because missing values in a dataset can generate bias that affects the quality of the learned patterns. To deal with this issue, this paper proposes some Imputation methods, which can impute missing values with negligible biased data. We experimentally evaluate our approach and demonstrate that it is much more efficient than the other available imputation methods.

Keywords: *KDD (Knowledge Discovery in Databases.) Data mining attribute missing values, Imputation methods, Sampling.*

1.0 INTRODUCTION

In recent years, the rapid development of data-mining techniques has enabled successful knowledge discovery applications in various industries (Han and Kamber 2000). Data pre-processing is a critical task in the knowledge discovery process for ensuring quality of mined patterns. Missing values could generate serious problems on knowledge extraction and on Data Mining algorithms application (Pyle (1999), Liu et al. (2005) and Fujikawa (2001)). Since many of data analysis algorithms can work only with complete data. Therefore different strategies to work with data that contains missing values, and to fill in missing values in the data are developed (Liu et al. (1997), Kalton and Kasprzyk (1982), Little and Rubin (1987), McQueen (1967), Pyle (1999), Ragel et al. (1998,1999), and Lee et al. (1976)).

The problem of missing value handling has been studied for many years with numerous methods proposed. The existing methods can be categorized into two types: imputation-based and data mining-based methods. The former types of methods are primarily for handling missing values of (Tseng et al.(2002)) numerical data, while the latter (2003) for categorised data. The principle of imputation methods is to estimate the missing values by using the existing values as an auxiliary base. The underlying assumption is that there exist certain correlations between different data tuples over all attributes. Rubin (1976) addressed three missing observation concepts: missing at random (MAR), observed at random (OAR) and parameter distribution (PD). They have shown its application to 'mass' imputation under two-phase sampling and deterministic imputation for missing data. Ahmed et al. (2006) generated several generalized structure of imputation procedure and their corresponding estimators of the population mean. Shukla and Thakur (2008) suggested the use of factor-type (F-T) estimator as a tool of imputation for non-responding units in the sample. Shukla et al. (2010) proposed some imputation methods to treat missing values in knowledge discovery in Data warehouse.

ISSN 1512-1232

Data mining-based methods area has different approaches like to ignore the tuple that contains missing values, fill in the missing values manually, use a global constant to fill in the missing values, use the attribute mean to fill in the missing values, use the most probable value or default value (Kantardzic 2003) to fill in the missing values. During the knowledge discovering process on a database, the substitution by a default value can introduce distorted information, which is not present on the event and circumstances that generate these instances (Pyle 1999).The instances (records) or attributes elimination could result in loss of important information related to present values (Fujikawa 2001). These techniques are applied only when the number of missing values is short. Moreover, important point on knowledge discovery process on database, requiring careful value predictions using more advanced and elaborated techniques and procedures, together with the tacit knowledge of a problem domain expert and the pre-processing of the database (Hofmann et al. 2003). Techniques such as associations (Ragel and Cremilleux 1998), clustering (Lee et al. 1976), and classifications (Liu et al. 1997) are used to discover the similar patterns between data tuples to predict the missing values.

For the existing methods, it was observed that specialized methods are needed for different types of missing data, and no single method can handle well to all kinds of missing data sets(Tseng et al. 2003).

In this work, we confine the study to handling missing values of the numerical type and suggest some imputation methods. To evaluate the performance of the proposed method, an artificial experiment is conducted under data set with missing ness. The empirical results show that the proposed methods deliver accuracy in recovering the missing values.

Let $\Omega = \{1, 2,..., N\}$ be a finite database with Y_i as a variable of main interest and X_i $(i = 1,2,..., N)$ an auxiliary variable. As usual, $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$, $\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$ are database means, $\overline{X}$ is assumed known and $\overline{Y}$ under investigation.

Consider preliminary large sample S' of size n' drawn from Dataset Ω by SRSWOR and a secondary sample of size n ($n < n'$) drawn in either of the following manners:

Case I : as a sub-sample from sample S' (denote by F_1) as in fig. 1(a),

Case II: draw independently to sample S' (denote by F_2) as in fig. 1(b) without replacing S'.

The sample S of n units contains r responding units ($r < n$) forming a subspace R and $(n - r)$ non-responding with sub-space R^C in $S = R \bigcup R^C$. For every $i \in R$, the y_i is observed available. For $i \in R^C$, the y_i values are missing and imputed values are to be derived. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R^C$. Assume for S, the data $x_s = \{x_i : i \in S\}$ and for $i' \in S'$, the data $\{x_{i'} : i' \in S'\}$ are known with mean $\overline{x} = (n)^{-1}\sum_{i'=1}^{n} x_{i'}$ and $\overline{x}' = (n')^{-1}\sum_{i'=1}^{n'} x_{i'}$ respectively.

ISSN 1512-1232

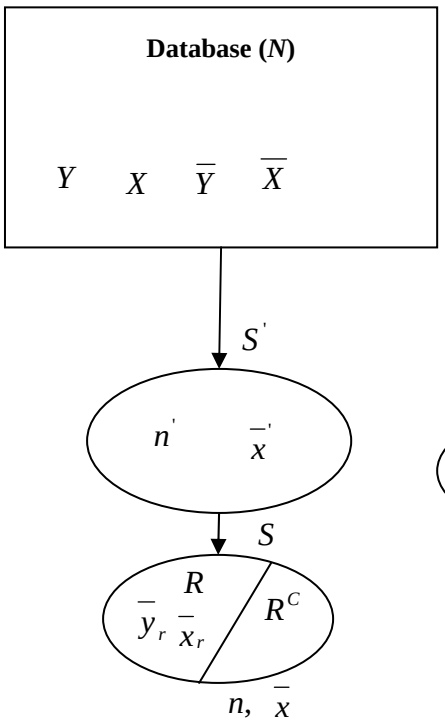

Fig. 1.1.1(a) [F_1]

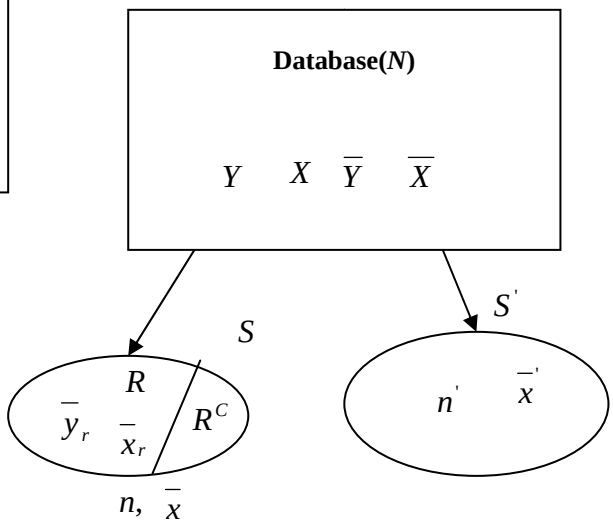

Fig. 1.1.1(b) [F_2]

2.0 PROPOSED METHODS OF IMPUTATION

Let y'_{ji} denotes the i^{th} available observation for the j^{th} imputation. We suggest the following imputation methods for missing data in database:

(1)
$$y'_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r \left(\dfrac{\bar{x}'}{\bar{x}} \right)^{\beta_1} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad \ldots(2.1)$$

where β_1 is suitably chosen constant, such that the variance of the resultant estimator is minimum. Under this method, the point estimator of $\bar{Y}$ is

$$t'_1 = \bar{y}_r \left(\frac{\bar{x}'}{\bar{x}} \right)^{\beta_1} \qquad \ldots(2.2)$$

(2)
$$y'_{2i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r \left(\dfrac{\bar{x}}{\bar{x}_r} \right)^{\beta_2} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad \ldots(2.3)$$

ISSN 1512-1232

where β_2 is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of $\bar{Y}$ is

$$t_2^{'} = \bar{y}_r \left(\frac{\bar{x}}{\bar{x}_r} \right)^{\beta_2} \quad \ldots(2.4)$$

(3)

$$y_{3i}^{'} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{x}^{'}}{\bar{x}_r} \right)^{\beta_3} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \quad \ldots(2.5)$$

where β_3 is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of $\bar{Y}$ is

$$t_3^{'} = \bar{y}_r \left(\frac{\bar{x}^{'}}{\bar{x}_r} \right)^{\beta_3} \quad \ldots(2.6)$$

When $\beta_3 = 1, \text{ then } t_3^{'} = \bar{y}_r \left(\dfrac{\bar{x}^{'}}{\bar{x}_r} \right)$ …(2.7)

(Ratio type estimator in two-phase sampling)

and when $\beta_3 = -1, \text{ then } t_3^{'} = \bar{y}_r \left(\dfrac{\bar{x}_r}{\bar{x}^{'}} \right)$ …(2.8)

(Product type estimator in two-phase sampling)

This natural analogue of the ratio estimator which is called the product estimator when an auxiliary variate *X* has a negative correlation with *Y*, where *X* and *Y* are variates that take only positive values. (Cochren, 2005).

3.0 AHMED METHODS OF IMPUTATION

For the case where y_{ji} denotes the i^{th} available observation for the j^{th} imputation method. Ahmed et al. (2006) suggested the following:

(A) :

$$y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \quad \ldots(3.1)$$

Under this, the point estimator of $\bar{Y}$ is

$$t_1 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} \quad \ldots(3.2)$$

ISSN 1512-1232

Lemma 3.1:

(1) The bias of t_1 is :

$$B(t_1)=\left(\frac{1}{n}-\frac{1}{N}\right)\overline{Y}\left[\frac{\beta_1(\beta_1+1)}{2}C_x^2-\beta_1\rho C_Y C_X\right] \quad \ldots(3.3)$$

(2) The m.s.e. of t_1 is :

$$M(t_1)=\overline{Y}^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)C_Y^2+\left(\frac{1}{n}-\frac{1}{N}\right)\beta_1^2C_x^2-\left(\frac{1}{n}-\frac{1}{N}\right)2\beta_1\rho C_Y C_X\right] \quad \ldots(3.4)$$

(3) The minimum m.s.e. of t_1 is :

$$M(t_1)_{\min}=\left(\frac{1}{r}-\frac{1}{N}\right)S_Y^2-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{XY}^2}{S_X^2} \quad \ldots(3.5)$$

for the optimum value of β_1 which is given by $\beta_1=\rho\frac{C_Y}{C_X}$.

(B) : $$y_{2i}=\begin{cases} y_i & if \quad i\in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^{\beta_2}-r\bar{y}_r\right] & if \quad i\in R^C \end{cases} \quad \ldots(3.6)$$

Under this, the point estimator of $\overline{Y}$ is

$$t_2=\bar{y}_r\left(\frac{\bar{x}_n}{\bar{x}_r}\right)^{\beta_2} \quad \ldots(3.7)$$

Lemma 3.2:

(4) The bias of t_2 is :

$$B(t_2)=\left(\frac{1}{r}-\frac{1}{n}\right)\overline{Y}\left[\frac{\beta_2(\beta_2+1)}{2}C_x^2-\beta_2\rho C_Y C_X\right] \quad \ldots(3.8)$$

(5) The m.s.e. of t_2 is :

$$M(t_2)=\overline{Y}^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)C_Y^2+\left(\frac{1}{r}-\frac{1}{n}\right)\beta_2^2C_x^2-\left(\frac{1}{r}-\frac{1}{n}\right)2\beta_2\rho C_Y C_X\right] \quad \ldots(3.9)$$

(6) The minimum m.s.e. of t_2 is :

$$M(t_2)_{\min}=\left(\frac{1}{r}-\frac{1}{N}\right)S_Y^2-\left(\frac{1}{r}-\frac{1}{n}\right)\frac{S_{XY}^2}{S_X^2} \quad \ldots(3.10)$$

for the optimum value of β_2 which is given by $\beta_2=\rho\frac{C_Y}{C_X}$.

(C) : $$y_{3i}=\begin{cases} y_i & if \quad i\in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\overline{X}}{\bar{x}_r}\right)^{\beta_3}-r\bar{y}_r\right] & if \quad i\in R^C \end{cases} \quad \ldots(3.11)$$

ISSN 1512-1232

Under this, the point estimator of $\overline{Y}$ is

$$t_3 = \overline{y}_r \left(\frac{\overline{X}}{\overline{x}_r} \right)^{\beta_3} \quad \ldots(3.12)$$

As special cases :

when $\beta_3 = 1$, then $t_{Ratio} = \overline{y}_r \left(\frac{\overline{X}}{\overline{x}_r} \right)$ …(3.13)

and when $\beta_3 = -1$, then $t_{\Pr oduct} = \overline{y}_r \left(\frac{\overline{x}_r}{\overline{X}} \right)$ …(3.14)

Lemma 3.3:

(7) The bias of t_3 is :

$$B(t_3) = \left(\frac{1}{r} - \frac{1}{N} \right) \overline{Y} \left[\frac{\beta_3 (\beta_3 + 1)}{2} C_x^2 - \beta_3 \rho C_Y C_X \right] \quad \ldots(3.15)$$

(8) The m.s.e. of t_3 is :

$$M(t_3) = \overline{Y}^2 \left(\frac{1}{r} - \frac{1}{N} \right) \left[C_Y^2 + \beta_3^2 C_x^2 - 2\beta_3 \rho C_Y C_X \right] \quad \ldots(3.16)$$

(9) The minimum m.s.e. of t_3 is :

$$M(t_3)_{\min} = \left(\frac{1}{r} - \frac{1}{N} \right) S_Y^2 (1 - \rho^2) \quad \ldots(3.17)$$

for the optimum value of β_3 which is given by $\beta_3 = \rho \frac{C_Y}{C_X}$.

4.0 PROPERTIES OF SUGGESTED STRATEGIES

Let *B*(.) and *M*(.) denote the bias and mean squared error (*M.S.E.*) of an estimator under a given sampling design. The large sample approximations are

$$\overline{y}_r = \overline{Y}(1 + e_1);\ \overline{x}_r = \overline{X}(1 + e_2);\ \overline{x} = \overline{X}(1 + e_3) \text{ and } \overline{x}' = \overline{X}(1 + e_3')$$

Using the concept of two-phase sampling, following Rao and Sitter (1995) and the mechanism of MCAR, for given *r*, *n* and n', we have:

(i) Under design F_1

$$E(e_1) = E(e_2) = E(e_3) = E(e_3') = 0;\ E(e_1^2) = \delta_1 C_Y^2;\ E(e_2^2) = \delta_1 C_X^2;\ E(e_3^2) = \delta_2 C_X^2;$$
$$E(e'^2_3) = \delta_3 C_x^2;\ \phi = -K_{YZ}\ E(e_1 e_3) = \delta_2 \rho\ C_Y C_X;\ E(e_1 e_3') = \delta_3 \rho C_Y\ C_X;$$
$$E(e_2 e_3) = \delta_2\ C_X^2;\ E(e_2 e_3') = \delta_3\ C_X^2;\ E(e_3 e_3') = \delta_3\ C_X^2;$$

(ii) Under design F_2

$$E(e_1) = E(e_2) = E(e_3) = E(e_3') = 0;\ E(e_1^2) = \delta_4 C_Y^2;\ E(e_2^2) = \delta_4 C_X^2;\ E(e_3^2) = \delta_5 C_X^2;$$
$$E(e'^2_3) = \delta_3 C_x^2;\ E(e_1 e_2) = \delta_4 \rho\ C_Y\ C_X;\ E(e_1 e_3) = \delta_5 \rho\ C_Y C_X;\ E(e_1 e_3') = 0;$$
$$E(e_2 e_3) = \delta_5\ C_X^2;\ E(e_2 e_3') = 0;\ E(e_3 e_3') = 0$$

where $\delta_1 = \left(\frac{1}{r} - \frac{1}{n'} \right);\ \delta_2 = \left(\frac{1}{n} - \frac{1}{n'} \right);\ \delta_3 = \left(\frac{1}{n'} - \frac{1}{N} \right);\ \delta_4 = \left(\frac{1}{r} - \frac{1}{N - n'} \right);\ \delta_5 = \left(\frac{1}{n} - \frac{1}{N - n'} \right)$

ISSN 1512-1232

Theorem 4.1:

(1) Estimator $t_1^{'}$ in terms of e_i; $i=1,2,3$ and $e_3^{'}$, could be expressed :

$$t_1^{'}=\overline{Y}\left[1+e_1+\beta_1\left\{e_3^{'}-e_3-e_1e_3+e_1e_3^{'}-\beta_1e_3e_3^{'}+\frac{\beta_1+1}{2}e_3^2+\frac{\beta_1-1}{2}e_3^{'2}\right\}\right] \quad \ldots(4.1)$$

with ignorance of terms $E\left[e_i^r e_j^s\right]$, $E\left[e_i^r\left(e_j^{'}\right)^s\right]$ for $(r+s)>2$, where $r,s=0,1,2,\ldots$ and $i=1,2,3;\ j=2,3$ which is first order of approximation.

Proof: $t_1^{'}=\overline{y}_r\left(\frac{\overline{x}^{'}}{\overline{x}}\right)^{\beta_1}$

$$=\overline{Y}(1+e_1)(1+e_3^{'})^{\beta_1}(1+e_3)^{-\beta_1}$$

$$=\overline{Y}(1+e_1)\left(1+\beta_1e_3^{'}+\frac{\beta_1(\beta_1-1)}{2}e_3^{'2}\right)\left(1-\beta_1e_3+\frac{\beta_1(\beta_1+1)}{2}e_3^2\right)$$

$$=\overline{Y}\left[1+e_1+\beta_1\left\{e_3^{'}-e_3-e_1e_3+e_1e_3^{'}-\beta_1e_3e_3^{'}+\frac{\beta_1+1}{2}e_3^2+\frac{\beta_1-1}{2}e_3^{'2}\right\}\right]$$

(2) Bias of $t_1^{'}$ under design F_1 and F_2 is:

(i) $$B\left(t_1^{'}\right)_I=\overline{Y}\beta_1(\delta_2-\delta_3)\left(\frac{\beta_1+1}{2}C_X^2-\rho C_YC_X\right) \quad \ldots(4.2)$$

(ii) $$B\left(t_1^{'}\right)_{II}=\overline{Y}\beta_1\left[\frac{1}{2}\{\beta_1(\delta_3+\delta_5)-(\delta_3-\delta_5)\}C_X^2-\delta_5\rho C_YC_X\right] \quad \ldots(4.3)$$

Proof:

(i) $B\left(t_1^{'}\right)_I=E\left[t_1^{'}-\overline{Y}\right]_I$

$$=\overline{Y}\,E\left[1+e_1+\beta_1\left\{e_3^{'}-e_3-e_1e_3+e_1e_3^{'}-\beta_1e_3e_3^{'}+\frac{\beta_1+1}{2}e_3^2+\frac{\beta_1-1}{2}e_3^{'2}\right\}-1\right]$$

$$=\overline{Y}\beta_1(\delta_2-\delta_3)\left(\frac{\beta_1+1}{2}C_X^2-\rho C_YC_X\right)$$

(ii) $B\left(t_1^{'}\right)_{II}=E\left[t_1^{'}-\overline{Y}\right]_{II}$

$$=\overline{Y}\beta_1\left[\frac{1}{2}\{\beta_1(\delta_3+\delta_5)-(\delta_3-\delta_5)\}C_X^2-\delta_5\rho C_YC_X\right]$$

(3) Mean squared error of $t_1^{'}$ under design F_1 and F_2 respectively, upto first order of approximation could be written as:

(i) $$M\left(t_1^{'}\right)_I=\overline{Y}^2\left[\delta_1C_Y^2+(\delta_2-\delta_3)\left(\beta_1^2C_X^2-2\beta_1\rho C_YC_X\right)\right] \quad \ldots(4.4)$$

(ii) $$M\left(t_1^{'}\right)_{II}=\overline{Y}^2\left[\delta_4C_Y^2+(\delta_3+\delta_5)\beta_1^2C_X^2-2\delta_5\beta_1\rho C_YC_X\right] \quad \ldots(4.5)$$

Proof: $M\left(t_1^{'}\right)=E\left[t_1^{'}-\overline{Y}\right]^2=\overline{Y}^2E\left[1+e_1+\beta_1\left(e_3^{'}-e_3\right)-1\right]^2$

$$=\overline{Y}^2E\left[e_1^2+\beta_1^2\left(e_3^{'2}+e_3^2-2e_3e_3^{'}\right)+2\beta_1\left(e_1e_3^{'}-e_1e_3\right)\right] \quad \ldots(4.6)$$

(i) Under F_1 (Using (4.6))

ISSN 1512-1232

$$M\left(t_1^{'}\right)_I = \overline{Y}^2\left[\delta_1 C_Y^2 + \left(\delta_2 - \delta_3\right)\left(\beta_1^2 C_X^2 - 2\beta_1 \rho C_Y C_X\right)\right]$$

(ii) Under F_2 (Using (4.6))

$$M\left(t_1^{'}\right)_{II} = \overline{Y}^2\left[\delta_4 C_Y^2 + \left(\delta_3 + \delta_5\right)\beta_1^2 C_X^2 - 2\delta_5\beta_1\rho C_Y C_X\right]$$

(4) Minimum mean squared error of $t_1^{'}$ is :

(i) $$\left[M\left(t_1^{'}\right)_I\right]_{Min} = \left[\delta_1 - \left(\delta_2 - \delta_3\right)\rho^2\right]S_Y^2 \quad \text{when} \quad \beta_1 = \rho\frac{C_Y}{C_X} \qquad \ldots(4.7)$$

(ii) $$\left[M\left(t_1^{'}\right)_{II}\right]_{Min} = \left[\delta_4 - \left(\delta_3 + \delta_5\right)^{-1}\delta_5^2\rho^2\right]S_Y^2 \text{ when } \beta_1 = \delta_5\left(\delta_3 + \delta_5\right)^{-1}\rho\frac{C_Y}{C_X} \qquad \ldots(4.8)$$

Proof:

(i) $$\frac{d}{d\beta_1}\left[M\left(t_1^{'}\right)_I\right] = 0 \Rightarrow \beta_1 = \rho\frac{C_Y}{C_X}$$

$$\left[M\left(t_1^{'}\right)_I\right]_{Min} = \left[\delta_1 - \left(\delta_2 - \delta_3\right)\rho^2\right]S_Y^2$$

(ii) $$\frac{d}{d\beta_1}\left[M\left(t_1^{'}\right)_{II}\right] = 0 \;\; \beta_1 = \delta_5\left(\delta_3 + \delta_5\right)^{-1}\rho\frac{C_Y}{C_X}$$

$$\left[M\left(t_1^{'}\right)_{II}\right]_{Min} = \left[\delta_4 - \left(\delta_3 + \delta_5\right)^{-1}\delta_5^2\rho^2\right]S_Y^2$$

Theorem 4.2:

(1) The estimator $t_2^{'}$ in terms of e_1, e_2, e_3 and $e_3^{'}$ is

$$t_2^{'} = \overline{Y}\left[1 + e_1 + \beta_2\left\{e_3 - e_2 + e_1e_3 - e_1e_2 - \beta_2e_2e_3 + \frac{\beta_2+1}{2}e_2^2 + \frac{\beta_2-1}{2}e_3^2\right\}\right] \qquad \ldots(4.9)$$

Proof: $$t_2^{'} = \overline{y}_r\left(\frac{\overline{x}}{\overline{x}_r}\right)^{\beta_2} = \overline{Y}\left(1+e_1\right)\left(1+e_3\right)^{\beta_2}\left(1+e_2\right)^{-\beta_2}$$

$$= \overline{Y}\left(1+e_1\right)\left(1 + \beta_2e_3 + \frac{\beta_2\left(\beta_2-1\right)}{2}e_3^{'2}\right)\left(1 - \beta_2e_2 + \frac{\beta_2\left(\beta_2+1\right)}{2}e_2^2\right)$$

$$= \overline{Y}\left[1 + e_1 + \beta_2\left\{e_3 - e_2 + e_1e_3 - e_1e_2 - \beta_2e_2e_3 + \frac{\beta_2+1}{2}e_2^2 + \frac{\beta_2-1}{2}e_3^2\right\}\right]$$

(2) The bias of $t_2^{'}$ under design F_1 and F_2 is

(i) $$B\left(t_2^{'}\right)_I = \overline{Y}\beta_2\left(\delta_1 - \delta_2\right)\left(\frac{\beta_2+1}{2}C_X^2 - \rho C_Y C_X\right) \qquad \ldots(4.10)$$

(ii) $$B\left(t_2^{'}\right)_{II} = \overline{Y}\beta_2\left(\delta_4 - \delta_5\right)\left[\frac{1}{2}\left(\beta_2+1\right)C_X^2 - \rho C_Y C_X\right] \qquad \ldots(4.11)$$

Proof:

(i) $$B\left(t_2^{'}\right)_I = E\left[t_2^{'} - \overline{Y}\right]_I$$

$$= \overline{Y}\,E\left[1 + e_1 + \beta_2\left\{e_3 - e_2 + e_1e_3 - e_1e_2 - \beta_2e_2e_3 + \frac{\beta_2+1}{2}e_2^2 + \frac{\beta_2-1}{2}e_3^2\right\} - 1\right]$$

$$= \overline{Y}\beta_2\left(\delta_1 - \delta_2\right)\left(\frac{\beta_2+1}{2}C_X^2 - \rho C_Y C_X\right)$$

(ii) $$B\left(t_2^{'}\right)_{II} = E\left[t_2^{'} - \overline{Y}\right]_{II}$$

ISSN 1512-1232

$$= \overline{Y}\beta_2(\delta_4 - \delta_5)\left[\frac{1}{2}(\beta_2 + 1)C_X^2 - \rho C_Y C_X\right]$$

(3) Mean squared error of $t_2^{'}$ under design F_1 and F_2 respectively is:

(i) $$M(t_2^{'})_I = \overline{Y}^2\left[\delta_1 C_Y^2 + (\delta_1 - \delta_2)(\beta_2^2 C_X^2 - 2\beta_2 \rho C_Y C_X)\right] \quad \ldots(4.12)$$

(ii) $$M(t_2^{'})_{II} = \overline{Y}^2\left[\delta_4 C_Y^2 + (\delta_4 - \delta_5)\{\beta_2^2 C_X^2 - 2\beta_2 \rho C_Y C_X\}\right] \quad \ldots(4.13)$$

Proof: $$M(t_2^{'}) = E\left[t_2^{'} - \overline{Y}\right]^2 = \overline{Y}^2 E\left[1 + e_1 + \beta_2(e_3 - e_2) - 1\right]^2$$

$$= \overline{Y}^2 E\left[e_1^2 + \beta_2^2(e_3^2 + e_2^2 - 2e_2e_3) + 2\beta_2(e_1e_3 - e_1e_2)\right] \quad \ldots(4.14)$$

(i) Under F_1 (Using (4.14))

$$M(t_2^{'})_I = \overline{Y}^2\left[\delta_1 C_Y^2 + (\delta_1 - \delta_2)(\beta_2^2 C_X^2 - 2\beta_2 \rho C_Y C_X)\right]$$

(ii) Under F_2 (Using (4.14))

$$M(t_2^{'})_{II} = \overline{Y}^2\left[\delta_4 C_Y^2 + (\delta_4 - \delta_5)(\beta_2^2 C_X^2 - 2\beta_2 \rho C_Y C_X)\right]$$

(4) The minimum m.s.e. of $t_2^{'}$ is

(i) $$\left[M(t_2^{'})_I\right]_{Min} = \left[\delta_1 - (\delta_1 - \delta_2)\rho^2\right]S_Y^2 \quad \text{when} \quad \beta_2 = \rho\frac{C_Y}{C_X} \quad \ldots(4.15)$$

(ii) $$\left[M(t_2^{'})_{II}\right]_{Min} = \left[\delta_4 - (\delta_4 - \delta_5)\rho^2\right]S_Y^2 \quad \text{when} \quad \beta_2 = \rho\frac{C_Y}{C_X} \quad \ldots(4.16)$$

Proof:

(i) $$\frac{d}{d\beta_2}\left[M(t_2^{'})_I\right] = 0 \Rightarrow \beta_2 = \rho\frac{C_Y}{C_X}$$

$$\left[M(t_2^{'})_I\right]_{Min} = \left[\delta_1 - (\delta_1 - \delta_2)\rho^2\right]S_Y^2$$

(ii) $$\frac{d}{d\beta_2}\left[M(t_2^{'})_{II}\right] = 0 \quad \beta_2 = \rho\frac{C_Y}{C_X}$$

$$\left[M(t_2^{'})_{II}\right]_{Min} = \left[\delta_4 - (\delta_4 - \delta_5)\rho^2\right]S_Y^2$$

Theorem 4.3:

(1) The estimator $t_3^{'}$ in terms of e_1, e_2, e_3 and $e_3^{'}$ is

$$t_3^{'} = \overline{Y}\left[1 + e_1 + \beta_3\left\{e_3^{'} - e_2 - e_1e_2 + e_1e_3^{'} - \beta_3 e_2 e_3^{'} + \frac{\beta_3 + 1}{2}e_2^2 + \frac{\beta_2 - 1}{2}e_3^{'2}\right\}\right] \quad \ldots(4.17)$$

Proof: $$t_3^{'} = \overline{y}_r\left(\frac{\overline{x}^{'}}{\overline{x}_r}\right)^{\beta_3}$$

$$= \overline{Y}(1 + e_1)(1 + e_3^{'})^{\beta_3}(1 + e_2)^{-\beta_3}$$

$$= \overline{Y}(1 + e_1)\left(1 + \beta_3 e_3^{'} + \frac{\beta_3(\beta_3 - 1)}{2}e_3^{'2}\right)\left(1 - \beta_3 e_2 + \frac{\beta_3(\beta_3 + 1)}{2}e_2^2\right)$$

$$= \overline{Y}\left[1 + e_1 + \beta_3\left\{e_3^{'} - e_2 - e_1e_2 + e_1e_3^{'} - \beta_3 e_2 e_3^{'} + \frac{\beta_3 + 1}{2}e_2^2 + \frac{\beta_2 - 1}{2}e_3^{'2}\right\}\right]$$

(2) Bias of $t_3^{'}$ under design F_1 and F_2 is:

ISSN 1512-1232

(i) $$B\left(t_3^{'}\right)_I = \overline{Y}\beta_3\left(\delta_1-\delta_3\right)\left(\frac{\beta_3+1}{2}C_X^2-\rho C_Y C_X\right) \quad \ldots(4.18)$$

(ii) $$B\left(t_3^{'}\right)_{II} = \overline{Y}\beta_3\left[\frac{1}{2}\left\{\beta_3\left(\delta_4+\delta_3\right)-\left(\delta_3-\delta_4\right)\right\}C_X^2-\delta_4\rho C_Y C_X\right] \quad \ldots(4.19)$$

Proof:

(i) $$B\left(t_3^{'}\right)_I = E\left[t_3^{'}-\overline{Y}\right]_I$$

$$= \overline{Y}\,E\left[1+e_1+\beta_3\left\{e_3^{'}-e_2-e_1e_2+e_1e_3^{'}-\beta_3e_2e_3^{'}+\frac{\beta_1+1}{2}e_2^2+\frac{\beta_1-1}{2}e_3^{'2}\right\}-1\right]$$

$$= \overline{Y}\beta_3\left(\delta_1-\delta_3\right)\left(\frac{\beta_3+1}{2}C_X^2-\rho C_Y C_X\right)$$

(ii) $$B\left(t_3^{'}\right)_{II} = E\left[t_3^{'}-\overline{Y}\right]_{II}$$

$$= \overline{Y}\beta_3\left[\frac{1}{2}\left\{\beta_3\left(\delta_4+\delta_3\right)-\left(\delta_3-\delta_4\right)\right\}C_X^2-\delta_4\rho C_Y C_X\right]$$

(3) Mean squared error of $t_3^{'}$ is:

(i) $$M\left(t_3^{'}\right)_I = \overline{Y}^2\left[\delta_1C_Y^2+\left(\delta_1-\delta_3\right)\left(\beta_3^2C_X^2-2\beta_3\rho C_Y C_X\right)\right] \quad \ldots(4.20)$$

(ii) $$M\left(t_3^{'}\right)_{II} = \overline{Y}^2\left[\delta_4C_Y^2+\left(\delta_3+\delta_4\right)\beta_3^2C_X^2-2\delta_4\beta_3\rho C_Y C_X\right] \quad \ldots(4.21)$$

Proof: $$M\left(t_3^{'}\right) = E\left[t_3^{'}-\overline{Y}\right]^2 = \overline{Y}^2E\left[1+e_1+\beta_3\left(e_3^{'}-e_2\right)-1\right]^2$$

$$= \overline{Y}^2E\left[e_1^2+\beta_3^2\left(e_3^{'2}+e_2^2-2e_2e_3^{'}\right)+2\beta_3\left(e_1e_3^{'}-e_1e_2\right)\right] \quad \ldots(4.22)$$

(i) Under F_1 (Using (4.22))

$$M\left(t_3^{'}\right)_I = \overline{Y}^2\left[\delta_1C_Y^2+\left(\delta_1-\delta_3\right)\left(\beta_3^2C_X^2-2\beta_3\rho C_Y C_X\right)\right]$$

(ii) Under F_2 (Using (4.22))

$$M\left(t_3^{'}\right)_{II} = \overline{Y}^2\left[\delta_4C_Y^2+\left(\delta_3+\delta_4\right)\beta_3^2C_X^2-2\delta_4\beta_3\rho C_Y C_X\right]$$

(4) The minimum m.s.e. of $t_3^{'}$ is:

(i) $$\left[M\left(t_3^{'}\right)_I\right]_{\min} = \left[\delta_1-\left(\delta_1-\delta_3\right)\rho^2\right]S_y^2 \quad \ldots(4.23)$$

(ii) $$\left[M\left(t_3^{'}\right)_{II}\right]_{\min} = \left[\delta_4-\delta_4^2\left(\delta_3+\delta_4\right)^{-1}\rho^2\right]S_y^2 \quad \ldots(4.24)$$

Proof:

(i) By differentiating (4.20) with respect to β_3 and equate to zero

$$\frac{d}{d\beta_3}\left[M\left(t_3^{'}\right)_I\right] = 0 \Rightarrow \beta_3 = \rho\frac{C_Y}{C_X}$$

$$\left[M\left(t_3^{'}\right)_I\right]_{Min} = \left[\delta_1-\left(\delta_1-\delta_3\right)\rho^2\right]S_Y^2$$

(ii) By differentiating (4.21) with respect to β_3 and equate to zero

$$\frac{d}{d\beta_3}\left[M\left(t_3^{'}\right)_{II}\right] = 0 \Rightarrow \beta_3 = \left(\frac{\delta_4}{\delta_3+\delta_4}\right)\rho\frac{C_Y}{C_X}$$

ISSN 1512-1232

$$\left[M\left(t_3^{'}\right)_{II}\right]_{Min}=\left[\delta_4-\delta_4^2\left(\delta_3+\delta_4\right)^{-1}\rho^2\right]S_Y^2 .$$

5.0 COMPARISON OF ESTIMATORS

(i) $\Delta_1=\min\left[M\left(\bar{y}_{d1}\right)_I\right]-\min\left[M\left(\bar{y}_{d2}\right)_I\right]=\left(\frac{1}{r}-\frac{1}{N}\right)\rho^2S_Y^2$

$\left(\bar{y}_{d2}\right)_{\mathrm{I}}$ is better than $\left(\bar{y}_{d1}\right)_{\mathrm{I}}$, if $\Rightarrow N>r$ which is always true.

(ii) $\Delta_2=\min\left[M\left(\bar{y}_{d1}\right)_{II}\right]-\min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$=\left(\frac{\delta_4^2}{\left(\delta_3+\delta_4\right)}-\frac{\delta_5^2}{\left(\delta_3+\delta_5\right)}\right)\rho^2S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\mathrm{II}}$ is better than $\left(\bar{y}_{d1}\right)_{\mathrm{II}}$, if $\Delta_2>0$

$$\Rightarrow\left(n-r\right)\left[N^3-\left(n^{'}n+n^{'}r+nr\right)N+2n^{'}nr\right]>0$$

(A) when $\left(n-r\right)>0\Rightarrow n>r$ and

(B) $N^3-\left(n^{'}n+n^{'}r+nr\right)N+2n^{'}nr>0$

if $n^{'}\approx N\ \left[i.e.\ n^{'}\rightarrow N\right]$

then $N\left[N^2-\left(n-r\right)N+nr\right]>0$

$$\Rightarrow N^2-\left(n-r\right)N+nr>0$$

$$\Rightarrow\left(N-n\right)\left(N-r\right)>0$$

We get $\left(N-n\right)>0\Rightarrow N>n$ and $N-r>0\Rightarrow N>r$

The ultimate result is $N>n>r$, which is always true.

(iii) $\Delta_3=\min\left[M\left(\bar{y}_{d2}\right)_I\right]-\min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$=\frac{\left(\delta_1-\delta_4\right)\left(\delta_3+\delta_4\right)+\left(\delta_4^2+\delta_3^2-\delta_1\delta_3-\delta_1\delta_4+\delta_3\delta_4\right)}{\left(\delta_3+\delta_4\right)}\rho^2S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\mathrm{II}}$ is better than $\left(\bar{y}_{d2}\right)_{\mathrm{I}}$, if $\Delta_3>0$

$$\Rightarrow\rho^2>\left[\frac{1+m}{1+2m}\right]\text{ where }m=\left[\frac{r\left(N-n^{'}\right)}{n^{'}\left(N-r\right)}\right]$$

$$\Rightarrow -1<\rho<-\sqrt{\frac{1+m}{1+2m}}\ \text{ or }\ \sqrt{\frac{1+m}{1+2m}}<\rho<1 .$$

6.0 EMPIRICAL STUDY

The attached appendix A has a generated artificial database of size N = 200 containing values of main variable Y and auxiliary variable X. Parameter of this are given below in table 6.0.

Table 6.0 Dataset Parameters

$\bar{Y}$	$\bar{X}$	S_Y^2	S_X^2	ρ	C_X	C_Y	$V=\rho\frac{C_Y}{C_X}$
42.485	18.515	199.0598	48.5375	0.8652	0.3763	0.3321	0.7635

ISSN 1512-1232

Under design-I, we draw a preliminary random sample S' of size $n' = 110$ to compute $\bar{x}'$ and further draw a random sample *S* of size $n = 50$ such that $S \subset S'$ by *SRSWOR*. The *V* is a stable quantity which may assume to be known [see Reddy (1978)].

7.0 SIMULATION

The bias and optimum m.s.e. of estimators under both designs are computed based on 50,000 repeated samples *n*, n' as per design. These computations given in table 7.1 where efficiency measurement is considered as $E(\bar{y}_s)_t = \frac{M(\bar{y}_s)_t}{M(\bar{y})_w}$ with $M(\bar{y}_s)_t$ the mean squared error of estimator $\bar{y}_s$, $s = d_1, d_2; t = I, II$.

For design *I* and *II* the simulation procedure contains as following steps :

Step 1: Draw a random sample S' of size $n' = 110$ from the Dataset of *N* = 200 by SRSWOR.

Step 2: Again draw a random sub-sample of size $n = 50$ from S' for design *I* and independent sample *n*= 50 under design *II*.

Step 3: Drop down 5 units randomly from each sample corresponding to *Y* in both I, II.

Step 4: Compute and impute the dropped units of *Y* with the help of proposed methods and available methods.

Step 5: Repeat the above steps 50,000 times, which provides multiple sample based estimates $\hat{y}_1, \hat{y}_2, \hat{y}_3 ..... \hat{y}_{50000}$ for t_1', t_2' and t_3'.

Step 6: Bias of $\hat{y}_1$ is obtained by $B(\hat{y}_s)_t = \frac{1}{50000} \sum_{i=1}^{50000} \left[(\hat{y}_{is})_t - \bar{Y} \right]$.

Step 7: *M.S.E.* of $\hat{y}$ is computed by $M(\hat{y}_s)_t = \frac{1}{50000} \sum_{i=1}^{50000} \left[(\hat{y}_{is})_t - \bar{Y} \right]^2$.

Table 7.1

Estimator	Design F_1		Design F_2	
	Bias	***MSE***	***Bias***	***MSE***
t_1'	-0.1285	2.5936	0.3794	2.8056
t_2'	-0.1237	2.9321	0.6239	13.9020
t_3'	-0.2826	2.1707	0.3823	3.0217

8.0 CONCLUSIONS

The content of this chapter has a comparative approach for the three estimators examined under two different strategies of imputation in two-phase sampling under data mining environment. The estimator t_3' is better in terms of mean squared error than other estimators under design *I*. Moreover in design *II*, the estimator t_1' is found better over other estimators. All the methods of imputation are capable enough to replace the values of missing observations in data warehouse. Therefore, suggested strategies are good enough in application for data forming. These suggested methods replace the values not available in the database. Estimation procedure is scientific and based on sampling theory based approach.

ISSN 1512-1232

References

1. Ahmed, M. S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W. (2006): Estimation of a population mean using different imputation methods, Statistics in Transition, Vol.7, No.6, pp.1247-1264.
2. Cochran, W. G. (2005): Sampling Techniques, John Wiley and Sons, New York.
3. Rubin, D. B. (1976): Inference and missing data, Biometrica, Vol.63, pp.581-593.
4. Shukla, D. and Thakur, N. S. (2008): Estimation of mean with imputation of missing data using factor type estimator, Statistics in Transition, Vol.9, No.1, pp.33-48.
5. Shukla, D. and Thakur, N. S. (2010): Some Imputation Methods to Treat Missing Values in Knowledge Discovery in Data warehouse, IJDE, Vol1, No.2.
6. Han, J., and Kamber, M. (2000): Data Mining: Concepts and Techniques, San Mateo, CA: Morgan.Kaufmann.
7. Kalton, G., and Kasprzyk, D. (1982): Imputing for missing survey response, In Proc. Sect. Survey Res.Meth. Amer. Statist. Assoc., pp. 22–23.
8. Lee,R. C. T., Slagle, and Mong, C. T. (1976): Application of clustering to estimate missing data and improve data integrity. In Proc. Int'l Conf. Software Engineering, San Francisco, CA, IEEE Press, pp. 539–544.
9. Little, R. J. A., and Rubin, D. B. (1987): Statistical Analysis with Missing Data. New York, NY: John Wiley and Sons Publishers.
10. Liu, W. Z., White, A. P., Thompson, S. G. and Bramer, M. A. (1997): Techniques for dealing with missing values in classification, In 2nd Int. Symp. Intelligent Data Analysis, pp. 527–536.
11. Ragel, A., and Cremilleux, B. (1998): Treatment of missing values for association rules, In Proc. 2nd Pacific-Asia Conf. On Knowledge Discovery and Data Mining, pp. 258–270.
12. Ragel, A., and Cremilleux, B. (1999): MVC: A preprocessing method to deal with missing values,Knowledge-Base System Vol.12, No.5, pp.205–332.
13. Tseng, S.-M., and Kao, C.P. (2002): Efficient clustering methods for gene expression mining: A performance evaluation, In Proc. Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 432–437.
14. Tseng, S.M., Howang, k., and Lee, C.I.(2003): A pre-processing method to deal with missing values by integrating clustering and regression techniques, Applied Artificial Intelligence, taylor & Francis Group, Vol.17,pp.535–544.
15. Pyle, D.(1999): Data Preparation for Data Mining, Morgan Kaufmann, USA.
16. Liu, P., Lei, L., and Wu, N.,(2005): A quantitative study of the effect of missing data in classifiers, in CITOS: Proceedings of the 2005 Fifth international conference on Computer and Information Technology.
17. Fujikawa, Y.(2001): Efficient Algorithms for Dealing with Missing values in Knowledge Discovery, School of Knowledge Science – Japan Advanced Institute of Science and Technology. Japan, 2001.
18. Kantardzic, M.(2003): Data Mining - Concepts, Models, Methods and Algorithms, IEEE Press, USA.
19. Hofmann, M. and Tierney, B.(2003): The involvement of human resources in large scale data mining projects**,** in Proceedings of the 1st international symposium on Information and communication technologies, Ireland, pp. 103 - 109.

ISSN 1512-1232

Appendix A
Dataset (N = 200)

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19

Article received: 2011-01-05