Some Imputation Methods to Treat Missing Values in Knowledge Discovery in Data warehouse

D. Shukla diwakarshukla@rediffmail.com
Deptt. of Mathematics and Statistics,
Dr. H.S.G. Central University, Sagar (M.P.), India.

Rahul Singhai singhai_rahul@hotmail.com
Iinternational Institute of Professional Studies,
Devi Ahilya Vishwavidyalaya, Indore (M.P.) India.

Narendra Singh Thakur nst_stats@yahoo.co.in

B.T. Institute of Research and Technology,
Sironja, Sagar (M.P.) India.

Naresh Dembla nareshdembla@gmail.com
Iinternational Institute of Professional Studies,
Devi Ahilya Vishwavidyalaya, Indore (M.P.) India.

Abstract

One major problem in the data cleaning & data reduction step of KDD process is the presence of missing values in attributes. Many of analysis task have to deal with missing values and have developed several treatments to guess them. One of the most common method to replace the missing values is the mean method of imputation. In this paper we suggested a new imputation method by combining factor type and compromised imputation method, using two-phase sampling scheme and by using this method we impute the missing values of a target attribute in a data warehouse. Our simulation study shows that the estimator of mean from this method is found more efficient than compare to other.

Keywords: KDD (Knowledge Discovery in Databases), Data mining, Attribute, Missing values, Imputation methods, Sampling.

1. INTRODUCTION

“Data mining”, often also referred to as “Knowledge Discovery in Databases” (KDD), is a young sub-discipline of computer science aiming at the automatic interpretation of large datasets. The classic definition of knowledge discovery by Fayyad et al.(1996) describes KDD as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad et al. 1996). Additionally, they define data mining as “a step in the KDD process consisting of applying data analysis and discovery algorithms. In order to be able to “identify valid, novel patterns in data”, a step of pre-processing of the data is almost always required. This preprocessing has a significant impact on the runtime and on the results of the subsequent data mining algorithm.

The knowledge discovery in database is more than pure pattern recognition, Data miners do not simply analyze data, and they have to bring the data in a format and state that allows for this analysis. It has been estimated that the actual mining of data only makes up 10% of the time required for the complete knowledge discovery process (Pyle 1999). In our opinion, the precedent time-consuming step of preprocessing is of essential importance for data mining (Han and Kamber 2001). It is more than a tedious necessity: The techniques used in the preprocessing step can deeply influence the results of the following step, the actual application of a data mining algorithm (Hans et al.(2007). We therefore feel that the role of the impact on and the link of data preprocessing to data mining will gain steadily more interest over the coming years.

Thus Data pre-processing is one of the essential issue of KDD process in Data mining. Since data warehouse is a large database that contains data that is collected and integrated from multiple heterogeneous data sources. This may lead to irrelevant, noisy inconsistent, missing and vague data. So it is required to apply different data pre-processing techniques to improve the quality of patterns mined by data mining techniques. The data mining pre-processing methods are organised into four categories: Data cleaning, data integration and transportation, data reduction, descritization and concept hierarchy generation.

Since the goal of knowledge discovery can be vaguely characterized as locating interesting regularities from large databases (Fayyad et al. &. Krishnamurthy R. et al.) For large collections of data, sampling is a promising method for knowledge discovery: instead of doing complicated discovery processes on all the data, one first takes a small sample, finds the regularities in it, and then possibly validates these on the whole data

Sampling is a powerful data reduction technique that has been applied to a variety of problems in database systems. Kivinen and Mannila (1994) discuss the general applicability of sampling to data mining, and Zaki, et al.(1996) employ a simple random sample to identify association rules. Toivonen (1996) uses sampling to generate candidate itemsets but still requires a full database scan. John and Langley (1996) give a dynamic sampling method that selects the sample size based on the observed behavior of the data-mining algorithm. Traditionally, random sampling is the most widely utilized sampling strategy for data mining applications. According to the Chernoff bounds, the consistency between the population proportion and the sample proportion of a measured pattern can be probabilistically guaranteed when the sample size is large (Domingo et al.(2002) and Zaki et al.(1997)). Kun-Ta Chuang et al.(2007) proposed a novel sampling algorithm (PAS) to generate a high quality online sample with the desired sample rate.

Presence of missing data is one of the critical problem in data cleaning and data reduction approach. While using sampling techniques to obtain reduced representation of large database, it often possible that the sample may contains some missing values.Missing data are a part of most of the research, and missing data can seriously affect research results (Robert 1996). So, it has to be decided how to deal with it. If one ignores missing data or assumes that excluding missing data is acceptable, there is a risk of reaching invalid and non-representative conclusions. There are a number of alternative ways of dealing with missing data (Joop 1999). There are many methods of imputation (Litte and Rubin 1987) like Mean Imputation,regression imputation, Expectation maximization etc. Imputation of missing data minimizes bias and allows for analysis using a reduced dataset. In general the imputation methods can be classified into single & multiple imputations. The single imputation method always imputes the same value, thereby ignoring the variance associated with the imputation process. The multiple imputations method imputes several imputed values and the effect of the chosen imputed values on the variance can be taken into account.

Both the single-imputation and MI methods can be divided into three categories: 1) data driven; 2) model based; and 3) ML based (Laxminarayan et al.(1999), Little and Rubin(1987), Oh (1983)). Data-driven methods use only the complete data to compute imputed values. Model-based methods use some data models to compute imputed values. They assume that the data are generated by a model governed by unknown parameters. Finally, ML-based methods use the entire available data and consider some ML algorithm to perform imputation. The data-driven methods include simple imputation procedures such as mean, conditional mean, hot-deck, cold-deck, and substitution imputation (Laxminarayan et al. (1999), Sarle(1998)). Several model-based imputation algorithms are described by Little and Rubin (1987). The leading methods include regression-based, likelihood-based, and linear discriminant analysis (LDA)-based imputation. In regression-based methods, missing values for a given record are imputed by a regression model based on complete values of attributes for that record. The likelihood-based methods can be considered to impute values only for discrete attributes. They assume that the data are described by a parameterized model, where parameters are estimated by maximum likelihood or maximum a posteriori procedures, which use different variants of the EM algorithm (Cios(1998), Little and Rubin(1987)). A probabilistic imputation method that uses probability density estimates and Bayesian approach was applied as a preprocessing step for an independent module analysis system (Chan K et al.(2003)). Neural networks were used to implement missing data imputation methods (Freund and Schapire (1996), Tresp (1995)). An association rule algorithm, which belongs to the category of algorithms encountered in data mining, was used to perform MIs of discrete data (Zhang (2000)). Recently, algorithms of supervised ML were used to implement imputation. In this case, imputation is performed one attribute at a time, where the selected attribute is used as a class attribute. Several different families of supervised ML algorithms, such as decision trees, probabilistic, and decision rules (Cios et al.(1998)) can be used; however, the underlying methodology remains the same. For example, a decision tree C4.5 (Quinlan(1992),(1986), and a probabilistic algorithm A decision rule algorithm CLIP4 (Cios(1998)) and a probabilistic algorithm Naïve Bayes were studied in (Farhangfar et al.(2004). A k-nearest neighbor algorithm was used by Batista and Monard(2003). Backpropagation Neural Network (BPNN) is one of the most popular neural network learning algorithms. Werbos (1974) proposed the learning algorithm of the hidden layers and applied to the prediction in the economy. Classification is another important technique in data mining. A decision tree approach to classification problems were described by Friedman 1997. Let $A=\{x,y,z....\}$ is a finite attribute set of any database, where target attribute domain Y consist of $Y_i;(i=1,2,.......N)$ values of main interest and attribute domain X consist of $X_i;(i=1,2,.......N)$ auxiliary values, that is highly associated with attribute domain Y. Suppose target attribute Domain Y has some missing values. Let $\bar{y}$ be the mean of finite attribute set Y under consideration for estimation $\left[\bar{Y}=N^{-1}\sum_{i=1}^{N}Y_i\right]$ and $\bar{X}$ be the mean of reference attribute set X. When $\bar{X}$ is unknown, the two-phase sampling is used to estimate the main data set missing values (Shukla, 2002).

2. PROPOSED IMPUTATION TECHNIQUES FOR MISSING ATTRIBUTE VALUES

Consider preliminary large sample $S'=\{X_i;i=1,2,3,.....,n'\}$ of size n' drawn from attribute data set A by SRSWOR and a secondary sample of size n $(n<n')$ drawn in the following manner (fig. 1).

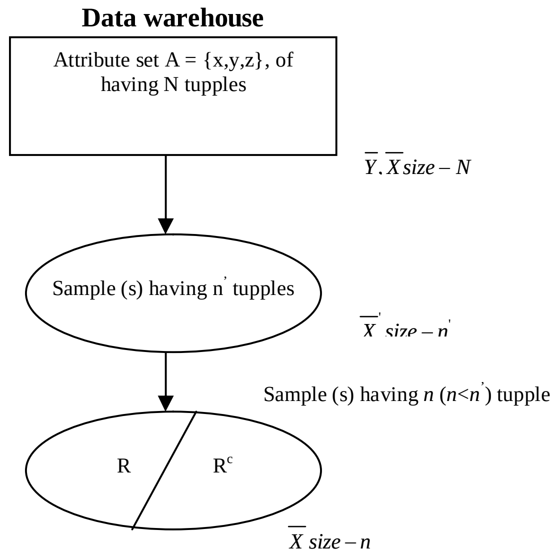
Data warehouse
Attribute set A = {x,y,z}, of having N tupples
$\bar{Y}.\bar{X}\,size-N$
Sample (s) having n' tupples
$\bar{X}'\,size-n'$
Sample (s) having n (n<n') tupple
R
R^c
$\bar{X}\,size-n$

FIGURE 1.

The sample S of n units contains r available values (r < n) forming a subspace R and (n – r) missing values with subspace R^C in $S = R \bigcup R^C$. For every $i \in R$, the y_i's are available values of attribute Y and for $i \in R^C$, the y_i values are missing and imputed values are to be derived, to replace these missing values.

2.1.0 F-T-C Imputation Strategies:

For $y_{ji}\left(j = 1,2,3\right)$

$$y_{ji} = \begin{cases} \left(\frac{kn}{r}\right) y_i + (1-k)\phi_j^{'}(k) & if\ i \in R \\ \left(1-k\right)\phi_j^{'}\left(k\right) & if \quad i \in R^C \end{cases} \qquad \ldots(2.1)$$

where, $\phi_1^{'}(k) = \bar{y}_r \left[\frac{(A+C)\bar{x}^{'} + fB\bar{x}}{(A+fB)\bar{x}^{'} + C\bar{x}}\right]$; $\phi_2^{'}(k) = \bar{y}_r \left[\frac{(A+C)\bar{x} + fB\bar{x}_r}{(A+fB)\bar{x} + C\bar{x}_r}\right]$;

$$\phi_3^{'}(k) = \bar{y}_r \left[\frac{(A+C)\bar{x}^{'} + fB\bar{x}_r}{(A+fB)\bar{x}^{'} + C\bar{x}_r}\right]; A = \left(k-1\right)\left(k-2\right);\ \ B = \left(k-1\right)\left(k-4\right);$$

$$C = \left(k-2\right)\left(k-3\right)\left(k-4\right);\ \ 0 < \mathrm{k} < \infty$$

2.1.1 Properties of $\phi_j\left(k\right)$:

(i) At k = 1; A = 0; B = 0; C = - 6

$$\phi_1^{'}\left(1\right) = \bar{y}_r \frac{\bar{x}^{'}}{\bar{x}}; \quad \phi_2^{'}\left(1\right) = \bar{y}_r \frac{\bar{x}}{\bar{x}_r}; \quad \phi_3^{'}\left(1\right) = \bar{y}_r \frac{\bar{x}^{'}}{\bar{x}_r}$$

(ii) At k = 2; A = 0; B = -2; C = 0

$$\phi_3^{'}\left(2\right) = \bar{y}_r \frac{\bar{x}}{\bar{x}^{'}}; \quad \phi_2^{'}\left(2\right) = \bar{y}_r \frac{\bar{x}_r}{\bar{x}}; \quad \phi_3^{'}\left(2\right) = \bar{y}_r \frac{\bar{x}_r}{\bar{x}^{'}}$$

(iii) At k = 3; A = 2; B = - 2; C = 0

$$\phi_1^{'}\left(3\right) = \bar{y}_r \left[\frac{\bar{x}^{'} - f\bar{x}}{\left(1-f\right)\bar{x}^{'}}\right]; \phi_2^{'}\left(3\right) = \bar{y}_r \left[\frac{\bar{x} - f\bar{x}_r}{\left(1-f\right)\bar{x}}\right]; \ \phi_3^{'}\left(3\right) = \bar{y}_r \left[\frac{\bar{x}^{'} - f\bar{x}_r}{\left(1-f\right)\bar{x}^{'}}\right]$$

(iv) At k = 4; A = 6; B = 0; C = 0

$$\phi_1^{'}\left(4\right) = \phi_2^{'}\left(4\right) = \phi_3^{'}\left(4\right) = \bar{y}_r$$

Theorem 2.1: The point estimate for S of $\bar{Y}$ are:

$$(\bar{y}_{FTC}^{'})_j = k\bar{y}_r + (1-k)\phi_j^{'}(k);\ j = 1,2,3 \qquad \ldots(2.2)$$

Proof: $\left(\bar{y}_{FTC}^{'}\right)_j = \left(\bar{y}_s\right)_j = \frac{1}{n}\sum_{i\in S}(y_{ji})$

$$= \frac{1}{n}\left[\sum_{i\in R}(y_{ji}) + \sum_{i\in R^c}(y_{ji})\right]$$

$$= \frac{1}{n}\left[\sum_{i\in R}\left\{\left(\frac{kn}{r}\right)y_i + (1-k)\phi_j^{'}(k)\right\} + \sum_{i\in R^c}(1-k)\,\phi_j^{'}(k)\right]$$

$$\left(\bar{y}'_{FTC}\right)_j = k\bar{y}_r + (1-k)\phi'_j(k); \qquad j = 1,2,3$$

2.2.0 Some Special Cases:

At $k=1$, $$\left(\bar{y}'_{FTC}\right)_j = \bar{y}_r \quad j=1,2,3 \qquad \ldots(2.3)$$

At $k=2$, $$\left(\bar{y}'_{FTC}\right)_1 = \bar{y}_r\left(2 - \frac{\bar{x}}{\bar{x}'}\right) \qquad \ldots(2.4)$$

$$\left(\bar{y}'_{FTC}\right)_2 = \bar{y}_r\left(2 - \frac{\bar{x}_r}{\bar{x}}\right) \qquad \ldots(2.5)$$

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{y}_r\left(2 - \frac{\bar{x}_r}{\bar{x}'}\right) \qquad \ldots(2.6)$$

At $k=3$, $$\left(\bar{y}'_{FTC}\right)_1 = \bar{y}_r\left(3 - \frac{2(\bar{x}' - f\bar{x})}{(1-f)\bar{x}'}\right) \qquad \ldots(2.7)$$

$$\left(\bar{y}'_{FTC}\right)_2 = \bar{y}_r\left(3 - \frac{2(\bar{x} - f\bar{x}_r)}{(1-f)\bar{x}}\right) \qquad \ldots(2.8)$$

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{y}_r\left(3 - \frac{2(\bar{x}' - f\bar{x}_r)}{(1-f)\bar{x}'}\right) \qquad \ldots(2.9)$$

At $k=4$, $$\left(\bar{y}'_{FTC}\right)_j = \bar{y}_r \qquad j=1,2,3 \qquad \ldots(2.10)$$

3. BIAS AND MEAN SQUARED ERROR

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

$$\bar{y}_r = \bar{Y}(1+e_1);\ \bar{x}_r = \bar{X}(1+e_1), \bar{x} = \bar{X}(1+e_3);\ \ \bar{x}' = \bar{X}(1+e'_3) \qquad \ldots(3.1)$$

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

$$\left.\begin{aligned} &E(e_1) = E(e_2) = E(e_3) = E(e'_3) = 0 \\ &E(e_1^2) = \delta_1 C_Y^2;\ E(e_2^2) = \delta_1 C_X^2;\ E(e_3^2) = \delta_2 C_X^2; E(e_3'^2) = \delta_3 C_X^2; \\ &E(e_1e_2) = \delta_1\rho C_Y C_X; E(e_1e_3) = \delta_2\rho C_Y C_X; E(e_1e'_3) = \delta_3\rho C_X C_Y; \\ &E(e_2e_3) = \delta_2 C_X^2;\ E(e_2e'_3) = \delta_3 C_X^2;\ E(e_3e'_3) = \delta_3 C_X^2 \end{aligned}\right\} \qquad \ldots(3.2)$$

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

Theorem 3.1: Estimator $\left(\bar{y}'_{FTC}\right)_j;\ \ j = 1,2,3$ in terms of $e_i, i = 1,2,3$ and e'_i could be expressed as:

(i) $$\left(\bar{y}'_{FTC}\right)_1 = \bar{Y}\left[1 + e_1 + (1-k)P\{e_3 - e'_3 + e_1e_3 - e_1e'_3 - (\theta_3 - \theta_4)e_3e'_3 - \theta_4 e_3^2 + \theta_3 e_3'^2\}\right] \qquad \ldots(3.3)$$

(ii) $$\left(\bar{y}'_{FTC}\right)_2 = \bar{Y}\left[1 + e_1 + (1-k)P\{e_2 - e_3 + e_1e_2 - e_1e_3 - (\theta_3 - \theta_4)e_2e_3 - \theta_4 e_2^2 + \theta_3 e_3^2\}\right] \qquad \ldots(3.4)$$

(iii) $$\left(\bar{y}'_{FTC}\right)_3 = \bar{Y}\left[1 + e_1 + (1-k)P\{e_2 - e'_3 + e_1e_2 - e_1e'_3 - (\theta_3 - \theta_4)e_2e'_3 - \theta_4 e_2^2 + \theta_3 e_3'^2\}\right] \qquad \ldots(3.5)$$

Proof :

(i) $\left(\bar{y}'_{FTC}\right)_1 = k\bar{y}_r + (1-k)\phi_1(k)$

Since

$$\phi_1'(k) = \bar{y}_r\left[\frac{(A+C)\bar{x}' + fB\bar{x}}{(A+fB)\bar{x}' + C\bar{x}}\right] = \bar{Y}(1+e_1)\left[\frac{(A+fB+C)+(A+C)e_3' + fBe_3}{(A+fB+C)+(A+fB)e_3' + Ce_3}\right]$$

$$= \bar{Y}(1+e_1)\left[\frac{1+\theta_1 e_3' + \theta_2 e_3}{1+\theta_3 e_3' + \theta_4 e_3}\right] = \bar{Y}(1+e_1)(1+\theta_1 e_3' + \theta_2 e_3)(1+\theta_3 e_3' + \theta_4 e_3)^{-1}$$

$$\left[Note:-Binomial\,theorem\,(1+\alpha e)^{-1} = 1-\alpha e+\alpha^2 e^2 - \alpha^3 e^3 + \ldots\ldots\ldots.\right]$$

$$= \bar{Y}(1+e_1)(1+\theta_1 e_3' + \theta_2 e_3)[1-(\theta_3 e_3' + \theta_4 e_3) + (\theta_3 e_3' + \theta_4 e_3)^2 + \ldots\ldots.]$$

$$\phi_1(k) = \bar{Y}\left[1+e_1+P\left\{e_3 - e_3' + e_1 e_3 - e_1 e_3' - (\theta_3-\theta_4)e_3 e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2\right\}\right]$$

Therefore,

$$(\bar{y}'_{FTC})_1 = \bar{Y}\left[1+e_1+(1-k)P\left\{e_3 - e_3' + e_1 e_3 - e_1 e_3' - (\theta_3-\theta_4)e_3 e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2\right\}\right]$$

(ii): $\left(\bar{y}'_{FTC}\right)_2 = k\bar{y}_r + (1-k)\phi_2(k)$

$$\phi_2(k) = \bar{Y}_r\left[\frac{(A+C)\bar{x} + fB\bar{x}_r}{(A+fB)\bar{x} + C\bar{x}_r}\right] = \bar{Y}(1+e_1)\left[\frac{1+\theta_1 e_3 + \theta_2 e_2}{1+\theta_3 e_3 + \theta_4 e_2}\right]$$

$$= \bar{Y}(1+e_1)\left[(1+\theta_1 e_3 + \theta_2 e_2)(1+\theta_3 e_3 + \theta_4 e_2)^{-1}\right]$$

$$= \bar{Y}(1+e_1)\Big[1+(\theta_1-\theta_3)e_3 + (\theta_2-\theta_4)e_2 + (\theta_4-\theta_2)\theta_4 e_2^2$$
$$-(\theta_2\theta_3 + \theta_1\theta_4 - 2\theta_3\theta_4)e_2 e_3 - (\theta_1-\theta_3)\theta_3 e_3^2\Big]$$

$$= \bar{Y}\left[1+e_1+P(e_2 - e_3 - \theta_4 e_2^2 + \theta_3 e_3^2 - (\theta_3-\theta_4)e_2 e_3 + e_1 e_2 - e_1 e_3)\right]$$

$$= \bar{Y}\left[1+e1+P(e_2 - e_3 + e_1 e_2 - e_1 e_3 - (\theta_3-\theta_4)e_2 e_3 - \theta_4 e_2^2 + \theta_3 e_3^2)\right]$$

Hence $\left(\bar{y}_{FTC}\right)_2 = \bar{Y}\left[(1+e_1)+(1-k)P(e_2 - e_3 + e_1 e_2 - e_1 e_3 - (\theta_3-\theta_4)e_2 e_3 - \theta_4 e_2^2 + \theta_3 e_3^2)\right]$

(iii) : $\left(\bar{y}_{FTC}\right)_3 = k\bar{y}_r + (1-k)\phi_3(k)$

$$\phi_3(k) = \bar{y}_r\left[\frac{(A+C)\bar{x}' + fB\bar{x}_r}{(A+fB)\bar{x}' + \bar{x}_r}\right] = \bar{Y}(1+e_1)\left[(1+\theta_1 e_3'\theta_2 e_2)(1+\theta_3 e_3' + \theta_4 e_2)\right]$$

$$= \bar{Y}(1+e_1)\left[1-Pe_3' + Pe_2 + P\theta_3 e_3'^2 - P\theta_4 e_2^2 - P(\theta_3-\theta_4)e_2 e_3'\right]$$

$$= \bar{Y}\Big[1+P(e_2 - e_3' + \theta_3 e_3'^2 - \theta_4 e_2^2 - (\theta_3-\theta_4)e_2 e_3') + e_1$$
$$+P(e_1 e_2 - e_1 e_3' + \theta_3 e_1 e_3'^2 - \theta_4 e_1 e_2^2 - (\theta_3-\theta_4)e_1 e_2 e_3')\Big]$$

$$= \bar{Y}\left[1+e_1+P(e_2 - e_3' + e_1 e_2 - e_1 e_3' - (\theta_3-\theta_4)e_2 e_3' - \theta_4 e_2^2 + \theta_3 e_3'^2)\right]$$

Hence,

$$\left(\bar{y}'_{FTC}\right)_3 = \bar{Y}\left[(1+e_1)+(1-k)P(e_2 - e_3' + e_1 e_2 - e_1 e_3' - (\theta_3-\theta_4)e_2 e_3' - \theta_4 e_2^2 + \theta_3 e_3'^2)\right]$$

Theorem (3.2): The bais of the estimators $\left(\bar{y}'_{FTC}\right)_j$ is given by

(i) $$B\left[\left(\bar{y}'_{FTC}\right)_1\right] = -\bar{Y}P(1-k)(\delta_2-\delta_3)\left[\theta_4 C^2{}_X - \rho C_Y C_X\right]$$

(ii) $$B\left[\left(\bar{y}'_{FTC}\right)_2\right] = -\bar{Y}(1-k)P(\delta_1-\delta_2)\left[\theta_4 C_x^2 - \rho C_Y C_X\right]$$

(iii) $$B\left[\left(\bar{y}'_{FTC}\right)_3\right] = -\bar{Y}(1-k)P(\delta_1-\delta_3)\left[\theta_4 C_X^2 - \rho C_Y C_X\right]$$

Proof:

(i): $$B\left[\left(\bar{y}'_{FTC}\right)_1\right] = E\left[\left(\bar{y}'_{FTC}\right)_1 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{1+e_1+(1-k)P(e_3-e_3^{'}+e_1e_3-e_1e_3^{'}-(\theta_3-\theta_4)e_3e_3^{'}-\theta_4e_3^2+\theta_3e_3^{'2})\right\}-\bar{Y}\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_2-\delta_3)\rho C_Y C_X - \left\{(\theta_3-\theta_4)\delta_3+\theta_4\delta_2\right\}C_X{}^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_2-\delta_3)\rho C_Y C_X - (\delta_2-\delta_3)\theta_4 C_X^2\right]$$

$$= -\bar{Y}P(1-k)(\delta_2-\delta_3)\left[\theta_4 C^2{}_X - \rho C_Y C_X\right] \qquad \ldots(3.6)$$

(ii) $$B\left[\left(\bar{y}'_{FTC}\right)_2\right] = E\left[\left(\bar{y}'_{FTC}\right)_2 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{1+e_1+(1-k)P(e_2-e_3+e_1e_2-e_1e_3-(\theta_3-\theta_4)e_2e_3-\theta_4e_2^2+\theta_3e_3^2)\right\}-\bar{Y}\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - \left\{(\theta_3-\theta_4)\delta_2+\theta_4\delta_1-\theta_3\delta_2\right\}C_X^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - \left\{\theta_3\delta_2-\theta_4\delta_2+\theta_4\delta_1-\theta_3\delta_2\right\}C_X^2\right]$$

$$= \bar{Y}(1-k)P\left[(\delta_1-\delta_2)\rho C_Y C_X - (\delta_1-\delta_2)\theta_4 C_X^2\right]$$

$$= -\bar{Y}(1-k)P(\delta_1-\delta_2)\left[\theta_4 C_x^2 - \rho C_Y C_X\right] \qquad \ldots(3.7)$$

(iii) $$B\left[\left(\bar{y}'_{FTC}\right)_3\right] = E\left[\left(\bar{y}'_{FTC}\right)_3 - \bar{Y}\right]$$

$$= E\left[\bar{Y}\left\{(1+e_1)+(1-k)P(e_2-e_3^{'}+e_1e_2-e_1e_3^{'}-(\theta_3-\theta_4)e_2e_3^{'}-\theta_4e_2^2+\theta_3e_3^{'2})\right\}-\bar{Y}\right]$$

$$= \bar{Y}\left[(1-k)P\left[(\delta_1-\delta_3)\rho C_Y C_X - \left\{(\theta_3-\theta_4)\delta_3+\theta_4\delta_1-\theta_3\delta_3\right\}C_x^2\right]\right]$$

$$= -\bar{Y}(1-k)P(\delta_1-\delta_3)\left[\theta_4 C_X^2 - \rho C_Y C_X\right] \qquad \ldots(3.8)$$

Theorem 3.3: The m.s.e. of the estimators $\left(\bar{y}'_{FTC}\right)_j$ is given by:-

(i) $$M\left[(\bar{y}'_{FTC})_1\right] = \bar{Y}\left[\delta_1 C_Y^2 + (1-k)^2P^2(\delta_2-\delta_3)C_x^2 + 2(1-k)P(\delta_2-\delta_3)e^{C_YC_X}\right] \qquad \ldots(3.9)$$

(ii) $$M\left[(\bar{y}'_{FTC})_2\right] = \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2P^2(\delta_1-\delta_2)C_X^2 + 2(1-k)P(\delta_1-\delta_2)\rho C_Y C_X\right] \ldots(3.10)$$

(iii) $$M\left[(\bar{y}'_{FTC})_3\right] = \bar{Y}^2\left[\delta_1 C_Y^2 + (1-k)^2P^2(\delta_1-\delta_3)C_X^2 + 2(1-k)P(\delta_1-\delta_3)\rho C_Y C_X\right] \ldots(3.11)$$

Proof:

(i): $$M\left[(\bar{y}'_{FTC})_1\right] = E\left[(\bar{y}_{FTC})_1 - \bar{Y}\right]^2$$

Using equation (3.3)

$$= \bar{Y}^2 E\left[e_1+(1-k)P\left\{e_3-e_3^{'}+e_1e_3-e_1e_3^{'}-(\theta_3-\theta_4)e_3e_3^{'}-\theta_4e_3^2+\theta_3e_3^{'2}\right\}\right]^2$$

$$=\bar{Y}^2 E\left[e_1+(1-k)P(e_3-e_3')\right]^2$$
$$=\bar{Y}^2 E\left[e_1^2+(1-k)^2P^2(e_3-e_3')^2+2(1-k)P(e_3-e_3')e_1\right]$$
$$=\bar{Y}\left[\delta_1 C_Y^2+(1-k)^2P^2(\delta_2-\delta_3)C_x^2+2(1-k)P(\delta_2-\delta_3)\rho C_Y C_X\right]$$

(ii) $M\left[(\bar{y}'_{FTC})_2\right]=E\left[(\bar{y}'_{FTC})_2-\bar{Y}\right]^2$

From using equation (3.4)

$$=E\left[Y\left\{1+e_1+(1-k)P\left\{e_2-e_3+e_1e_2-e_1e_3-(\theta_3-\theta_4)e_2e_3-\theta_4e_2^2+\theta_3e_3^2\right\}\right\}-\bar{Y}\right]^2$$
$$=\bar{Y}^2 E\left[e_1^2+(1-k)^2P^2(e_2-e_3)^2+(1-k)P(e_2-e_3)e_1\right]$$
$$=\bar{Y}^2 E\left[e_1^2+(1-k)^2P^2(e_2^2+e_3^2-2e_2e_3)+2(1-k)P(e_1e_2-e_1e_3)\right]$$
$$=\bar{Y}^2\left[\delta_1C_Y^2+(1-k)^2P^2(\delta_1-\delta_2)C_X^2+2(1-k)P(\delta_1-\delta_2)\rho C_Y C_X\right]$$

(iii) $M\left[\left(\bar{y}'_{FTC}\right)_3\right]=E\left[\left(\bar{y}'_{FTC}\right)_3-\bar{Y}\right]^2$

$$=\bar{Y}^2 E\left[e_1+(1-k)P\left\{e_2-e_3'\right\}\right]^2$$
$$=\bar{Y}^2 E\left[e_1^2+(1-k)^2P^2\left\{e_2-e_3'\right\}+2(1-k)P(e_2-e_3')e_1\right]^2$$
$$=\bar{Y}^2\left[\delta_1C_Y^2+(1-k)^2P^2(\delta_1-\delta_3)C_X^2+2(1-k)P(\delta_1-\delta_3)\rho C_Y C_X\right]$$

Theorem 3.4: The minimum m.s.e of $\left(\bar{y}'_{FTC}\right)_j$ is

(i) $M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{mim}=\left[\delta_1-(\delta_2-\delta_3)\rho^2\right]S_Y^2$ …(3.13)

(ii) $M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min}=\left[\delta_1-(\delta_1-\delta_2)\rho^2\right]S_Y^2$ …(3.14)

(iii) $M\left[\left(\bar{Y}'_{FTC}\right)_3\right]_{\min}=\left[\delta_1-(\delta_1-\delta_3)\rho^2\right]S_Y^2$ …(3.15)

Proof:

(i): $\dfrac{d}{d(1-k)P}M\left[(\bar{y}'_{FTC})_1\right]=0$

From equation (3.9)

$\Rightarrow$ $(1-k)PC_X+\rho C_Y=0$ $\Rightarrow$ $(1-k)P=-\rho\dfrac{C_y}{C_x}$

Therefore from equation (3.9). we have

$M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{\min}$ $=\bar{Y}^2\left[\delta_1C_Y^2-(\delta_2-\delta_3)\rho^2C_Y^2\right]$ $\because C_Y^2=\left(\dfrac{S_Y}{\bar{Y}}\right)^2$

Therefore

$M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{mim}=\left[\delta_1-(\delta_2-\delta_3)\rho^2\right]S_Y^2$

(ii) $\dfrac{d}{d\left[(1-k)P\right]}M\left[\left(\bar{y}'_{FTC}\right)_2\right]=0$

From equation (3.10)

$\Rightarrow$ $(1-k)PC_x+\rho C_Y=0\Rightarrow$ $(1-k)P=-\rho\dfrac{C_Y}{C_X}$

Therefore

$$M\left[\left(\bar{y}'_{FTC}\right)^{2}\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_2)\rho^2\right]S_Y^2$$

(iii) $\frac{d}{d\left[(1-k)P\right]} M\left[\left(\bar{y}'_{FTC}\right)_3\right] = 0$ From equation (3.11)

$\Rightarrow$ $(1-k)P = -\rho\frac{C_Y}{C_X}$ ...(3.16)

Therefore $M\left[\left(\bar{Y}'_{FTC}\right)_3\right]_{\min} = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2$

3.1 Multiple Choices of k :

The optimality condition $P = -V$ provides the equation

$$k^4 - (f - V)k^3 - \left[(4f + 15) - (f - 8)V\right]k^2 + \left[(f - 10) - (5f - 23)V\right]k + \left[(4f + 24) + (4f - 22)V\right] = 0 \qquad \text{...(3.17)}$$

which fourth degree polynomial in terms of k. One can get at most four values of k like k_1, k_2, k_3, k_4 for which m. s. e. is optimal. The best choice criteria is

Step I: Compute $\left|B(T_{FTi})_{k_j}\right|$ for i = 1, 2, 3; j = 1, 2, 3, 4.

Step II: For given i, choose k_j as $\left|B(T_{FTi})_{k_j}\right| = \overset{\min}{j = 1,2,3,4}\left[\ \left|B(T_{FTi})_{k_j}\right|\ \right]$

This ultimately gives bias control at the optimal level of m.s.e.

Note 3.1: For given pair of values of (V, f), $0 < V < \infty$; $0 < f < 1$, one can generate a trivariate table of k_1, k_2, k_3, k_4 so as to achieve solution quickly.

Remark 3.2: Reddy (1978) has shown that quantity $V = \rho\frac{C_Y}{C_X}$ is stable over moderate length time period and could be priorly known or guessed by past data. Therefore, pair (f, V) be treated as known and equation (3.13) generates maximum of four roots (some may imaginary) on which optimum level of m.s.e. will be attained.

4. COMPARISON

(i) Let $D_1 = M\left[(\bar{y}'_{FTC})_1\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min} = [\delta_1 - 2\delta_1 + \delta_3]\rho^2\delta_Y^2$

Thus $\left(\bar{y}'_{FTC}\right)_2$ is better than $\left(\bar{y}'_{FTC}\right)_1$ if:

$D_1 > 0 \Rightarrow [\delta_1 - 2\delta_2 + \delta_3]e^2\delta_Y^2 > 0 \Rightarrow \delta_1 - 2\delta_2 + \delta_3 > 0$ ...(4.1)

(ii) Let $D_2 = M\left[\left(\bar{y}'_{FTC}\right)_1\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_3\right]_{\min} = [-\delta_2 + \delta_3 + \delta_1 - \delta_3]\rho^2\delta_Y^2$

$= (\delta_1 - \delta_2)\rho^2\delta_Y^2$

Thus $\left(\bar{y}'_{FTC}\right)_3$ better than $\left(\bar{y}'_{FTC}\right)_1$ if

$D_2 > 0 \Rightarrow (\delta_1 - \delta_2)\rho 2\delta_Y^2 > 0 \Rightarrow \frac{1}{r} - \frac{1}{n} > 0 \Rightarrow \frac{1}{r} > \frac{1}{n} \Rightarrow n > r$ ...(4.2)

i.e. the size of sample domain is greater than the size of auxiliary data.

(iii) $D_3 = M\left[\left(\bar{y}'_{FTC}\right)_2\right]_{\min} - M\left[\left(\bar{y}'_{FTC}\right)_3\right]_{\min} = [(\delta_2 - \delta_3)\rho^2]\delta_Y^2 \quad = (\delta_2 - \delta_3)\rho^2\delta_Y^2$

Thus $\left(\bar{y}'_{FTC}\right)_3$ is better than $\left(\bar{y}'_{FTC}\right)_2$ if

$D_3 > 0 \Rightarrow (\delta_2 - \delta_3) > 0 \Rightarrow \delta_1 -> \delta_3 \Rightarrow \frac{1}{n} - \frac{1}{n'} > \frac{1}{n'} - \frac{1}{N}$ If $n' \to N$

Then $\frac{1}{n} - \frac{1}{N} > 0 \Rightarrow \frac{1}{n} > \frac{1}{N} \Rightarrow N > n$ …(4.3)

i.e. the size of total data set is greater than the size of sample data set.

5. EMPIRICAL STUDY

The attached appendix A has generated artificial population of size N = 200 containing values of main variable Y and auxiliary variable X. Parameter of this are given below:

$\bar{Y}$ = 42.485; $\bar{X}$ = 18.515; S_Y^2 = 199.0598; S_X^2 = 48.5375; ρ = 0.8652; C_X = 0.3763; C_Y = 0.3321. Using random sample SRSWOR of size n = 50; r = 45; f = 0.25, α = 0.2365. Solving optimum condition $\theta = -V$ [see (3.13)] the equation of power four in k provides only two real values k_1 = 0.8350; k_2 =4.1043. Rest other two roots appear imaginary.

6. SIMULATION

The bias and optimum m.s.e. of proposed estimators under both designs are computed through 50,000 repeated samples n, n' as per design. Computations are in table 6.1.

The simulation procedure has following steps :

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

Step 2: Draw a random sub-sample of size $n = 50$ from S'.

Step 3: Drop down 5 units randomly from each second sample corresponding to Y.

Step 4: Impute dropped units of Y by proposed methods and available methods and compute the relevant statistic.

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

Step 6: Bias of $\hat{y}$ is $B(\hat{y}) = \frac{1}{50000}\sum_{i=1}^{50000}\left[\hat{y}_i - \bar{Y}\right]$

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

Table 6.1 : Comparisons of Estimators

Estimator	***Bias* (.)**	***M*(.)**
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_1}$	0.3313	13.5300
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_2}$	0.0489	3.4729
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_3}$	---	---
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_1}$	0.2686	4.6934

$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_2}$	0.0431	3.2194
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_3}$	---	---
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_1}$	0.5705	14.6633
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_2}$	0.0639	3.5274
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_3}$	---	---

TABLE 1: Bias and Optimum m.s.e. at $k = k_i\ (i = 1,2)$

7. CONCLUDING REMARKS

The content of this paper has a comparative approach for the three estimators examined under two-phase sampling. The estimator $\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_2}$ is best in terms of mean squared error than other estimators. We can also choose an appropriate value of k for minimum bias from available values of k. Equation (4.1), (4.2) and (4.3) shows the general conditions for showing better performance of any estimator. All suggested methods of imputation are capable enough to obtain the values of missing observations in data warehouse. These methods are useful in the case where two attributes are in quantitative manner and linearly correlate with each other, like, Statistical Database, agricultural database (yield and area under cultivation), banking database (saving and interest),Spatial Databases etc. Therefore, suggested strategies are found very effective in order to replace missing values during the data preprocessing in KDD, so that the quality of the results or patterns mined by data mining methods can be improved.

8. REFERENCES

[1]. U Fayyad, Piatetsky-Shapiro, P.Smyth. ”Knowledge discovery and data mining: Towards a unifying framework”,In Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD), Portland, OR, pp 82–88.1996.

[2]. Piatetsky, Shapiro and J.William, Frawley. “Knowledge discovery in databases”,AAAI Press/MIT Press,1991.

[3]. R.Krishnamurthy, and T.Imielinski. “Research directions in Knowledge Discovery”, SIGMOD Record,20(3):76-78,1991.

[4]. D.Pyle. “Data preparation for data mining”, Morgan Kaufmann Publishers Inc, (1999).

[5]. J. Han, M. Kamber. “Data mining: concepts and techniques”, Academic Press, San Diego, (2001).

[6]. H. P. Kriegel, Karsten, M. Borgwardt, P. Kröge, A. Pryakhin, M. Schubert, A. Zimek, “Future trends in data mining”, Data Min Knowl Disc 15:87–97 DOI 10.1007/s10618-007-0067-9,2007.

[7]. J. Kivinen and H.Mannila. “The power of sampling in knowledge discovery”, In Proc. Thirteenth ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Sys., pages 77–85. ACM Press,1994.

[8]. M. J. Zaki, S. Parthasarathy, W. Lin, and M. Ogihara. “Evaluation of sampling for data mining of association rules”, Technical Report 617, University of Rochester, Rochester, NY,1996.

[9]. H. Toivonen. “Sampling large databases for association rules”, In Proc. 22nd VLDB 1996.

[10]. G. H. John and P. Langley. “Static versus dynamic sampling for data mining”, In Proc. Second Intl. Conf. Knowledge Discovery and Data Mining, pages 367–370. AAAI Press,1996.

[11]. C. Domingo, R. Gavalda and Q. Watanabe. “Adaptive Sampling Methods for Scaling Up Knowledge Discovery Algorithms”, Data Mining and Knowledge Discovery,2002.

[12]. M. Zaki, S. Parthasarathy, W. Li and M. Ogihara. “Evaluation of Sampling for Data Mining of Association Rules”, Proc. Int’l Workshop Research Issues in Data Eng,1997.

[13]. K.T. Chuang, K. P. Lin, and M. S. Chen. “Quality-Aware Sampling and Its Applications in Incremental Data Mining”, IEEE Transactions on knowledge and data engineering,vol.19, no. 4,2007.

[14]. K.Lakshminarayan, S. A. Harp and Samad. “Imputation of missing data in industrial databases, Appl. Intell., vol. 11, no. 3, pp. 259–275, Nov./Dec1999.

[15]. R. J. Little and D. B. Rubin. “Statistical Analysis With Missing Data”, Hoboken, NJ: Wiley, (1987).

[16]. H. L. Oh, and F. L. Scheuren. “Weighting adjustments for unit nonresponse, incomplete data in sample survey”, in Theory and Bibliographies, vol. 2, W. G. Madow, I. Olkin, and D. B. Rubin, Eds. New York: Academic, pp. 143–183,1983.

[17]. W. S. Sarle. “Prediction with missing inputs”, in Proc. 4th JCIS, vol. 2, pp. 399–402,1998.

[18]. K. J. Cios, W. Pedrycz, ,and R. Swiniarski. “Data Mining Methods for Knowledge Discovery”,Norwell, MA: Kluwer,(1998).

[19]. K. Chan, T. W. Lee, and T. J. Sejnowski. “Variational Bayesian learning of ICA with missing data, Neural Comput”, vol. 15, no. 8, pp. 1991–2011,2003.

[20]. Y. Freund and R. E. Schapire. “Experiments with a new boosting algorithm”, in Proc. 13th Int. Conf. Mach. Learn., pp. 146–148,1996.

[21]. V. Tresp, R. Neuneier, and S. Ahmad. “Efficient methods for dealing with missing data in supervised learning”, in Advances in Neural Information Processing Systems 7, G. Cambridge, MA: MIT Press, pp. 689–696,1995.

[22]. W. Zhang. “Association based multiple imputation in multivariate datasets”, A summary, in Proc. 16th ICDE, pp. 310–311,2000.

[23]. J. R. Quinlan. ”C4.5: Programs for Machine Learning”, San Mateo, CA: Morgan Kaufmann,1992.

[24]. J. R. Quinlan. “Induction of decision trees, Mach. Learn”, vol. 1, no. 1, pp. 81–106, 1986.

[25]. A. Farhangfar, L. A. Kurgan, and W. Pedrycz. “Novel framework for imputation of missing values in databases”, Comput.: Theory and Appl. II Conf., Conjunction with SPIE Defense and Security Symp. (formerly AeroSense), Orlando, FL, pp. 172–182,2004.

[26]. G. Batista and M. Monard. “An analysis of four missing data treatment methods for supervised learning”, Appl. Artif. Intell., vol. 17, no. 5/6, pp. 519–533,2003

[27]. W. G. Cochran. “Sampling Techniques”, John Wiley and Sons, New York, (2005).

[28]. D. F. Heitjan and S. Basu. “Distinguishing ‘Missing at random’ and ‘missing completely at random”, The American Statistician, 50, 207-213,1996.

[29]. V. N. Reddy. "A study on the use of prior knowledge on certain population parameters in estimation", Sankhya, C, 40, 29-37,1978.

[30]. D. Shukla. "F-T estimator under two-phase sampling", Metron, 59, 1-2, 253-263,2002.

[31]. S. Singh, and S. Horn. "Compromised imputation in survey sampling", Metrika, 51, 266-276,2000.

[32]. Li.Liu, Y. Tu, Y. Li, and G. Zou. "Imputation for missing data and variance estimation when auxiliary information is incomplete", Model Assisted Statistics and Applications, 83-94,2005.

[33]. S.Singh. "A new method of imputation in survey sampling", Statistics, Vol. 43, 5 , 499 – 511,2009.

Appendix A (Artificial 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