International Journal of Chemical Studies

P-ISSN: 2349–8528
E-ISSN: 2321–4902
www.chemijournal.com
IJCS 2020; 8(6): 1467-1471

Received: 11-09-2020
Accepted: 27-10-2020

**Srikanth Bairi**
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

**AR Rao**
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

**Corresponding Author:**
**Srikanth Bairi**
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

# Identification of a suitable technique for imputation of incomplete genotyping by sequencing (GBS) data

**Srikanth Bairi and AR Rao**

**DOI:** https://doi.org/10.22271/chemi.2020.v8.i6u.10967

**Abstract**
In the field of DNA sequencing, Genotype by sequencing is to discover SNPs in order to perform Genotyping studies. A most commonly occurring problem in GBS is the presence of missing observations. Quite often, the standard statistical models may not handle such missing data situation also known as incomplete data situations. An alternative to deal with incomplete data situation is to impute missing data for further downstream analysis. Hence a study is conducted with the objectives (i) to impute missing GBS data by various imputation techniques, based on both supervised and unsupervised learning algorithms at different levels of missingness, (ii) to identify suitable imputation technique to deal with incomplete GBS data situation. Based on correlation coefficient and mean squared prediction error (MSPE) between imputed value and true response, the accuracy of imputation technique was assessed. Different imputation techniques, viz., Mean Allele Frequency Imputation (MNI), Singular Value Decomposition Imputation (SVDI), k-Nearest Neighbour Imputation (kNNI), locally weighted linear regression imputation (LWI), Expectation Maximization Imputation (EMI) and Random Forest Imputation (RFI) were applied on incomplete GBS data of mice, a model animal organism, to assess their performance. The results revealed that RFI was found to be most accurate imputation technique. Besides, the performance of RFI in terms correlation coefficient at 5%, 10%, 15% and 20% missing data situation was observed to be 0.778, 0.765, 0.750 and 0.735 respectively. A Similar trend was also observed for RFI in terms of mean square prediction errors. Thus, it is suggested to use RFI technique to deal with incomplete GBS data situation and prior to the application of genomic selection models for breeding value estimation.

**Keywords:** Incomplete data, Genotyping by sequencing, Imputation techniques, prediction error

## Introduction
A milestone for genomic studies in molecular biology is largely concerned with understanding how DNA regions regulating the chemical processes leading in controlling the traits of an organism. Hence association of different regions or order of nucleotides of the DNA involved in controlling traits by regulating several chemical processes called ‘genes’ (Wilhelm Johannsen, 1909) [6] and totality of single copy of all genes called genome (Hans winkler, 1920) is of highest importance in selection processes.

In recent past Elshire *et al*. (2011) [4] described called genotyping by sequencing (GBS), to discover single nucleotide polymorphisms (SNPs) in order to perform genotyping studies. GBS is a prominently vigorous, modest, and inexpensive procedure for SNP discovery and mapping. GBS is gaining popularity is it provides genome wide marker coverage (Poland and rife, 2012) [10, 11]. GBS data with SNP information of genotypes generally containing missing values, because i) the fraction of the chromosomal segment which is re-sequenced is not exactly the same between two individuals. ii) Random fragments of the genome are sequenced at low depth. The proportion of missing data mainly based on two factors: (a) Depth or coverage of sequencing and (b) library complexity (Sims *et al*. 2014) [16]. Coverage (or depth) in DNA sequencing is the number of reads that include a given nucleotide in the reconstructed sequence. Whereas library complexity is number of unique molecules in library. The unique molecules number is inversely proportional to the library complexity (Elshire *et al.* 2011) [4].

Most of the statistical models require a complete data set and therefore marker imputation is necessary step before the data can be used for further downstream analysis for true estimation

of breeding values in genomic selection. Imputation is a statistical technique that is often used to increase the power and resolution of genomic association studies without a major loss in accuracy of breeding values estimated through GS models (Marchini *et al*, 2007; Marchini and Howei, 2010) [8, 7].
Several statistical imputation methods are used to make incomplete as complete data. Methods are like k- Nearest Neighbours imputation (k-NNI) based on k-nearest neighbour algorithm, Singular Value Decomposition Imputation (SVDI) based on singular value decomposition algorithm, Mean Allele Frequency Imputation (MNI) based on mean allele frequency, Locally weighted linear regression imputation (LWI) based on regression technique and Expectation maximization Imputation (EMI) based on Expectation maximization algorithm and random forest imputation technique (RFI) based on random forest algorithm.

## Materials and methods

The mice data set freely available in BGLR package of R software was considered initially. The data set consists of 1814 genotypes, each genotyped for 10346 polymorphic markers for body mass index (BMI) (Valdar *et al*. 2006a; 2006b) [20, 21]. To assess the performance of different imputation techniques, various levels of missingness viz, 5%, 10%, 15% and 20% were created by randomly generating missing observations from data set. This process was repeated 100 times at each level of missingness.
The imputation techniques described below were used for assessing their performance in imputing the missing values.

### Mean allele frequency Imputation (MNI)

Here, the missing value of any genotype for a given marker is imputed by the mean allele frequency of that marker on remaining genotypes. This method of imputation maintains the sample size and is easy to use, but the variance estimate gets underestimated (Charmet G, 2020) [3]. MNI function was used to impute missing values from BWGS package of R 3.6.0 software.

### Singular value decomposition imputation (SVDI)

This method obtains a set of mutually orthogonal (independent) marker patterns that can be linearly combined to approximate the effect of all markers in the data set. These patterns are identical to the principle components of the marker data matrix and are further referred to as eigengenes (Alter *et al.*, 2000) [1].

$$M_{nxp} = U_{nxr}\Sigma_{rxr}V^{T}_{rxp}$$

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}_{(nxp)} = \begin{bmatrix} u_{11} & \cdots & u_{n1} \\ \vdots & \ddots & \vdots \\ u_{1n} & \cdots & u_{nn} \end{bmatrix}_{(nxr)} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p \end{bmatrix}_{(rxr)} \begin{bmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pp} \end{bmatrix}_{(rxp)}$$

Where, M is (nxp) dimensional marker matrix with entries i.e., 0's, 1's and 2's (0 for AA, 1 for Aa and 2 for aa). U=$\{u_{ij}\}$ is (nxr) dimensional matrix of order with 'n' eigen vectors and V=$\{v_{ij}\}$ is (rxp) dimensional matrix with 'p' eigen vectors, *U* and *V* are orthogonal matrices with orthonormal eigenvectors i.e., $UU^{T}$or $U^{T}U = I = VV^{T}$or $V^{T}V$ chosen from $MM^{T}$ and $M^{T}M$ respectively(i =1….n) and (j = 1….p). $\sum$ is a (rxr) diagonal matrix where $r \leq min\{n, p\}$ is the rank of M matrix and $\sum$with '*r'* elements equal to the square root of the positive eigenvalues $\sigma_i = \sqrt{\lambda_i}$ of $M^{T}M$ or $MM^{T}$ (both matrices have the same positive eigenvalues anyway). The diagonal elements are composed of singular values. Therefore, initially the missing values of $j^{th}$ marker column are imputed by the average of available values of $j^{th}$ marker column, to make it complete. This procedure is performed for all the marker columns having missing values so as to make M matrix complete. SVD was then utilized to impute the missing value in M, and this procedure is repeated on the newly obtained matrix, until the total change in the matrix falls below the empirically determined threshold of 0.01. The *impute. svd* function available in BCV package of R 3.6.0 software was utilized for this imputation.

### k-Nearest Neighbour Imputation (kNNI)

This imputation is based on k-nearest neighbour algorithm (k-NN). It is a non-parametric method mostly used in pattern recognition and also used for handling two problems such as classification and regression. kNN assigns weights to the contributions of the neighbours, so that the nearer neighbors contribute more to the average than the more distant ones (Hastie, 2001) [19]. kNN algorithm involves the following steps:
1) 'k' gets decided based on the square root of number of genotypes(k=$\sqrt{n}$). It is a positive integer and must be an odd number.
2) Calculate distance between the query genotype (with missing marker) and the genotype under consideration by excluding the corresponding missing markers with prominent distance measure is Euclidean distance: $d(X,Y) = \sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2}$

Where, X and Y are two different genotypes and d (X, Y) is distance between two different genotypes measured over 'p' number of markers without missing observations.
3) From $^{n}C_2$ distances, 'k' number of nearest genotypes are considered by giving weights as inverse of distance (1/d), where larger weights are given for nearest neighbor genotypes. The missing markers in query genotype are then replaced with the weighted mean of the k most similar genotypes based on Euclidean distance between standardized observations. kNN imputation was implemented by using *impute.knn* function of IMPUTE package in R 3.6.0.

### Locally weighted linear regression method Imputation (LWI)

This is a non-parametric regression algorithm proposed by Pyeye *et al.*, 2016 [13], where the model does not learn a fixed set of parameters as is done in ordinary linear regression. Relatively, the parameters (βs) are measured individually for each missing value (z). While computing 'β', a higher preference is given to the points in the training set lying in the vicinity of missing value (z) than the points far away from missing value (z). LWI was executed under *locfit* package in R 3.6.0.

### Expectation maximization imputation (EMI)

This imputation technique is based on Expectation-Maximization algorithm and used for predicting the values of latent variables with the condition that the general form of

probability distribution governing those latent variables is known to us (Poland J, 2012) [10, 11]. It is used to find the local maximum likelihood parameters of a statistical model in the cases where latent variables are involved and the data is missing or incomplete. Under this method, non-missing marker data is used to obtain the maximum likelihood estimates of the vector of means ($\bar{x}$) and covariance matrix(ŝ)of the individuals based on the markers, called 'E-step' for the 'Expected' distribution. These estimates fed into the model to obtain multiple linear regression estimates of the missing marker values called 'M-step' for the maximization of the distribution. $\bar{x}$ and ŝ were then re-estimated and over again used to re-estimate the missing marker values. This process was repeated until the difference between the new estimate and the previous estimate of $\bar{x} + \hat{s} * \hat{s}^T$was less than or equal to 0.02. *EMI* function was used from *BWGS* package of R 3.6.0 (Charmet G, 2020) [3].

**Random forest imputation (RFI)**

Random forest imputation is based on random forest algorithm, which is a supervised machine learning algorithm mainly used for the purpose of solving classification and regression problems. The final decision of classifying the instances is made on the basis of the majority of the decision trees (Brieman *et al*., 1984).

The steps involved in this procedure: (1) construct bootstrapped data sets from training data set with missingness, (2) construct decision trees using bootstrapped datasets, wherein which marker should go to root node was decided randomly and leaf nodes can be selected with the help of Gini Index (GI) or Gini split $GI = 1 - \sum_{i=1}^{C}(p_i)^2$ (i = 0, 1 and 2 for number of SNPs) (Gini, 1912), (3) repeat the step 1 and 2 to get defined number of decision trees then missing observation was decided based on majority of decision trees by aggregating them called 'Bagging'. The *missForest* function was utilized from *miss Forest* package of R 3.6.0.

**Parametric values of different imputation techniques**

The missForest of R package was used for imputing the missing values by RFI. The *nodesize*, *maxiter*, *ntree* and *mtry* were kept as 5, 10, 200 and 57 (=sqrt (3285)) respectively. In case of EMI, *tol*=0.01 was set, i.e., threshold between true and imputed value. The *maxiter* and *tol* as 10 and 0.01 respectively were set for SVDI and *k*, *rowmax* and *colmax* as 10, 0.5 and 0.8 for kNNI respectively. LWI and MNI were used directly with the default values of parameters

**Prediction accuracy of imputation techniques**

The prediction accuracy of a given imputation technique at a given level of missingness was computed as average correlation coefficient ($\bar{r}$), where average is taken over the estimated correlation coefficients ($\hat{r}_i$'s) obtained from 100 simulated data sets (*i*=1,2,….100) and the correlation coefficient from each data set is estimated between imputed missing values and the original flagged values. The $\hat{r}_i$ for an $i^{th}$ simulated data set is given by $\hat{r}_i = \frac{cov(X,Y)}{\sqrt{v(X).v(Y)}}$, where *X* and *Y* are vectors of flagged values and imputed values. The standard error of correlation coefficient (SE(*r*)) is estimated as square root of variance among $\hat{r}_i$'s. Student's t-test is used for testing the $H_0$: ρ=0 at α (=0.01) level of significance. The *t* statistics is calculated as ratio of average correlation coefficient to SE (*r*). On the other hand, Mean Squared Prediction Error (MSPE) from $i^{th}$ simulated data set was estimated by $MSPE_i = \frac{\sum_{j=1}^{n}(X_{ij}^{flagged} - Y_{ij}^{imputed})^2}{n}$, (Schmitt *et al*. 2015) [15]. Where *n* is the number of missing values. The average MSPE is then estimated over 100 simulated data sets.

**Results and Discussion**

The mice data set described in materials and methods was used for creating missing values randomly at various levels of missingness, viz, 5%, 10%, 15% and 20%. Under each level of missingness, 100 data sets were simulated by randomly deleting the observations. However, the original values of the deleted observation were recorded and marked as "flagged values". Six imputation techniques viz, EMI, MNI, LWI, SVDI, k-NNI and RFI were applied on each of the 100 simulated data sets for a given level of missingness. Subsequently, all the imputed values were recorded and Pearson correlation coefficient between the imputed missing values and the original flagged values was computed for each of the simulated data set for a given level of missingness and that too under a given imputation technique. The mean and standard error of correlation coefficient over 100 simulated data sets were estimated and presented in Table 1. This procedure was repeated for 10%, 15% and 20% of missing values and the corresponding estimated mean and standard error of correlation coefficient were presented in Table 1. Similarly, for all other imputation techniques, the estimated mean and standard error of correlation coefficients were presented in Table 1.

The mean square prediction errors from each of the simulated data sets under six imputation techniques: EMI, MNI, LWI, SVDI, k-NNI and RFI at different levels of missingness were computed and presented in Table 2. In addition, the estimated correlation coefficients along with standard errors under each level missingness over different imputation techniques were graphically represented in Fig. 1.

Table. 1 reveal that at 5% level of missing observations, RFI has shown highest correlation between flagged and imputed values, followed by EMI and SVDI, KNNI whereas LWI has shown lowest correlation. Whereas, the standard error of correlation coefficient from RFI at 5% level of missingness was found to be lowest and from LWI it was observed as highest (Fig. 1). Such trends in terms of correlation coefficients and standard errors were observed under 10%, 15% and 20% missingness. Besides, RFI among different imputation techniques has shown lowest MSPE value under 5% missingness (Table 1). However, the values of MSPE for EMI, SVDI, kNNI, MNI and LWI were increased in the order the techniques arranged. Such trend in the performance of RFI over other imputation techniques in terms of MSPE was observed under 10%, 15% and 20% missing data situations.

Troyanskaya *et al.* (2001) showed that kNNI outperforms SVDI while using DNA micro array data with 1 to 20% missingness. However, in the present study, SVDI shown a slightly higher performance than kNNI with GBS data, where marker response is 0, 1, and 2 or -1, 0 and +1. Stekhoven and Buhlmann, (2012) [18] presented that RFI outperforms on mixed data such as continuous as well as categorical data compared to k-NN imputation and multivariate imputation using chained equations (MICE). Further, RFI has no need for tuning parameters nor does it require assumptions about distributional aspects of the data. Our results also revealed that RFI has shown high imputation accuracy than kNNI in GBS data situation with various levels of missing observations. Charmet *et al*. (2020) [3] reported that the imputation accuracy of EMI technique was greater than MNI

technique when presence of missingness in GBS data upto 80% while using GS models for prediction of breeding values. The results from the present study also showed a similar trend in imputation accuracy between EMI and MNI even without using GS models.

**Table 1:** Estimates of Correlation coefficient between flagged values and imputed missing values for various levels of missing (LoM) observations under different imputation techniques

| LoM | Imputation Technique | | | | | |
|---|---|---|---|---|---|---|
| | RFI | EMI | SVDI | KNNI | MNI | LWI |
| 5% | 0.7788±0.0037 | 0.6785±0.0043 | 0.6583±0.0044 | 0.6455±0.0045 | 0.4791±0.0051 | 0.2541±0.0057 |
| 10% | 0.7659±0.0027 | 0.6565±0.0031 | 0.6385±0.0032 | 0.6254±0.0032 | 0.4594±0.0037 | 0.2456±0.004 |
| 15% | 0.7509±0.0022 | 0.6387±0.0026 | 0.6145±0.0027 | 0.6158±0.0027 | 0.4455±0.003 | 0.2354±0.0033 |
| 20% | 0.7353±0.002 | 0.6254±0.0023 | 0.5915±0.0024 | 0.5855±0.0024 | 0.4124±0.0027 | 0.2195±0.0029 |

**Table 2:** Mean Square prediction error (MSEP) between flagged and imputed values at various levels of missing observations under different imputation techniques

| LOM | Imputation Technique | | | | | |
|---|---|---|---|---|---|---|
| | RFI | EMI | SVDI | KNNI | MNI | LWI |
| 5% | 0.010 | 0.0112 | 0.0137 | 0.0141 | 0.0199 | 0.028 |
| 10% | 0.0107 | 0.0119 | 0.0146 | 0.0148 | 0.0209 | 0.0301 |
| 15% | 0.0129 | 0.0126 | 0.0151 | 0.0154 | 0.0214 | 0.0334 |
| 20% | 0.0136 | 0.0144 | 0.0158 | 0.0168 | 0.0222 | 0.0344 |

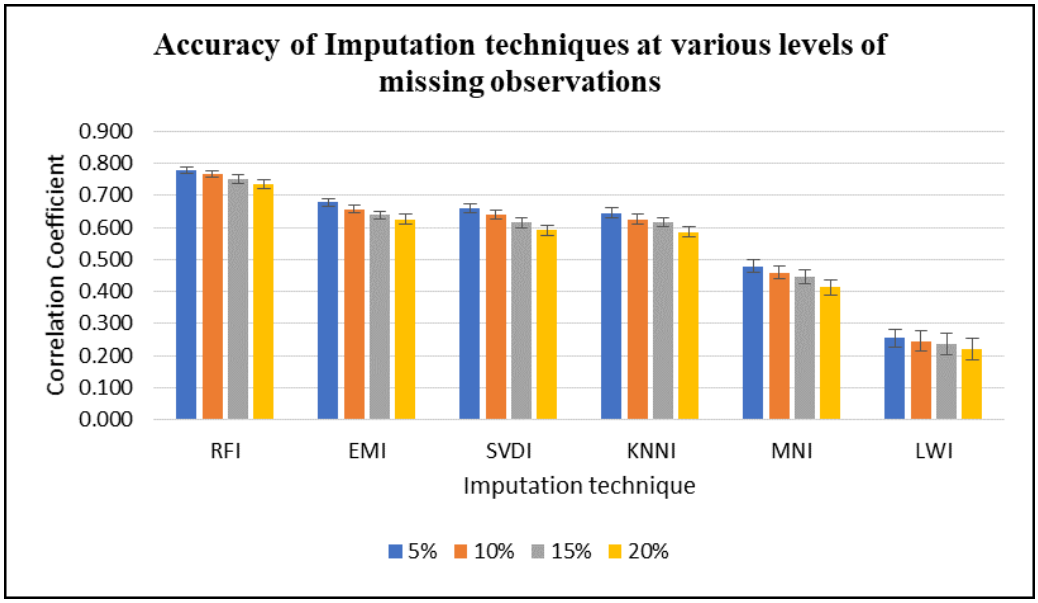


**Fig 1:** Imputation accuracy of different imputation techniques measured as estimated correlation coefficients between flagged and imputed values at varying levels of missingness

## Conclusion

The present study revealed that RFI outperformed EMI, SVDI, kNNI, MNI and LWI imputation techniques while using GBS data at different levels of missing observations, i.e., 5%, 10%, 15% and 20%. The performance of imputation techniques is assessed in terms of imputation accuracy measured as high significant correlation between observed (flagged) and imputed values as well as in terms of Mean Squared Prediction Error.

## Acknowledgement

SB is thankful to PG School, IARI and ICAR-IASRI for providing computational facilities and support for conducting the study. SB also acknowledges the RJNF-SC fellowship received from University Grants Commission, New Delhi.

## References

1. Alter O, Brown OP, Botstein D. Singular value decomposition for genome-wide expression data processing and modelling. PNAS 2000;97(18):10101-10106.
2. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software 1984.
3. Charmet G, Tran LG, Auzanneau J, Renaud Rincent R, Bouchet S. BWGS: A R package for genomic selection and its application to a wheat breeding programme. PLoS One 2020;15(4):e0222733.
4. Elshire RJ, Glaubitz JC, Qi S, Poland JA, Kawamoto Ken *et al.* A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 2011;6(5):e19379.
5. Gini C. The origins of the Gini index: extracts from Variability and Mutability (1912). The Journal of Economic Inequality 2012;10:421-443.
6. Johannsen WL. The elements of an exact theory of heredity. Arizona State University. School of Life Sciences. Center for Biology and Society. Embryo Project Encyclopedia 1909.

7. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Review Genetics 2010;11(7):499-511.
8. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 2007;39(7):906-13.
9. Pearson K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 1895;58:240-242.
10. Poland JA, Rife TW. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome 2012a;5:92-102.
11. Poland J, Endelman JB, Dawson J, Rutkoski JE, Wu S, Manes Y, *et al.* Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome 2012b;5:103-113.
12. Ponnapalli SP, Saunders MA, Van Loan CF, Alter O. A Higher-Order Generalized Singular Value Decomposition for Comparison of Global mRNA Expression from Multiple Organisms. PLoS ONE 2011;6(12):e28072.
13. Pyeye S, Syengo CK, Odongo L, Orwa GO, Odhiambo RO. Imputation Based on Local Linear Regression for Nonmonotone Non-respondents in Longitudinal Surveys. Open Journal of Statistics 2016;6:1138-1154.
14. Sanger F, Thompson EOP. Kitai R. The amide groups of insulin. Biochemical Journal 1955;59(3):509-518.
15. Schmitt P, Mandel J, Guedj M. A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics 2015;6:1.
16. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 2014;15:121-132.
17. Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. Journal of Molecular Biology 1970;51:379-91.
18. Stekhoven DJ, Peter Buhlmann P. Miss Forest-non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112-118.
19. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520-525.
20. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P *et al.* Genome wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics 2006a;38:879-887.
21. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlis JNP *et al.* Genetic and environmental effects on complex traits in mice. Genetics 2006b;174(2):959-984.
22. Watson JD, Crick FHC. Genetical Implications of the Structure of Deoxyribonucleic Acid. Nature 1953;171:964-967.
23. Yoon KP, Hwang C. Multiple Attribute Decision Making: An Introduction. California: SAGE publications 1995.