Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data

R Florez-Lopez

To cite this article: R Florez-Lopez (2010) Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data, Journal of the Operational Research Society, 61:3, 486-501, DOI: 10.1057/jors.2009.66

To link to this article: https://doi.org/10.1057/jors.2009.66

Journal of the Operational Research Society (2010) 61, 486–501

www.palgrave-journals.com/jors/

Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data

R Florez-Lopez*

University of Leon, Leon, Spain

The 2004 Basel II Accord has pointed out the benefits of credit risk management through internal models using internal data to estimate risk components: probability of default (PD), loss given default, exposure at default and maturity. Internal data are the primary data source for PD estimates; banks are permitted to use statistical default prediction models to estimate the borrowers' PD, subject to some requirements concerning accuracy, completeness and appropriateness of data. However, in practice, internal records are usually incomplete or do not contain adequate history to estimate the PD. Current missing data are critical with regard to low default portfolios, characterised by inadequate default records, making it difficult to design statistically significant prediction models. Several methods might be used to deal with missing data such as list-wise deletion, application-specific list-wise deletion, substitution techniques or imputation models (simple and multiple variants). List-wise deletion is an easy-to-use method widely applied by social scientists, but it loses substantial data and reduces the diversity of information resulting in a bias in the model's parameters, results and inferences. The choice of the best method to solve the missing data problem largely depends on the nature of missing values (MCAR, MAR and MNAR processes) but there is a lack of empirical analysis about their effect on credit risk that limits the validity of resulting models. In this paper, we analyse the nature and effects of missing data in credit risk modelling (MCAR, MAR and NMAR processes) and take into account current scarce data set on consumer borrowers, which include different percents and distributions of missing data. The findings are used to analyse the performance of several methods for dealing with missing data such as likewise deletion, simple imputation methods, MLE models and advanced multiple imputation (MI) alternatives based on MarkovChain-MonteCarlo and re-sampling methods. Results are evaluated and discussed between models in terms of robustness, accuracy and complexity. In particular, MI models are found to provide very valuable solutions with regard to credit risk missing data.

Journal of the Operational Research Society (2010) **61,** 486–501. doi:10.1057/jors.2009.66
Published online 2 September 2009

Keywords: banking; credit risk; management; forecasting; missing data; scarce data

Introduction

The 2004 Basel II Accord has pointed out the benefits of credit risk management through internal ratings-based models (IRB approach) that use each bank's internal data to categorize borrowers into risk grades and to estimate risk components: probability of default (PD), loss given default, exposure at default and effective maturity (BCBS, 2004). Internal data are considered to be the primary source of information for the estimation of PD. Banks are permitted to use statistical default prediction models to estimate borrowers' PD, subject to the fulfilment of some requirements related to accuracy, completeness and appropriateness of data, which must be representative of the total number of existing borrowers. However, in practice, internal records are usually incomplete or do not contain adequate history for the estimation of PD (Carey and Hrycay, 2001; ECB, 2004). The problem of current missing values is more critical with regard to low default portfolios (LDPs), characterized by inadequate default records, which make it difficult to design a statistically significant prediction model. Nevertheless, this problem has been largely ignored in credit risk analysis (OeNB, 2004; BCBS, 2005a).

Several methods might be used to deal with missing values such as listwise deletion (CC), application-specific CC, substitution techniques, maximum likelihood (ML) or multiple imputation (MI) models. CC is an easy-to-use method widely applied by social scientists but with existing sparse data sets generates a substantial loss of cases and reduces the diversity of information resulting in a bias in the model's parameters, results and inferences (King *et al*, 2001).

**Correspondence: R Florez-Lopez, Faculty of Economics and Business Administration, University of Leon, Leon 24071, Spain.*
E-mail: raquel.florez@unileon.es

Table 1 Risk components for IRB foundation and advanced approaches

Exposure	*IRB - FOUNDATION*		*IRB - ADVANCED*	
	Internal estimates	*Supervisory estimates*	*Internal estimates*	*Supervisory estimates*
Corporate, sovereign and bank exposures	PD	LGD, EAD, M	PD, LGD, EAD,M	—
Retail exposures	PD, LGD, EAD	M	PD, LGD, EAD	M
Equity exposures*	Potential loss (*market-based approach*) PD, LGD (*PD/LGD approach*)			

*For equity exposures two specific alternatives are proposed, named market-based approach and PD/LGD approach (BCBS, 2004: paragraphs 340–361).

The selection of the best method for solving the missing data problem largely depends on the nature of the missing values. Nevertheless, there is a lack of empirical analysis about their effect on credit risk, which limits the validity of consequent models. In this paper, we analyse the nature (MCAR, MAR and MNAR processes) and effects of missing data in credit risk modelling. We consider a well-known sparse data set on retail borrowers (Quinlan, 1979), which include different percents and impact of missing data for categorical and continuous features. Findings on the nature of missing values are used to analyse the performance of several methods for dealing with missing data such as likewise deletion, mean/median substitution (MS), ML methods and advanced MI alternatives based on MarkovChain-MonteCarlo (MCMC) and re-sampling methods. Results are evaluated and discussed between models in terms of robustness, accuracy and complexity. In particular, some MI models are found to provide very valuable solutions with regard to credit risk missing data. The paper is structured as follows: The effect of missing data on credit risk scoring is analysed in the next section with particular consideration to LDPs. In the subsequent section, main approaches for dealing with missing values are presented, a discussion of their weaknesses and strengths and an in-depth analysis is carried out on the nature of missing data. The penultimate section includes the comparative empirical analysis of previous methods with main conclusions summarised in the final section.

Effects of missing data on credit risk scoring

PD is the most significant credit risk component to be estimated for banks in the development of IRB systems both in fundamental and advanced approaches (Table 1) (BCBS, 2004: paragraph 391). Basel II establishes some overall requirements about rating system design, estimation and validation of PD measures that must be fulfilled to use IRB models. Also, some specific rules are established for broad classes of assets: (a) corporate, (b) sovereign, (c) bank, (d) retail, and (e) equity. With existing corporate, sovereign and bank exposures, the PD estimate could be based on internal default experience, mapping to external data and/or statistical default models (BCBS, 2004, paragraph 417). With regard to retail exposures, internal data must be considered as the foundation data set for PD estimates. In particular, credit-scoring models, based on statistical techniques and historical databases, are usually considered as the primary or partial basis of rating assignments and PD estimates. To use a credit-scoring model, the bank needs to prove the procedure is stable and has a good predictive capacity, the exogenous variables are a reasonable set of predictors, there are no known material biases and a regular validation process is performed.

In addition, the accuracy, completeness and appropriateness of the data set used to build the model needs to be demonstrated, so an extensive database with sufficient cases must be used to obtain statistically valid results. However, in practice, records are usually incomplete or do not have adequate history for estimating the PD (Carey and Hrycay, 2001), particularly if portfolios with a limited numbers of default cases are considered. Existing insufficient default history can cause average observed default rates to be statistically unreliable estimators of PD and could result in those considered LDPs being excluded from IRB treatment (Benjamin *et al*, 2006). Also, a lack of quality in the IRB components derived from insufficient or lower quality data relative to what is used to estimate internal parameters could give rise to supervisory concern (BCBS, 2006). However, this exclusion could greatly affect the banks' requirements for credit risk due to many entities indicating that at least 50% of wholesale assets and a material proportion of their retail portfolios could be considered as LDP (BBA, 2004). To avoid this undesired result, Basel II has approved a directive suggesting the most appropriate management for LDPs (BCBS, 2005b).

Even if there is no clear definition of LDP, several categories of portfolios may have low default numbers (BBA, 2004; FSA, 2005; BCBS, 2005b): (1) *type I portfolios* that historically have experienced a low number of defaults and are considered to be low-risk (eg banks, sovereigns, insurance companies, highly rated companies); (2) *type II portfolios* that have a low number of counterparties (eg train operating companies), or represent small markets of the counterparty and exposure type (eg niche markets); (3) *type III portfolios* that have a lack of historical data (eg caused by being a new entrant into a market or operating in an emerging market); (4) *type IV portfolios* that may not have incurred recent losses but past experience might suggest there is a greater likelihood of losses than captured in recent records (eg retail mortgages in a number of jurisdictions). Table 2 categorizes previous portfolios in consideration of their duration (long term *versus* short term) and their nature (systemic *versus* institution specific).

Table 2 LDP categories

Nature/Duration	*Long term*	*Short term*
Systemic	LDP type I	LDP type IV
Institution Specific	LDP type II	LDP type III

The BCBS directive focuses on type I portfolios, but suggestions are also applicable to small portfolios or sparse data portfolios (LDP type III). It recommends alternative data sources and data-enhancing tools for risk quantification and IRB systems' validation. Among data-enhancing tools, pooling data with other banks, models for improving data input quality and the use of more complex validation tools are found to be very useful. Rating assignments should be based on historical experience and empirical evidence and not purely on human judgement (BCBS, 2005b, paragraph 49). Also, different rating categories could be combined, and PDs estimated for the combined category could enhance default data without necessarily disrupting the predictability of the rating systems. Also, the PD estimation should be based on very prudent principles such that the upper limit of the PD confidence interval could be considered a pessimistic approach to the real PD (Pluto and Tasche, 2004; FSA, 2005; Schuermann and Hanson, 2005; BCBS, 2005b).

In particular, the existence of missing values in LDP historical data sets dramatically reduces the number and quality of observation data and also affects the statistical estimator and confidence intervals. Missing data affect the frequency of an exogenous variable (indicator), which is calculated in relation to the overall sample of cases (OeNB, 2004). In multivariate analyses, a value must be disposable for each case to be processed otherwise a rating cannot be obtained for the case. Thus, it is necessary to handle missing values appropriately and four different approaches could be considered (OeNB, 2004, pp 77):

- Excluding cases in which an indicator cannot be calculated; often impracticable because it excludes too many data records, samples become smaller and could be rendered invalid.
- Excluding indicators that do not attain a minimum level of availability. In practice, if an indicator can be calculated in less than approximately 80% of cases, it is not possible for missing values to be statistically handled and the indicator should be excluded from the analysis.
- Including missing indicators as a separate category in the analysis; very difficult to apply in the development of scoring functions from quantitative data because missing values do not constitute independent information and cannot be directly substituted for any significant quantitative value.
- Replacing missing values with estimated values specific to each group. The selection of the best method for solving the missing data problem largely depends on the nature of the missing values.

The last approach is the most suitable and statistically valid procedure for handling missing values for quantitative analysis. As a result, a final database is obtained in which a valid value can be found for each indicator in each case; it could be used for performing multivariate analysis in the development of credit scoring models (OeNB, 2004, pp 78).

Dealing with missing values. A methodological analysis

In social science, clinical trials and other observational studies, complete data are often not available for every case (Schafer, 1997; Little and Rubin, 2002). Missing data could be caused by many circumstances, some due to research design and some to chance: items non-response,[1] missing by design, partial non-response, previous data aggregation, loss of data, etc. (Ibrahim *et al*, 2005; Horton and Kleinman, 2007). Models that deal with missing data are of particular interest to many existing problems. The absence of just a few percent of a number of predictors (or covariates) could lead to a large number of cases with missing information (Horton and Kleinman, 2007); with the existence of sparse data sets, missing values lead to a higher loss of information because of the limited amount of data available in the sample. The development of methods for handling missing values has been an active area of research; see Schafer (1997), Little and Rubin (2002), Ibrahim *et al* (2005) for a review. Many of these methods have been focused on generalized regression models with missing covariate data both categorical and continuous (Little and Rubin, 2002; Ibrahim *et al*, 2005; Horton and Kleinman, 2007).

Models that incorporate partially observed predictors greatly depend on three assumptions about the process through which the data have come to be missing (Rubin, 1976). To analyse such assumptions, we will use the notation proposed by Little and Rubin (2002). Let D be the data matrix collected on a sample of n subjects and that principle interest relates to the parameters governing the conditional distribution $f(Y_i|X_i, \beta)$. For a given subject, X could be partitioned into components denoting observed variables (X_{obs}) and missing values (X_{mis}). Denote R a set of response indicators (ie, $R_j = 1$ if the j-th element of X is observed, 0 otherwise) could be analysed in terms of probability models for R (Table 3). The missing completely at random (MCAR) considers that:

$$P(R|Y, X) = P(R|Y, X_{obs}, X_{mis}) = P(R|\phi),$$

where ϕ and β are presumed distinct. Briefly, this assumption considers that missing values are not related to any factor,

[1] A complete analysis on causes, prevention and treatment of item non-response can be consulted in De Leeuw *et al* (2003).

Table 3 Missingness assumptions

Assumption	*Acronym*	*R could be predicted by*
Missing completely at random	MCAR	—
Missing at random	MAR	X_{obs}
Not missing at random (non-ignorable missing data)	MNAR, NINR, NI	X_{obs} and X_{mis}

Source: King *et al* (2001, p 50).

known or unknown, of the study (Horton and Kleinman, 2007).

Otherwise, the **missing at random assumption (MAR)** considers that:

$$P(R|Y, X) = P(R|Y, X_{obs}, \phi).$$

Heuristically, this assumption states that missing values depend only on observed quantities, outcomes and predictors. Therefore, conditional on the observed data, the failure to observe a value does not depend on the unobserved data. MAR is a more realistic assumption than MCAR, and some authors have noted that the inclusion of a relatively abundant set of predictors in the model (Collins *et al*, 2001) or information regarding the outcome (Moons *et al*, 2006), the MAR assumption could be made more plausible.

Finally, if the probability $P(R|Y, X)$ cannot be simplified, the process is named as non-ignorable (NI), non-ignorable non-response (NINR) or **missing not at random (MNAR)**. Thus, the probability of missing data depends on the unobserved value of the missing response and a correct specification of the missingness law should be given in order to obtain consistent estimates of the regression parameters (Little and Rubin, 2002).

Formally, it is possible to test the MCAR assumption against the alternate hypothesis of MAR (Little, 1988; Hair *et al*, 1999); the analysis of correlations between all pairs of predictors that include missing values previously dichotomized,[2] allows to assume or not a MCAR process (no significant correlations are observed). If an MCAR process is not observed, the analysis of correlations between all pairs of predictors with and without missing values (real values) permits the assumption of an MAR process (a majority of significant correlations are observed) or MNAR behaviour (otherwise). By definition, the presence or absence of MNAR can never be demonstrated using only the observed data. Thus, without additional information such as theory, logic or prior data, it is impossible to test whether MAR or MNAR holds. But again, the analysis of dichotomous correlations between predictors with and without missing values could suggest the presence of an MNAR process (no significant correlations are observed) or an MAR assumption (a majority of significant correlations are observed). Another significant concept regarding missing values is related to the monotony of missing data. If a hierarchy of missing values could be observed within the data matrix, so that observing a particular variable X_b for a subject implies that X is observed, for $a < b$, then the missing value is said to be monotone; if the missing value is not monotone, models should include covariate, which are also missing values (Horton and Kleinman, 2007).

For practice purposes, methods to handle missing values could be classified into six different groups (King *et al*, 2001; Ibrahim *et al*, 2005; Horton and Kleinman, 2007): (1) CC; (2) substitution approaches and other 'ad hoc' methods; (3) ML; (4) MI; (5) weighting methods; and (6) fully Bayesian (FB) approaches.

Listwise deletion, also called *complete-case* analysis (CC), is the technique most commonly used in the presence of missing values. It involves analysis of the set of completely observed subjects so that entire observations are deleted when any one variable remains missing (King *et al*, 2001). Advantages of this approach are simplicity and comparability of univariate statistics since they are calculated on a common sample base of subjects (Little and Rubin, 2002). Limitations arise from the loss of information derived from deleting incomplete cases that could be quite high if a large number of predictors is considered. In addition, when the MCAR assumption is violated, CC could generate biased parameter estimates; it could result in errors of magnitude and signs of causal and descriptive inferences that largely depend on the original sample size,[3] being more important for sparse data sets (Anderson *et al*, 1983).

As a variant, *application-specific listwise deletion* (also called *available-case analysis*) proposes to include all cases where the variable of interest in a specific application is present. This alternative uses all available values, but the sample changes from variable to variable that creates practical problems when tables are computed for various conceptual sample bases (Little and Rubin, 2002). The bias problem is observed if data are not MCAR. *Substitution approaches* are simple imputation methods that make use of correlation among predictors to deal with missing values. Suppose a case with X_j missing contains the value of another variable X_k

[2] Variables are dichotomized though the substitution of observed values by 1 and missing values by 0.

[3] The degree on mean square error will often be more than one standard error and its direction will depend on the application, pattern of missing data and model estimated (Sherman, 2000).

that is highly correlated with X_j. It is possible to predict the missing value of X_j from X_k and then include the substituted (or imputed) value in analyses involving X_j (Little and Rubin, 2002). The Variants are as follows.

(i) Imputing unconditional mean: involves estimation of missing values using mean imputation (average of the observed values). In the presence of categorical variables, the unconditional mean is substituted by the unconditional median or mode. As a consequence, the average of observed and imputed values does not change but the variance of the observed and imputed values are underestimated.
(ii) Imputing conditional mean (Buck's method): is a form of imputation that substitute means that are conditioned on the variables recorded in the incomplete case. If $X_1, \ldots X_k$ are multivariable normally distributed with mean μ and covariance matrix Σ, then missing values have linear regressions of the observed variables with coefficients that are functions of μ and Σ. Buck's method first estimates μ and Σ based on complete cases and then uses such parameters to calculate the linear regression of the missing values.
(iii) Last value carried forward: for longitudinal studies, the last observed value is used for imputation.

As principal inconveniences, previous methods induce bias as well as understanding variability, their performance is unreliable, and they often require ad hoc adjustments to yield satisfactory estimates (Little and Rubin, 2002; Carpenter *et al*, 2004; Jansen *et al*, 2006). Other *ad-hoc methods* include recoding missing values to some common value or category, creation of an indicator of missing values as a new variable, including both these variables together with their interaction in the regression model. Also, it is possible to drop variables that have a large percentage of missing cases but could be derived by excluding important factors in the model. Nevertheless, ad-hoc approaches induce bias and are not recommended (Horton and Kleinman, 2007).

Maximum Likelihood (ML) methods have been largely used for making statistical inferences about missing values in many model-based procedures. ML methods implicitly assume that missing values are MAR with the main aim related to the regression parameters that govern the conditional distribution $f(Y|X, \beta)$. When some predictors have missing values, information could be inferred through estimating the distribution of covariates $f(X|\gamma)$, see Ibrahim (1990). Finally, the joint distribution is maximized $f(X, Y|\beta, \gamma)$ through the use of some specific methods such as the expectation-maximization (EM) algorithm (Dempster *et al*, 1977). EM is a fast deterministic algorithm for finding the maximum of the likelihood function through two steps (E—expectation and M—maximization); the E step finds the conditional expectation of the missing data giving the observed data and current estimated parameters (mean vector μ and covariance matrix Σ), and then substitutes these expectations for the 'missing data':

$$\tilde{X}_{mis} = ML(X_{mis}|X_{obs}, \hat{\mu}, \hat{\Sigma}).$$

The M step performs ML estimation of parameters as if there were no missing data:

$$\hat{\mu}, \hat{\Sigma} = ML(\mu, \Sigma|X_{obs}, \tilde{X}_{mis}).$$

The process iterates such that both the imputations and parameters are the single maximum posterior values. The advantages of EM are its fast and deterministic convergence; also, likelihood theory and the EM algorithm could be adapted for models with known NI missing-data mechanisms in which some observations are rounded and grouped into categories (Schafer, 1997; Little and Rubin, 2002). Main limitations refer to the possibility of settling on a local maximum; more significantly, the algorithm only obtains maximum values but does not estimate the entire density. Ignoring estimation uncertainty means that standard errors are generally biased downward and point estimates could be biased (King *et al*, 2001).

Multiple imputation (MI) is based on replacing each missing value by $m > 1$ simulated values, where m could be as small as 5–10 (Rubin, 1987). The m set of imputations reflect uncertainty about the true values of missing data such that m plausible versions of the complete data set exist. Each of these data sets is analysed using complete-case methods and results are combined to produce only one inferential solution: to estimate some quantity of interest q (such as a variable mean or regression coefficient), the final point estimate $\hat{q}$ could be obtained as the average of the m separate estimates q_j.

$$\hat{q} = \sum_{j=1}^{m} \frac{q_j}{m}.$$

Let $\hat{S}E^2(q_j)$ denote the variance estimate of q_j for data set[4] j and let $\hat{B} = \sum_{j=1}^{m}(q_j - q^*)^2/(m-1)$ denote the sample variance across the m point estimates, then the variance of the MI point estimates is the weighted average of the estimated variances from within each completed data set plus the sample variance in the point estimates across the data sets; for $m = \infty$, it would be a simple average of the two sources of uncertainty:

$$\hat{S}E^2(\hat{q}) = \frac{1}{m}\sum_{j=1}^{m} \hat{S}E^2(q_j) + \hat{B} \cdot \left(1 + \frac{1}{m}\right).$$

The task of generating MIs has been traditionally problematic and research has focused on generating MIs using MCMC techniques. One of the most known MI algorithms based on MCMC is the Imputation-Posterior (IP) method (Schafer, 1997).

[4] Obtained by assuming that the imputed data set is the complete data set and calculating the usual variance estimate.

IP works like EM except that conditional expectations of parameters are replaced by random draws from the multivariate normal observed data posterior $P(X_{mis}|X_{obs})$, and random draws are obtained for $\hat{\mu}, \hat{\Sigma}$. Two steps are also carried out iteratively; in the imputation step (I-step), imputations are drawn from the conditional predictive distribution of the missing data:

$$\tilde{X}^{mis} \sim P(X^{mis}|X^{obs}, \hat{\mu}, \hat{\Sigma}).$$

In the posterior step (P-step), new values of the parameters are drawn from their posterior distribution which also depends on the observed data and imputer values of subjects:

$$\hat{\mu}, \hat{\Sigma} = P(\mu, \Sigma|X^{obs}, \tilde{X}^{mis}).$$

The process of a fixed number of steps is iterated until a stationary value of parameters is reached, which come increasingly from their actual distributions independent of the starting values. IP's main advantage is that distribution is exact but convergence only occurs after an infinite number of iterations; in practice, after a long 'burn-in period' where iterations are discarded, a finite number of iterations is imposed.

Also, simulation is a natural complement to the current tools for handling missing values and, in particular, the EM algorithm (Schafer, 1997, p 5). Markov Chains and other simulation techniques could be applied to precisely the same problems as EM and its implementation is similar to that of EM. King *et al* (2001) and Honaker and King (2006) provide three variants of EM based on MI that avoid previous estimation uncertainty problems:

(i) EM with sampling (EMs): This algorithm begins with EM and then adds back in estimation uncertainty so it gets draws from the correct posterior distribution on X_{mis}. To create imputations, the algorithm first runs EM to find the maximum posterior estimates of the parameters[5] $\hat{\theta} = vec(\tilde{\mu}, \tilde{\Sigma})$, and the variance matrix $V(\hat{\theta})$ is computed. Then, a simulation of θ is drawn from a normal with mean $\hat{\theta}$ and variance $V(\hat{\theta})$. Previous $\tilde{\theta}$ is used to compute $\tilde{\beta}$ deterministically; $\tilde{\varepsilon}$ is simulated from the normal such that:

$$\tilde{X}_{mis} = X_{obs} \cdot \tilde{\beta} + \tilde{\varepsilon}.$$

The entire process is repeated m times for MIs. EMs is very fast, works well in large samples, converges non-stochastically and obtains independent imputations. Nevertheless, for sparse data sets, data with many variables relative to the number of individuals, or highly skewed categorical data, the standard errors of MIs and predicted parameters could be biased.

(ii) EM with importance resampling (EMis): EMis is an improvement of EM through the inclusion of importance re-sampling, an iterative simulation technique that enhances small sample performance. EMis performs the same steps as EMs but draws of θ are treated as first approximation to the true posterior. Parameters are put on unbounded scales using the *log* for standard deviations and Fisher's z for correlations in order to obtain better results for the normal approximation in presence of small data sets. An acceptance-rejection algorithm is used by keeping simulations of θ with probability proportional to the 'importance-ratio':

$$IR = \frac{L(\theta|X_{obs})}{N(\tilde{\theta}|\tilde{\theta}, V(\tilde{\theta}))},$$

being $L(\cdot)$ a likelihood function. Finally, $\tilde{\theta}$ is used to compute $\tilde{\beta}$ and produce m imputations as it was in EMs. EMis gives draws from the same posterior distribution as IP but is much faster and seems to avoid convergence problems (King *et al*, 2001; Horton and Kleinman, 2007).

(iii) EM with bootstrapping (EMB): The EMB algorithm approaches the problem of sampling μ, Σ by missing theories of inference. In particular, the complicated process of drawing μ, Σ from their posterior density is replaced with a bootstrapping algorithm. As a result, EMB is able to impute data sets with many more variables and cases than the previous variants.

Other interesting MI methods for categorical and continuous variables are: (1) the Conditional Gaussian, popularized by Schafer (1997); (2) Chained Equations, which involve a variable-by-variable approach in which the imputation model is specified separately for each covariate, involving the rest of variables as predictors (van Buuren *et al*, 1999); (3) Predictive Matching Method for monotone data sets, which imputes a value randomly from a set of observed values whose values are closest to the predicted value from a regression model, etc. In MIs, a key issue is the appropriate specification of the imputation model. Usually, a multivariate normal specification is used, which implies that missing values are imputed linearly, but if some variables are not Gaussian it could lead to bias, mainly in presence of multiple categorical and continuous variables (Horton *et al*, 2003).

Also, a strategy for adjusting the bias of complete case selection is to assign them case weights for use in subsequent analyses, so-called *weighting methods* (Ibrahim *et al*, 2005; Carpenter *et al*, 2006). These approaches fit a model for the probability of missing values and the inverse of these probabilities are used as weights for the complete cases. An interesting weighting method is the Expectation Robust (ER) proposal, which modifies the M-step of the EM algorithm to include case weights d_i for $\hat{\mu}, \hat{\Sigma}$ estimates (Rocke, 1996); such weights are based on Mahalanobis distance in order to enhance cases near to the predictor's average and reduce the importance of outliers. Also, the ERTBS (Expectation Robust Translated

[5] Where $vec(\cdot)$ operator stacks the unique elements.

Table 4 Descriptive statistics for the Australian credit approval dataset

Var.	*Nature*	*Range*	*# Missing (%)*	*Mean (s.d.)* *Median (mode)*
Al	Categorical	{0, 1} (formerly: a,b)	12 (1.739%)	1 (1)
A2	Continuous	(13.75, 80.25)	12 (1.739%)	31.568 (11.958)
A3	Continuous	(0,28)	0	4.759 (4.978)
A4	Categorical	{1, 2, 3} (formerly: p,g,gg)	6 (0.869%)	2 (2)
A5	Categorical	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14} (formerly: ff, d, i, k, j, aa, m, c, w, e, q, r, cc, x)	9 (1.304%)	8 (8)
A6	Categorical	{1, 2, 3, 4, 5, 6, 7, 8, 9} (formerly: ff, dd, j, bb, v, n, o, h, z)	9 (1.304%)	5 (5)
A7	Continuous	(0,28.5)	0	2.223 (3.347)
A8	Categorical	{1, 0} (formerly: t, f)	0	1 (1)
A9	Categorical	{1, 0} (formerly: t, f)	0	0 (0)
A10	Continuous	(0,67)	0	2.400 (4.863)
All	Categorical	{1, 0} (formerly t, f)	0	0 (0)
A12	Categorical	{1, 2, 3} (formerly: s,g,p)	0	2 (2)
A13	Continuous	(0,2,000)	0	184.015 (173.807)
A14	Continuous	(0,100,000)	13 (1.884%)	1,017.386 (5,210.103)
A15	Categorical	1,2 class attribute (formerly: +, −)	—	—

Biweight S-estimator) algorithm departs from ER approach but considers both case weights and TBS estimator to assess multivariate means and covariates (Cheng and Victoria-Feser, 2000). As a main limitation, weighting methods tend to be untreatable in presence of multiple non-monotone missing variables.

Finally, *fully Bayesian* approaches (FB) could be applied more generally so that they are in MIs models. FB requires specific priors on all parameters as well as specific distributions for missing covariates but are easily implemented and do not require new techniques for inference. In fact, its implementation through uniform improper priors on all parameters leads to ML estimates, also, empirical results suggest that FB methods perform similarly to ML and MI methods (Ibrahim *et al*, 2005).

The performance of different methods depends on the existence of an MCAR, MAR, or MNAR process. In general, CC approaches obtain inefficient results regardless of the missing values assumption and are unbiased for MCAR. MI methods are more efficient than previously and they are unbiased under MCAR and MAR (Little and Schenker, 1995). All of them generate biased results in the presence of missing values following an MNAR process and cannot be dealt with without considering additional information. Finally, simple methods such as mean substitution and other '*ad hoc*' approaches could be used if the pattern is monotone but non-monotone subjects usually require MI models (Horton and Kleinman, 2007).

Empirical research: the Australian credit approval data set

In this paper we analyse the effect of missing data on a well-known authentic world data set on retail customers. The Australian credit approval data set is a two-class classification problem available from the UC Irvine Database Repository.[6] The file concerns credit card applications from 690 individuals involving 14 exogenous variables (six continuous and eight categorical features) and one class attribute (accepter or rejected). The sample includes 307 (44.5%) instances of creditworthy applicants and 383 (55.5%) instances of applicants who were not creditworthy. For mass market banking, retail credit risk databases typically include several tens of thousands (Jacobson and Roszbach, 2003; OeNB, 2004) or even hundreds of thousands of records (Staten and Cate, 2003), which depend on the bank's size and the country of operations. As a consequence, the previous data set can be considered 'not an extremely scarce-data retail portfolio' that represents a niche market, a new market or an emerging market for the Australian bank. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This data set is considered to be a good example of mixture attributes (continuous, nominal with small numbers of values and nominal with large numbers of values). Also, it includes 37 individuals with one or more missing value concerning six different features (Table 4) but they are replaced by defect by the mode of the attribute (categorical) or the mean of the attribute (continuous) (Quinlan, 1979, 1992; Baesens *et al*, 2000; Eggermont *et al*, 2004; Huang *et al*, 2006). The cause of missing values is the loss of information inside the data set.

Missing data affects A1, A2, A4, A5, A6 and A13 variables (40.00% of exogenous attributes). Nevertheless, only 37 individuals include missing values (5.36) and the highest number of missing data per variable only reaches 13 observations (1.88%). Some feature selection algorithms have been

[6] Information on this data set is available at http://mlearn.ics.uci.edu/databases/credit-screening/. Original data used in this paper can be obtained at http://mlearn.ics.uci.edu/databases/credit-screening/crx.data (information on missing values is included in the 'crx.data' file, see '?' symbols).

Table 5 Analyses of MCAR and MAR assumptions

Predictor	*First stage (MCAR assumption)*		*Second stage (MAR assumption)*	
	# pairs with significant correlation ($p = 0.05$)	*# pairs without significant correlation* ($p = 0.05$)	*# pairs with significant correlation* ($p = 0.05$)	*# pairs without significant correlation* ($p = 0.05$)
A1	2	3	—	—
A2	0	5	3	1
A4	3	2	—	—
A5	4	1	—	—
A6	4	1	—	—
A13	3	2	4	0
# pairs analysed	15		8	

previously applied to the Australian credit approval data set (Cavaretta and Chellapilla, 1999; Huang *et al*, 2006). Nevertheless, King *et al* (2001) establish, as a key point for the practice, that the imputation model should contain at least as much information as the analysis model. Also, for greater efficiency, any other variables could be added to the data set that would help predict the missing values. As far as computational purposes is concerned, the sample size must include at least $p(p+3)/2$ observations, where p is the number of exogenous variables ($p = 14$); in this data set, a minimum sample of 119 individuals will be required, which is more than fulfilled by the analysed data set.

From previous predictors, a binary logistic regression was considered to model the final class attribute. Logistic regression has been previously used for analysing the performance of several methods to deal with missing data (Ibrahim *et al*, 2005; Horton and Kleinman, 2007) and also provides quite flexible models for credit risk modelling. Six different methods for dealing with missing data are considered: (1) CC; (2) unconditional MS; (3) EM algorithm; (4) IP algorithm (10 000 iterations are considered for burning in purposes); (5) EM is algorithm (a very conservative 1/50 ratio of draws to iterations was included, see King *et al*, 2001); and (6) EMB algorithm. For MI methods (IP, EMs, EMis, EMB), 10 different imputations have been performed (including 10 000 iterations per one); also, the EM algorithm has been repeated 10 times from different starting points to obtain more information about ML convergence.

By applying MI methods we assume that data are MAR conditional on the imputation model and all of the variables in the model are jointly multivariate normal density. Nevertheless, using Kolmogorov–Smirnov tests, we have tested that all continuous predictors are non-normal, thus the multivariate normal hypothesis is not fulfilled; it could lead to biased results in presence of multiple categorical and continuous variables (Horton *et al*, 2003). To contrast the MCAR assumption, Pearson correlation coefficients were calculated between any pair of dichotomous continuous features with missing values; the existence of many correlated pairs of predictors allows the presence of an MCAR process to be rejected. To contrast the MAR assumption, real correlations between continuous variables with and without missing values were obtained; the majority presence of significant correlations allows the existence of an MAR process to be assumed (Table 5). Also, analysed missing values are non-monotone. The presence of an MAR process together with non-normal variables and non-monotone missing values suggest that MI methods could be particularly efficient.

Appendix Tables A1–A6 summarise results, including β estimates, standard errors, odds ratio, p values and 95% intervals for β estimates. For MI, standard errors were computed considering the weighted average of the estimated variances from within each completed data set plus the sample variance in the point estimates across the data sets. Table 6 also provides a comparison of main similarities and differences among methods in terms of significant β estimates (p values < 0.05).

From Table 6, we see that MI models produce quite similar and comparable results for this data set indicating that they yield a similar performance. In particular, EMis and EMB algorithms achieve similar results in terms of coefficients, signs and significant features because both approaches are based on similar hypotheses. Also, we notice that the average standard error in the EMis method is smaller than that on the EMB algorithm, but confidence intervals are similar. The greatest differences between both models appear in categorical attributes (A6, A12). Regarding the IP algorithm, some differences are observed in some categorical attributes (A4, A6, A12) even if similar significant features are achieved and average standard error is smaller than those from EMis and EMB alternatives. The EM algorithm generates results close to EMis in terms of signs and significant features but estimates and standard errors are greater indicating a higher instability of coefficients which could lead to worse prediction accuracy. On the contrary, the CC and mean substitution estimates are quite different from the other models as far as categorical attributes is concerned, indicating the presence of possible bias in predicted parameters. The CC method generates biased coefficients for categorical attributes (A5, A6, A8) together with a different sign of A12 categories. Also, the estimates and average standard error for both continuous and categorical attributes in the CC method are greater than those in the other methods. This result was also observed by Ibrahim *et al* (2005) who partially explain it considering that by omitting

Table 6 Models for dealing with missing values. Comparative results

Variable	*Listwise Deletion*	*Mean Substitution*	*EM*	*IP*	*EMis*	*EMB*
A10	0.131 (0.060)	0.125 (0.057)	0.126 (0.058)	0.128 (0.059)	0.127 (0.057)	0.128 (0.059)
A13	−0.003 (0.001)	−0.002 (0.001)		−0.002 (0.001)	−0.003 (0.001)	−0.003 (0.001)
A14	0.001 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
[A5 = 1]	−6.549 (2.257)	−6.328 (2.195)				
[A5 = 2]		−2.562 (1.032)	−2.181 (1.705)	−2.633 (1.031)	−2.633 (1.031)	−2.323 (1.648)
[A5 = 3]	−3.134 (1.007)	−2.901 (0.966)	−3.268 (1.629)	−3.051 (0.991)	−2.881 (0.959)	−2.705 (1.240)
[A5 = 4]	−3.189 (0.964)	−2.988 (0.923)	−2.834 (0.961)	−2.932 (0.915)	−2.879 (0.914)	
[A5 = 5]	−7.052 (2.365)	−6.881 (2.311)				
[A5 = 6]	−2.790 (0.931)	−2.748 (0.897)	−2.710 (0.903)	−2.785 (0.893)	−2.766 (0.901)	−2.317 (1.807)
[A5 = 7]	−2.599 (0.994)	−2.441 (0.961)	−2.702 (1.250)	−2.384 (0.957)	−2.531 (0.964)	−2.362 (1.121)
[A5 = 8]	−2.478 (0.878)	−2.423 (0.843)	−2.402 (0.866)	−2.451 (0.841)	−2.414 (0.834)	−2.299 (0.950)
[A5 = 9]	−1.880 (0.914)	−1.779 (0.881)	−1.913 (0.953)	−1.839 (0.883)	−1.823 (0.873)	−1.775 (0.890)
[A5 = 11]	−2.471 (0.916)		−2.135 (0.901)	−2.264 (0.878)	−2.236 (0.872)	−2.152 (0.967)
[A6 = 1]	5.890 (2.653)	−3.670 (0.341)				
[A6 = 3]	7.976 (2.785)					
[A6 = 5]	3.843 (1.793)					
[A6 = 6]	6.748 (2.386)		4.366 (2.344)	5.691 (2.361)		5.557 (2.658)
[A6 = 7]		−3.348 (0.971)				
[A6 = 8]	4.060 (1.775)	−3.302 (0.894)		3.511 (1.688)		3.480 (1.742)
[A8 = 0]	−3.836 (0.363)	5.384 (2.545)	−2.197 (3.618)	−3.681 (0.354)	−3.627 (0.337)	−3.705 (0.351)
[A9 = 0]		7.595 (2.672)				
[A11 = 0]		3.474(1.670)				
[A11 = 1]		6.285 (2.276)				
[A12 = 1]1	15.312 (0.513)		−3.650 (1.203)	−3.765 (1.315)	−3.479 (1.423)	
[A12 = 2]		3.729 (1.653)	−3.704 (1.120)	−3.760 (1.281)	−3.506 (1.417)	

Note: Only significant B are included (p value < 0.05) and their standard error (in brackets).

Table 7 Error estimates*

Method	*Overall error*	*Type I error*	*Type II error*
CC	84.00 (12.17%)	32.00 (10.42%)	52.00 (13.58%)
MS	82.00 (11.88%)	29.00 (9.45%)	53.00 (13.84%)
EM	80.40 (11.65%)	30.20 (9.84%)	50.20 (13.10%)
IP	82.67 (11.98%)	30.67 (9.99%)	52.00 (13.58%)
EMis	80.33 (11.64%)	30.00 (9.77%)	50.33 (13.14%)
EMB	80.00 (11.59%)	30.40 (9.90%)	49.60 (12.95%)

*Type I error refers to actual creditworthy applicants who were classified as not creditworthy; Type II error refers to actual not creditworthy applicants who were classified as creditworthy.

the case with missing values, much information is lost on the completely observed covariates resulting in less efficient and biased estimates for the regression coefficients. Concerning the mean substitution approach, it obtains the highest number of statistically significant features of all methods; also, differences on signs due to biased estimates are observed for A6, A8 and A12 variables. The 95% confidence intervals are very wide which indicates a high uncertainty in estimates that could affect final inference accuracy, as expected (see Table A2). To complete previous comparison of methods for handling missing data, Table 7 includes the errors obtained for each method.

Results from Table 7 show that EMis and EMB algorithms are the most accurate techniques for dealing with missing data in terms of the classification hit ratio. Also, they obtain the most balanced results in terms of type I and type II errors. The EM algorithm, which is the basic theory for both variants, also performed quite well even if parameter estimates could be partially biased. On the contrary, the CC approach generates the highest error rates. Mean substitution and IP alternatives only partially solved the missing value problem in terms of classification accuracy. Considering bias avoidance, stability of parameter estimates and accuracy measures, MI models are very promising techniques for dealing with missing values in the presence of sparse credit risk data sets. In particular, the EMis and EMB approaches are very robust techniques that permit accurate and robust models to be built even with the existence of categorical attributes with missing values.

Conclusions and remarks

The IRB approach for credit risk management allows banks to use their own internal measures for key risk components, as primary inputs to the minimum capital requirements calculate. The PD is the most significant credit risk component to be estimated for banks in developing IRB systems. If

the internal default experience is considered for PD estimates, an extensive database with a sufficient number of default experiences is necessary to obtain the requisite statistical validation. Nevertheless, internal databases are usually incomplete or do not contain adequate history to estimate the PD. The existence of missing values becomes more critical in the presence of sparse data portfolios because the limited default history can cause average observed default rates to be statistically unreliable estimators of PD for IRB systems.

In order to solve this problem, and to improve data quality and consistency, several methods could be applied to deal with missing values, grouped into six main categories: (1) CC; (2) substitution approaches and other 'ad hoc' methods; (3) ML; (4) MI; (5) weighting methods and (6) FB approaches. Even if CC has been profusely applied in practice because of its simplicity and comparability, the loss of information derived from deleting incomplete cases generates accuracy problems and biased parameter estimates. As alternatives, the remaining five categories could be used to deal with missing data. However, no theoretical rules are provided as to the best approach to be used for missing value analyses. Also, it is thought that final accuracy and practical significance greatly depends on the nature of analysed missing data in terms of randomness (MCAR, MAR, MNAR) and monotony (monotone or non-monotone).

In this paper, we analysed the nature of missing data together with the robustness, stability, bias and accuracy of a wide group of techniques for dealing with missing values with existing sparse data credit risk portfolios. Such techniques considered CC, mean substitution, ML and MI models, including Bayesian and MCMC algorithms. The empirical analysis was focused on the Australian credit approval data set, which includes many different categorical and continuous predictors related to a binary classification task.

Results indicated that the CC approach generated large bias and high error rates. Also, other simple approaches such as mean substitution did not perform well in presence of categorical attributes. Finally, the ML approach (EM algorithm) and several MIs techniques based on the EM theoretical approach (EMis and EMB models) obtained very promising results in terms of unbiased and stable parameter estimates, model robustness and classification accuracy.

Acknowledgements— The author gratefully acknowledges the helpful comments and questions of two anonymous reviewers.

References

Anderson AB, Basilevski A and Hum DPJ (1983). Missing data: A review of the literature. In: Rossi PE, Wright JD and Anderson AB (eds). *Handbook of Survey Research*. Academic Press: New York, pp 415–494.

Baesens B, Viaene S, van Gestel TM, Suykens JAL, Dedene G, de Moor B and Vanthienen J (2000). *An Empirical Assessment of KerneylType Performance for Least Squares SVM Classifiers*. Dept. Applied Economic Sciences, Katholieke Universiteit Leuven: Leuven, Belgium.

Basel Committee on Banking Supervision (BCBS) (2004). *International Convergence of Capital Measurements and Capital Standards. A Revised Framework*. Bank for International Settlements: Basel, June.

Basel Committee on Banking Supervision (BCBS) (2005a). *Studies on the Validation of Internal Rating Systems*. Working Paper no. 14, Bank for International Settlements, Basel, February.

Basel Committee on Banking Supervision (BCBS) (2005b). *Validation of low-default portfolios in the Basel II Framework*. Newsletter no. 6, Bank for International Settlements, Basel, September.

Basel Committee on Banking Supervision (BCBS) (2006). *The IRB Use Test*: *Background and Implementation*. Newsletter no. 9, Bank for International Settlements, Basel, September.

Benjamin N, Catheart A and Ryan K (2006). *Low default portfolios*: *A proposal for conservative estimation of default probabilities*. Working Paper, Financial Services Authority, London, April.

British Bankers' Association (BBA) (2004). *The IRB Approach for Low Default Portfolios (LDPs)*. BBA: London, August.

Carey M and Hrycay M (2001). Parameterizing credit risk models with rating data. *J Bank Financ* **25**: 197–270.

Carpenter J, Kenward M, Evans S and White I (2004). Last observation carry-forward and last observation analysis. *Stat Med* **23**: 3241–3244.

Carpenter J, Kenward M and Vansteelandt S (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J R Stat Soc Ser A* **169**: 571–584.

Cavaretta MJ and Chellapilla L (1999). Data mining using genetic programming: The implications of parsimony on generalization error. In: Angeline PJ, Michalewicz Z, Schoenauer M, Yao X and Zalzala A (eds) *Proceedings of the Congress on Evolutionary Computation* 2, IEEE Press: Washington, DC, pp 1330–1337.

Cheng TG and Victoria-Feser MP (2000). *Robust correlation estimation with missing data. Cahiers du department d'econométrie n*. 2000-5, University of Geneva.

Collins LM, Schafer JL and Kam CM (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* **6**: 330–351.

DeLeeuw E, Hox J and Huisman M (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics* **19**: 153–176.

Dempster AP, Laird NM and Rubin DB (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**: 1–22.

Eggermont J, Kok JN and Kosters WA (2004). Genetic programming for data classification: Partitioning the search space. In: Haddad HM, Omicini A, Wainwright RL and Liebrock LM (eds) *Proceedings of the 2004 Symposium on Applied Computing*, ACM Press: NewYork, pp 1001–1005.

European Central Bank (ECB) (2004). *Credit Risk Transfer by EU Banks*: *Activities*, *Risks and Risk Management*. European Central Bank: Frankfurt am Main, May.

Financial Services Authority (FSA) (2005). *Expert Group Paper on Low Default Portfolios*. FSA: London, August.

Hair JF, Anderson RE, Tatham RL and Black WC (1999). *Multivariate analysis*. Prentice-Hall: New York.

Honaker J and King G (2006). *What to do about missing values in time series cross-section data*. Working Paper, Harvard University, September.

Horton NJ and Kleinman KP (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* **61**: 79–90.

Horton NJ, Lipstitz SR and Parzen M (2003). A potential for bias when rounding in multiple imputation. *Am Stat* **57**: 229–232.

Huang CL, Chen MC and Wang CJ (2006). Credit scoring with a data mining approach based on support vector machines. *Expert Sys Appl* **33**: 847–856.

Ibrahim JG (1990). Incomplete data in generalized linear models. *J Am Stat Assoc* **85**: 765–769.

Ibrahim JG, Chen MH, Lipsitz SR and Herring AH (2005). Missing data methods for generalized linear models: A comparative review. *J Am Stat Assoc* **100**(469): 332–346.

Jacobson T and Roszbach K (2003). Bank lending policy, credit scoring and value-at-risk. *J Ban Finan* **27**: 615–633.

Jansen I, Beunckens C, Molenberghs G, Verbeke G and Malinckrodt C (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Stat Sci* **21**: 222–230.

King G, Honaker J, Joseph A and Scheve K (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am Polit Sci Rev* **95**(1): 49–69.

Little JR (1988). Missing-data adjustments in large surveys (with discussion). *J Bus Econ Stat* **6**: 287–301.

Little JR and Rubin D (2002). *Statistical Analysis with Missing Data*. Wiley: New York.

Little JR and Schenker N (1995). Missing data. In: Arminger G, Clogg CC and Sobel ME (eds). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum: New York, pp 39–75.

Moons KGM, Donders RA, Stijnen T and Harrell FE (2006). Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* **59**: 1092–1101.

Oesterreichische Nationalbank (OeNB) (2004). *Rating Models and Validation. Guidelines on Credit Risk Management*. Oesterreischische Nationalbank and Austrian Financial Authority: Vienna, November.

Pluto K and Tasche D (2004). *Estimating the probabilities of default for low default portfolios*. Working Paper, Deustche Bundesbank, April.

Quinlan RS (1979). Discovering rules by induction from large collections of examples. In: Michie D (ed). *Expert Systems in the Microelectronic Age*. Edinburgh University Press: Edinburgh, pp 168–201.

Quinlan RS (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann: New York.

Rocke DM (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Ann Stat* **24**: 1327–1345.

Rubin DB (1976). Inference and missing data. *Biometrika* **63**: 581–590.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons: New York.

Schafer JL (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall: New York.

Schuermann T and Hanson S (2005). *Confidence intervals for probabilities of default*. Working Paper, Federal Reserve Bank of New York, July.

Sherman RP (2000). Tests of certain types of ignorable nonresponse in surveys subject to item nonresponse or attrition. *Am J Polit Sci* **44**: 356–368.

Staten ME and Cate FH (2003). The impact of Opt-In Privacy Rules on retail credit markets: A case study of MBNA. *Duke Law J 52*: 745–786; June.

van Buuren S, Boshuizen HC and Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* **18**: 681–694.

Appendix A

Tables A1–A6.

Table A1 Models for dealing with missing values. Results for Listwise Deletion

	Listwise deletion (CC)					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
Intercept	−11.052	11.901	0.353			
A2	0.013	0.014	0.356	1.013	−0.015	0.041
A3	−0.024	0.030	0.420	0.976	−0.082	0.034
A7	0.081	0.055	0.146	1.084	−0.028	0.189
A10	**0.131**	**0.060**	**0.029**	**1.140**	**0.014**	**0.248**
A13	**−0.003**	**0.001**	**0.007**	**0.997**	**−0.004**	**−0.001**
A14	**0.001**	**0.001**	**0.005**	**1.001**	**0.000**	**0.001**
[A1 = 0]	0.078	0.323	0.810	1.081	−0.556	0.712
[A1 = 1]	0.000	.	.	.		
[A1 = 1]	−5.635	11.725	0.631	0.004	−28.616	17.345
[A4 = 2]	−4.790	11.721	0.683	0.008	−27.762	18.182
[A4 = 3]	0.000	.	.	.		
[A5 = 1]	**−6.549**	**2.257**	**0.004**	**0.001**	**−10.973**	**−2.125**
[A5 = 2]	−2.004	1.101	0.069	0.135	−4.161	0.153
[A5 = 3]	**−3.134**	**1.007**	**0.002**	**0.044**	**−5.109**	**−1.160**
[A5 = 4]	**−3.189**	**0.964**	**0.001**	**0.041**	**−5.079**	**−1.300**
[A5 = 5]	**−7.052**	**2.365**	**0.003**	**0.001**	**−11.687**	**−2.416**
[A5 = 6]	**−2.790**	**0.931**	**0.003**	**0.061**	**−4.615**	**−0.966**
[A5 = 7]	**−2.599**	**0.994**	**0.009**	**0.074**	**−4.548**	**−0.650**
[A5 = 8]	**−2.478**	**0.878**	**0.005**	**0.084**	**−4.198**	**−0.758**
[A5 = 9]	**−1.880**	**0.914**	**0.040**	**0.153**	**−3.672**	**−0.089**
[A5 = 10]	−0.712	1.358	0.600	0.491	−3.373	1.950
[A5 = 11]	**−2.471**	**0.916**	**0.007**	**0.084**	**−4.267**	**−0.676**

Table A1 (*Continued*)

	Listwise deletion (CC)					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
[A5 = 12]	−4.674	4.446	0.293	0.009	−13.389	4.040
[A5 = 13]	−1.360	1.037	0.190	0.257	−3.393	0.673
[A5 = 14]	0.000	.	.	.		
[A6 = 1]	**5.890**	**2.653**	**0.026**	**361.533**	**0.690**	**11.091**
[A6 = 2]	2.598	1.985	0.191	13.433	−1.294	6.489
[A6 = 3]	**7.976**	**2.785**	**0.004**	**2911.544**	**2.518**	**13.435**
[A6 = 4]	3.553	1.824	0.051	34.910	−0.023	7.128
[A6 = 5]	**3.843**	**1.793**	**0.032**	**46.688**	**0.330**	**7.357**
[A6 = 6]	**6.748**	**2.386**	**0.005**	**852.140**	**2.072**	**11.423**
[A6 = 7]	−11.478	7128.257	0.999	0.000		
[A6 = 8]	**4.060**	**1.775**	**0.022**	**57.957**	**0.582**	**7.538**
[A6 = 9]	0.000	.	.	.		
[A8 = 0]	**−3.836**	**0.363**	**0.000**	**0.022**	**−4.547**	**−3.124**
[A8 = 1]	0.000	.	.	.		
[A9 = 0]	−0.417	0.382	0.275	0.659	−1.165	0.332
[A9 = 1]	0.000	.	.	.		
[A11 = 0]	0.242	0.289	0.401	1.274	−0.323	0.808
[A11 = 1]	0.000	.	.	.		
[A12 = 1]	**15.312**	**0.513**	**0.00**	**4467302.058**	**14.307**	**16.317**
[A12 = 2]	15.407	0.000		4912834.337	15.407	15.407
[A12 = 3]	0.000				−0.015	0.041

In bold: significant attributes ($p < 0.05$).

Table A2 Models for dealing with missing values. Results for Mean Substitution

	Mean Substitution (MS)					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
Intercept	7.322	7.810	0.349			
A2	0.012	0.013	0.351	1.012	−0.013	0.037
A3	−0.018	0.029	0.542	0.983	−0.074	0.039
A7	0.057	0.049	0.245	1.059	−0.039	0.154
A10	**0.125**	**0.057**	**0.029**	**1.134**	**0.013**	**0.238**
A13	**−0.002**	**0.001**	**0.006**	**0.998**	**−0.004**	**−0.001**
A14	**0.000**	**0.000**	**0.014**	**1.000**	**0.000**	**0.001**
[A1 = 0]	0.082	0.308	0.790	1.085	−0.521	0.685
[A1 = 1]	0.000	.	.	.		
[A4 = 1]	−4.881	7.526	0.517	0.008	−19.631	9.869
[A4 = 2]	−4.078	7.519	0.588	0.017	−18.816	10.660
[A4 = 3]	0.000	.	.	.		
[A5 = 1]	**−6.328**	**2.195**	**0.004**	**0.002**	**−10.630**	**−2.026**
[A5 = 2]	**−2.562**	**1.032**	**0.013**	**0.077**	**−4.585**	**−0.539**
[A5 = 3]	**−2.901**	**0.966**	**0.003**	**0.055**	**−4.793**	**−1.008**
[A5 = 4]	**−2.988**	**0.923**	**0.001**	**0.050**	**−4.797**	**−1.179**
[A5 = 5]	**−6.881**	**2.311**	**0.003**	**0.001**	**−11.411**	**−2.351**
[A5 = 6]	**−2.748**	**0.897**	**0.002**	**0.064**	**−4.507**	**−0.989**
[A5 = 7]	**−2.441**	**0.961**	**0.011**	**0.087**	**−4.325**	**−0.557**
[A5 = 8]	**−2.423**	**0.843**	**0.004**	**0.089**	**−4.076**	**−0.770**
[A5 = 9]	**−1.779**	**0.881**	**0.043**	**0.169**	**−3.506**	**−0.053**
[A5 = 10]	−0.806	1.269	0.525	0.447	−3.294	1.681
[A5 = 11]	−2.259	0.880	0.010	0.104	−3.984	−0.533
[A5 = 12]	− 5.003	4.149	0.228	0.007	− 13.135	3.129
[A5 = 13]	−1.279	1.000	0.201	0.278	−3.238	0.681
[A5 = 14]	0.000	.	.	.		
[A6 = 1]	**−3.670**	**0.341**	**0.00**	**0.025**	**−4.339**	**−3.001**
[A6 = 2]	0.000	.	.	.		
[A6 = 3]	−0.620	0.369	0.093	0.538	−1.343	0.103

Table A2 (*Continued*)

	Mean Substitution (MS)					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
[A6 = 4]	0.000	.	.	.		
[A6 = 5]	0.270	0.279	0.333	1.310	−0.277	0.816
[A6 = 6]	0.000	.	.	.		
[A6 = 7]	**−3.348**	**0.971**	**0.001**	**0.035**	**−5.252**	**−1.444**
[A6 = 8]	**−3.302**	**0.894**	**0.00**	**0.037**	**−5.053**	**−1.550**
[A6 = 9]	0.000	.	.	.		
[A8 = 0]	**5.384**	**2.545**	**0.034**	**217.837**	**0.396**	**10.371**
[A8 = 1]	2.400	1.908	0.208	11.021	−1.339	6.139
[A9 = 0]	**7.595**	**2.672**	**0.004**	**1988.075**	**2.357**	**12.833**
[A9 = 1]	3.079	1.687	0.068	21.734	−0.228	6.385
[A11 = 0]	**3.474**	**1.670**	**0.037**	**32.280**	**0.201**	**6.748**
[A11 = 1]	**6.285**	**2.276**	**0.006**	**536.540**	**1.824**	**10.746**
[A12 = 1]	−13.627	0.000		0.000	−13.627	−13.627
[A12 = 2]	**3.729**	**1.653**	**0.024**	**41.651**	**0.490**	**6.968**
[A12 = 3]	0.000	.	.	.		

Table A3 Models for dealing with missing values. Results for EM algorithm

	EM					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
A2	0.005	0.014	0.696	1.005	−0.016	0.026
A3	−0.019	0.029	0.497	0.981	−0.067	0.028
A7	0.063	0.050	0.203	1.066	−0.018	0.145
A10	**0.126**	**0.058**	**0.028**	**1.134**	**0.032**	**0.220**
A13	-0.045	0.120	0.093	0.959	-0.137	0.047
A14	**0.000**	**0.000**	**0.010**	**1.000**	**0.000**	**0.001**
[A1 = 0]	0.074	0.261	0.601	1.079	−0.331	0.478
[A1 = 1]	0.419	13.686	0.878	8.126	−22.027	22.865
[A4 = 1]	−7.220	11.409	0.428	1.604	−19.358	4.919
[A4 = 2]	−7.006	9.677	0.480	0.006	−16.428	2.415
[A4 = 3]	−1.221	8.432	0.437	0.002	−14.149	11.708
[A5 = 1]	−4.183	3.955	0.225	0.044	−9.733	1.366
[A5 = 2]	**−2.181**	**1.705**	**0.008**	**0.066**	**−3.852**	**−0.510**
[A5 = 3]	**−3.268**	**1.629**	**0.024**	**0.046**	**−5.492**	**−1.043**
[A5 = 4]	**−2.834**	**0.961**	**0.004**	**0.060**	**−4.356**	**−1.312**
[A5 = 5]	−4.212	2.702	0.059	0.030	−7.253	−1.171
[A5 = 6]	**−2.710**	**0.903**	**0.002**	**0.067**	**−4.173**	**−1.248**
[A5 = 7]	**−2.702**	**1.250**	**0.018**	**0.073**	**−4.554**	**−0.851**
[A5 = 8]	**−2.402**	**0.866**	**0.005**	**0.092**	**−3.787**	**−1.017**
[A5 = 9]	**−1.913**	**0.953**	**0.035**	**0.153**	**−3.369**	**−0.456**
[A5 = 10]	−1.367	1.360	0.307	0.289	−3.278	0.544
[A5 = 11]	**−2.135**	**0.901**	**0.017**	**0.121**	**−3.564**	**−0.706**
[A5 = 12]	−2.610	3.441	0.417	0.140	−7.732	2.512
[A5 = 13]	−1.466	1.093	0.164	0.247	−3.053	0.120
[A5 = 14]	−0.595	4.375	0.470	0.051	−7.362	6.172
[A6 = 1]	1.848	3.603	0.374	40.376	−1.866	5.563
[A6 = 2]	2.081	2.241	0.154	14.086	−0.867	5.028
[A6 = 3]	4.584	3.245	0.123	387.712	0.379	8.789
[A6 = 4]	2.530	1.753	0.148	13.019	−0.299	5.359
[A6 = 5]	3.310	2.066	0.073	41.191	0.335	6.285
[A6 = 6]	**4.366**	**2.344**	**0.042**	**112.837**	**0.989**	**7.743**
[A6 = 7]	−11.795	963.142	0.526	4.865	−1595.945	1572.354
[A6 = 8]	3.415	1.834	0.052	34.015	0.539	6.290
[A6 = 9]	−3.178	7.946		0.000		
[A8 = 0]	**−2.197**	**3.618**	**0.006**	**6.717**	**−3.177**	**−1.216**
[A8 = 1]	0.000					

Table A3 (*Continued*)

	EM					
	B	*SE*	*p value*	*odds ratio*	95% *interval (lower limit)*	95% *interval (upper limit)*
[A9 = 0]	−1.192	1.534	0.085	0.447	−1.785	−0.599
[A9 = 1]	0.000	0.000				
[A11 = 0]	0.063	0.531	0.329	1.123	−0.420	0.546
[A11 = 1]	0.000	0.000				
[A12 = 1]	**−3.650**	**1.203**	**0.003**	**0.028**	**−5.493**	**−1.807**
[A12 = 2]	**−3.704**	**1.120**	**0.001**	**0.026**	**−5.431**	**− 1.978**
[A12 = 3]	0.000					

Table A4 Models for dealing with missing values. Results for IP algorithm

	IP					
	B	*SE*	*p value*	*odds ratio*	95% *interval (lower limit)*	95% *interval (upper limit)*
Intercept	16.578	8.083	0.075		9.430	23.727
A2	0.005	0.014	0.710	1.005	−0.017	0.027
A3	−0.019	0.029	0.509	0.981	−0.066	0.028
A7	0.062	0.051	0.020	1.064	−0.020	0.145
A10	**0.128**	**0.059**	**0.029**	**1.136**	**0.032**	**0.223**
A13	**−0.002**	**0.001**	**0.006**	**0.998**	**−0.004**	**−0.001**
A14	**0.000**	**0.000**	**0.006**	**1.000**	**0.000**	**0.001**
[A1 = 0]	0.124	0.308	0.686	1.133	−0.381	0.630
[A1 = 1]	0.000	.	.	.		
[A4 = 1]	−13.280	7.649	0.137	0.000	−18.132	−8.428
[A4 = 2]	−12.489	7.561	0.467	0.001	−16.976	−8.002
[A4 = 3]	0.000	.	.	.		
[A5 = 1]	−4.796	3.257	0.122	0.015	−8.959	−0.633
[A5 = 2]	**−2.633**	**1.031**	**0.011**	**0.072**	**−4.328**	**−0.937**
[A5 = 3]	**−3.051**	**0.991**	**0.002**	**0.048**	**−4.633**	**−1.470**
[A5 = 4]	**−2.932**	**0.915**	**0.001**	**0.053**	**−4.436**	**−1.428**
[A5 = 5]	−5.078	3.933	0.080	0.044	−8.342	−1.813
[A5 = 6]	**−2.785**	**0.893**	**0.002**	**0.062**	**−4.252**	**−1.318**
[A5 = 7]	**− 2.384**	**0.957**	**0.012**	**0.092**	**−3.947**	**−0.820**
[A5 = 8]	**−2.451**	**0.841**	**0.004**	**0.086**	**−3.832**	**−1.069**
[A5 = 9]	**−1.839**	**0.883**	**0.036**	**0.159**	**−3.281**	**−0.396**
[A5 = 10]	−0.763	1.262	0.545	0.467	−2.832	1.305
[A5 = 11]	**−2.264**	**0.878**	**0.010**	**0.104**	**−3.705**	**−0.822**
[A5 = 12]	−4.301	4.234	0.305	0.016	−11.136	2.534
[A5 = 13]	−1.288	0.998	0.197	0.276	−2.930	0.354
[A5 = 14]	0.000	.	.	.		
[A6 = 1]	3.821	3.407	0.232	93.536	−0.974	8.617
[A6 = 2]	2.039	1.941	0.293	7.717	−1.145	5.223
[A6 = 3]	5.489	3.707	0.067	660.203	1.511	9.466
[A6 = 4]	3.000	1.721	0.082	20.099	0.171	5.829
[A6 = 5]	3.214	1.708	0.060	24.980	0.415	6.012
[A6 = 6]	**5.691**	**2.361**	**0.012**	**331.527**	**2.019**	**9.363**
[A6 = 7]	−14.149	2417.436	0.996	0.000	−3990.475	3962.177
[A6 = 8]	**3.511**	**1.688**	**0.038**	**33.573**	**0.740**	**6.281**
[A6 = 9]	0.000	.	.	.		
[A8 = 0]	**−3.681**	**0.354**	**0.000**	**0.025**	**−4.243**	**−3.120**
[A8 = 1]	0.000	.	.	.		
[A9 = 0]	−0.564	0.373	0.130	0.569	−1.174	0.047
[A9 = 1]	0.000	.	.	.		
[A11 = 0]	0.254	0.284	0.366	1.290	−0.205	0.713
[A11 = 1]	0.000	.	.	.		
[A12 = 1]	**−3.765**	**1.315**	**0.002**	**0.026**	**−5.621**	**−1.909**
[A12 = 2]	**−3.760**	**1.281**	**0.002**	**0.026**	**−5.523**	**−1.998**
[A12 = 3]	0.000	.	.	.		

Table A5 Models for dealing with missing values. Results for EMis algorithm

	EMis					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
Intercept	9.324	8.527	0.258		−3.859	22.507
A2	0.008	0.013	0.560	1.008	−0.014	0.029
A3	−0.020	0.029	0.477	0.980	−0.067	0.026
A7	0.061	0.049	0.216	1.063	−0.020	0.142
A10	**0.127**	**0.057**	**0.026**	**1.135**	**0.033**	**0.220**
A13	**−0.003**	**0.001**	**0.005**	**0.997**	**−0.004**	**−0.001**
A14	**0.000**	**0.000**	**0.012**	**1.000**	**0.000**	**0.001**
[A1 = 0]	0.071	0.329	0.804	1.077	−0.430	0.572
[A1 = 1]	0.000	.	.	.		
[A4 = 1]	−6.058	7.906	0.439	0.004	−18.827	6.711
[A4 = 2]	−5.251	7.901	0.503	0.009	−18.011	7.509
[A4 = 3]	0.000	.	.	.		
[A5 = 1]	−4.750	3.152	0.120	0.015	−8.940	−0.559
[A5 = 2]	**−2.631**	**1.023**	**0.010**	**0.072**	**−4.307**	**−0.954**
[A5 = 3]	**−2.881**	**0.959**	**0.003**	**0.056**	**−4.448**	**−1.314**
[A5 = 4]	**−2.879**	**0.914**	**0.002**	**0.056**	**−4.376**	**−1.383**
[A5 = 5]	−5.070	3.963	0.100	0.047	−8.299	−1.841
[A5 = 6]	**−2.766**	**0.901**	**0.002**	**0.063**	**−4.224**	**−1.308**
[A5 = 7]	**−2.531**	**0.964**	**0.008**	**0.080**	**−4.093**	**−0.969**
[A5 = 8]	**−2.414**	**0.834**	**0.004**	**0.089**	**−3.785**	**−1.044**
[A5 = 9]	**−1.823**	**0.873**	**0.037**	**0.162**	**−3.255**	**−0.391**
[A5 = 10]	−1.157	1.338	0.370	0.330	−3.214	0.901
[A5 = 11]	**−2.236**	**0.872**	**0.010**	**0.107**	**−3.668**	**−0.804**
[A5 = 12]	−3.561	4.288	0.385	0.053	−9.992	2.870
[A5 = 13]	−1.298	0.988	0.189	0.273	−2.923	0.326
[A5 = 14]	0.000	.	.	.		
[A6 = 1]	3.567	3.514	0.259	90.085	−1.135	8.270
[A6 = 2]	2.799	1.896	0.131	18.012	−0.136	5.734
[A6 = 3]	5.024	4.154	0.128	656.505	1.254	8.794
[A6 = 4]	2.626	1.765	0.125	14.839	−0.128	5.379
[A6 = 5]	3.004	1.748	0.077	21.518	0.269	5.740
[A6 = 6]	4.633	2.730	0.056	161.101	1.080	8.186
[A6 = 7]	−15.686	2.529				
[A6 = 8]	3.263	1.719	0.052	27.585	0.556	5.971
[A6 = 9]	0.000	.	.	.		
[A8 = 0]	**−3.627**	**0.337**	**0.000**	**0.027**	**−4.178**	**−3.075**
[A8 = 1]	0.000	.	.	.		
[A9 = 0]	−0.595	0.367	0.105	0.552	−1.199	0.008
[A9 = 1]	0.000	.	.	.		
[A11 = 0]	0.231	0.282	0.407	1.261	−0.223	0.685
[A11 = 1]	0.000	.	.	.		
[A12 = 1]	**−3.479**	**1.423**	**0.007**	**0.035**	**−5.501**	**−1.458**
[A12 = 2]	**−3.506**	**1.417**	**0.004**	**0.035**	**−5.399**	**−1.613**
[A12 = 3]	0.000	.	.	.		

Table A6 Models for dealing with missing values. Results for EMB algorithm

	EMB					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
Intercept	13.456	9.826	0.128		5.929	20.983
A2	0.008	0.014	0.572	1.008	−0.014	0.029
A3	−0.020	0.029	0.480	0.980	−0.068	0.027
A7	0.062	0.051	0.218	1.064	−0.020	0.145

Table A6 (*Continued*)

	EMB					
	B	*SE*	*p value*	*odds ratio*	*95% interval (lower limit)*	*95% interval (upper limit)*
A10	**0.128**	**0.059**	**0.028**	**1.137**	**0.032**	**0.224**
A13	**−0.003**	**0.001**	**0.005**	**0.997**	**−0.004**	**−0.001**
A14	**0.000**	**0.000**	**0.007**	**1.000**	**0.000**	**0.001**
[A1 = 0]	0.069	0.318	0.803	1.074	−0.442	0.580
[A1 = 1]	0.000	.	.	.		
[A4 = 1]	−12.337	9.584	0.204	0.018	−18.566	−6.107
[A4 = 2]	−11.510	9.537	0.367	0.039	−17.563	−5.458
[A4 = 3]	0.000	.	.	.		
[A5 = 1]	−4.505	2.695	0.086	0.035	−7.914	−1.096
[A5 = 2]	**−2.323**	**1.648**	**0.042**	**0.361**	**−4.031**	**−0.615**
[A5 = 3]	**−2.705**	**1.240**	**0.040**	**0.108**	**−4.220**	**−1.189**
[A5 = 4]	−2.987	1.292	0.056	0.052	−5.077	−0.897
[A5 = 5]	−4.793	3.094	0.070	0.180	−8.064	−1.522
[A5 = 6]	**−2.317**	**1.807**	**0.005**	**0.724**	**−3.788**	**−0.845**
[A5 = 7]	**−2.362**	**1.121**	**0.039**	**0.118**	**−3.907**	**−0.817**
[A5 = 8]	**−2.299**	**0.950**	**0.014**	**0.113**	**−3.655**	**−0.942**
[A5 = 9]	**−1.775**	**0.890**	**0.040**	**0.175**	**−3.178**	**−0.371**
[A5 = 10]	−1.447	2.123	0.415	0.363	−3.650	0.757
[A5 = 11]	**−2.152**	**0.967**	**0.018**	**0.131**	**−3.547**	**−0.756**
[A5 = 12]	−3.585	4.073	0.346	0.095	−9.682	2.511
[A5 = 13]	−1.226	0.992	0.204	0.305	−2.783	0.331
[A5 = 14]	0.000	.	.	.		
[A6 = 1]	3.648	3.014	0.218	87.665	−0.486	7.781
[A6 = 2]	2.239	1.922	0.243	10.035	−0.852	5.330
[A6 = 3]	6.275	2.834	0.021	954.926	2.069	10.480
[A6 = 4]	2.808	1.772	0.111	17.935	−0.003	5.619
[A6 = 5]	3.214	1.762	0.066	26.943	0.427	6.001
[A6 = 6]	**5.557**	**2.658**	**0.046**	**373.967**	**1.785**	**9.328**
[A6 = 7]	−9.585	869.953	0.461	39.382	−1440.425	1421.256
[A6 = 8]	**3.480**	**1.742**	**0.044**	**35.087**	**0.722**	**6.239**
[A6 = 9]	0.000	.	.	.		
[A8 = 0]	**−3.705**	**0.351**	**0.000**	**0.025**	**−4.270**	**−3.140**
[A8 = 1]	0.000	.	.	.		
[A9 = 0]	−0.577	0.373	0.121	0.562	−1.188	0.033
[A9 = 1]	0.000	.	.	.		
[A11 = 0]	0.252	0.282	0.370	1.287	−0.208	0.712
[A11 = 1]	0.000	.	.	.		
[A12 = 1]	−1.742	5.185	0.057	12 013.938	−3.731	0.247
[A12 = 2]	−1.770	5.193	0.060	12 692.975	−3.595	0.055
[A12 = 3]	0.000	.	.	.		

Received December 2007;
accepted February 2009 after one revision