Article:
Stead, AD and Wheat, P (2020) The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using English local highway data. European Journal of Operational Research, 280 (1). pp. 59-77. ISSN: 0377-2217

https://doi.org/10.1016/j.ejor.2019.06.042

The case for the use of Multiple Imputation missing data methods in Stochastic Frontier Analysis with illustration using English Local Highway data

Alexander D. Stead and Phill Wheat

June 2019

Abstract

Multiple Imputation (MI) methods have been widely applied in economic applications as a robust statistical way to incorporate data where some observations have missing values for some variables. However in Stochastic Frontier Analysis (SFA), application of these techniques has been sparse and the case for such models has not received attention in the appropriate academic literature. This paper fills this gap and explores the robust properties of MI within the stochastic frontier context. From a methodological perspective, we depart from the standard MI literature by demonstrating, conceptually and through simulation, that it is not appropriate to use imputations of the dependent variable within the SFA modelling, although they can be useful to predict the values of missing explanatory variables. Fundamentally, this is because efficiency analysis involves decomposing a residual into noise and inefficiency and as a result any imputation of a dependent variable would be imputing efficiency based on some concept of average inefficiency in the sample. A further contribution that we discuss and illustrate for the first time in the SFA literature, is that using auxiliary variables (outside of those contained in the SFA model) can enhance the imputations of missing values. Our empirical example neatly articulates that often the source of missing data is only a sub-set of components comprising a part of a composite (or complex) measure and that the other parts that are observed are very useful in predicting the value.

Keywords: Production; stochastic frontier analysis; missing data; multiple imputation; efficiency analysis

---

A.D. Stead*
Room LG.06, Institute for Transport Studies, University of Leeds, Leeds, LS2 9LJ, UK
e-mail: a.d.stead@leeds.ac.uk
P. Wheat
Room 1.03, Institute for Transport Studies, University of Leeds, Leeds, LS2 9LJ, UK
e-mail: p.e.wheat@its.leeds.ac.uk
*Corresponding author

1 Introduction

Missing data is an issue that arises frequently in many statistical analyses, and many competing methods exist to deal with it. Our focus here is on the appropriateness of the application of the bank of techniques developed for missing data to stochastic frontier analysis (SFA), as missing data in this context has received relatively little attention to date. Our paper aims to make the case for Multiple Imputation (MI) methods in SFA. We highlight that MI methods are relatively easy to implement in SFA and explore, through a mix of theory, simulation and empirical example, the robustness of the parameter estimates and efficiency predictions obtained, compared to those from alternative methods.

By far the most common practice in the empirical literature is to proceed by only including observations if data on all explanatory variables are available, so called complete case analysis. However if data are taken from a variety of information sources, such a requirement can often be a high hurdle to clear, and it could be the case that an observation is excluded if only one of a number of explanatory factors is missing. This is the case in our empirical example which we draw on to illustrate the approach, where nearly 20% of observations have a missing value related to one attribute (road condition). Developing a solution, the focus of this paper, would ideally lead to more precise parameter estimates for the cost frontier and thus more precise efficiency predictions for the complete cases, and also provide efficiency predictions for the incomplete cases (observations with missing data).

The focus of this paper is on the plausibility of results obtained via MI, in terms of imputations, parameter estimates (both point estimate values and precision of estimation), and efficiency predictions. The plausibility of efficiency predictions is determined by the plausibility of the imputations, either directly in the case of those observations with missing values or indirectly as the imputations are used to estimate the frontier parameters. As such, a key consideration is whether the relationships identified between the observed variables and those variables with missing data are plausible. Indeed a positive feature of some MI methods, such as the Multiple Imputation by chained equations (MICE), is that they provide a clear way to evaluate such plausibility through estimation of imputation equations. Building on this, we explore the inclusion of auxiliary variables which have a clear *a priori* relationship with the missing data in the imputation model, and consider whether this results in improved imputations.

The layout of this paper is as follows: Section 2 discusses the key challenges that missing data imposes on efficiency analysis. Section 3 gives an overview of missing data methods in general, with a particular focus on imputation methods and why they are appropriate in the context of SFA. In Section 4, we undertake a Monte Carlo simulation study to explore the statistical robustness of MI techniques in SFA, focusing on comparison with results obtained via listwise deletion (LD), as well as considering whether imputed values of the dependent variable should be used in estimation. Section 5 outlines our empirical example of a stochastic cost frontier using data on English local authorities' highways maintenance

activities[1]. This sets out our missing data structure and the imputation models used, then discusses the parameter estimates resulting from the implementation of the various MI approaches and finally discusses the resulting efficiency predictions. Section 8 concludes and provides our recommendations for how MI methods should be used in SFA.

2 Missing data in efficiency analysis: Challenges

Efficiency analysis can be traced back to the seminal work by Farell (1957) who considered measurement of the extent to which decision making units failed to maximise output for given inputs (technical inefficiency) or when faced with a budget constraint (economic (or cost) efficiency).

Stochastic Frontier Analysis (SFA) is a parametric, statistical approach to efficiency analysis. Proposed by Aigner et al. (1977) and Meeusen and van Den Broeck (1977), a standard stochastic cost frontier model for cross sectional data can be written as

$$lnc_i = x_i'\beta + \varepsilon_i \quad i = 1, \dots, N \tag{1}$$

$$\varepsilon_i = u_i + v_i$$

where c is a cost variable x is a vector of independent variables, and β is a vector of coefficients. The cost frontier describes the efficient cost of producing a given vector of outputs with the cost-minimising combination of inputs, and is a function of output quantities and input prices (Shephard, 1953). Therefore x_i' should include ouput quantity and input price variables, as well as other 'hedonic' variables affecting costs. The error term ε is composed of a symmetric noise component v – usually assumed to follow a normal distribution – and an inefficiency component u, which assumed to follow some one-sided distribution, e.g. half normal or exponential. The subscript i denotes the observation number. Efficiency predictions are obtained following estimation of the frontier parameters, based on the distribution of $u_i|\varepsilon_i$. The usual approach is to take the mean of the distribution, as proposed by Jondrow et al. (1982) and Battese and Coelli (1988).

The formulation of SFA and the nature of a comparative efficiency benchmarking exercise present a number of challenges for dealing with missing data:

1) The analysis primarily focuses on predicting the efficiency for a given observation. This in turn requires prediction of the error in the model. In practice, the residual for an observation is used as a proxy for the unobserved realised error. Methods thus need to be able to predict the residual with precision, which requires precise parameter estimates. Furthermore, methods that do not allow for prediction of the residual for those observations with missing data are of limited use in this context.

[1] We provide a Stata .do file with a code template that can be used for the application of MI to SFA at https://www.its.leeds.ac.uk/bear

Deletion methods in response to missing data are the most commonly used methods for estimation in the literature. However, these do not use all the data available and so may not produce the most precise parameter estimates, and these estimates maybe biased if the data is not missing completely at random (MCAR) (see Section 3). As such listwise deletion (LD) may not predict the residual with greatest precision i.e. efficiency predictions even for observations with no missing data maybe imprecise. However, they are easily understood and implementable hence their popularity.

In the case of LD, there does also exist a pragmatic approach at the prediction stage of the modelling exercise to predict efficiency for observations with missing data by using some heuristics, such as linear interpolation. This has been used in applications such as earlier work on the dataset used in this study. Given the popularity of this approach, in our empirical example we contrast this approach with that from the multiple imputation approaches.

2) Given the focus on the residual, it is difficult to justify predicting efficiency for observations with missing values for the dependent variable in the context of performance benchmarking. This is because the difference between the dependent variable and the modelled relationship ($x_i'\beta$) is the residual prediction and which is, in turn, monotonically related to the efficiency prediction. Any imputation approach for the dependent variable would distort this difference (there would be some element of the difference that would reflect imputation error) and as such distort the efficiency measure.
3) However, as discussed in Section 3, there may be some benefit in including the dependent variable in the imputation phase of the missing data methods as the dependent variable could be a useful predictor for the missing independent variables. In our empirical study we do include the dependent variable as a covariate in the imputations models for the missing variables.
4) A more vexing question is whether the imputed missing dependent variable data should be used in the SFA model estimation i.e. to estimate the cost or production frontier, even if it is clear that efficiency should not be predicted for those observations. There may be precision gains to be had with respect to the frontier parameters (as there is more observations to base the estimation on), however the variance parameters are of central interest in efficiency analysis, and following reasoning along that discussed in 2) above, introducing observation with imputed dependent variables could bias the estimates of the variance parameters and distort efficiency predictions. We undertake a simulation study to investigate this issue.
5) In efficiency analysis, and specifically benchmarking, models are often developed with engineering stakeholders and comprise a simplified representation of the underlying cost relationship. The variables used in the model are often proxies for wider phenomena (e.g. asset condition in this application) and are underpinned by aggregation of more detailed data. Missing values can often arise because some, but not all, of the underlying data is missing for an observation. This is the case in our example where 'road condition' is an aggregate measure of the condition of various categories of road. As such there are often natural "auxiliary" predictors of missing data which are

ripe for exploitation. This is a compelling reason to use MI, since it exploits all the information available to the researcher.

3 Missing Data Methods

In this section we review the literature on missing data. We start by considering the broad set of methods available and motivate MI techniques. We review key results from the literature on the statistical properties of these techniques. We then discuss the two MI techniques that we consider appropriate for SFA. We also begin to address the issues considered in section 2, namely the use of additional auxiliary variables for imputation and whether observations with imputed values for the dependent variable should be used in the analysis stage. Finally in this section, we review the limited applications of MI in SFA applications.

3.1 Overview of Missing Data Methods

In general, the suitability of a given missing data method depends on the missing data mechanism and the pattern of the missing data. The missing data mechanism describes the relationship between 'missingness' and the missing and observed values. Data may be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). If data are MCAR, missingness is unrelated to the missing or observed values. If data are MAR, then missingness is unrelated to the missing values, but related to observed values. Finally, if the data are MNAR, missingness is related to the missing values, and possibly also to the observed values. Patterns of missing data can be univariate, meaning only one variable has missing values, monotone, meaning that when the value of variable p is missing, variables $p+1, \ldots, k$ are also missing, or there may be a general pattern of missing data.

Missing data approaches to linear modelling are grouped by Little (1992) into six categories: complete-case analysis, available-case analysis, analysis of imputed data, maximum likelihood (ML) methods, Bayesian methods, and multiple imputation (MI). The first two of these, also known as listwise deletion (LD) and pairwise deletion (PD), are based on the exclusion of missing data, while ML, Bayesian methods, and MI are described by Little (1992) and Little and Rubin (2002) as model based approaches. Given our context of efficiency analysis, and the desire to obtain efficiency predictions for observations with missing data, we mainly discuss imputation procedures, following a brief discussion of deletion-based and model-based procedures.

3.1.1 Deletion-based and Model-based Methods

The most common approach to dealing with missing data is LD, in which we simply exclude all observations containing missing values from the analysis. The advantage of this approach, aside from its simplicity, is that the resulting parameter estimates are not biased if the data are MCAR, since ordinary least squares (OLS) regression analysis – and this also applies to SFA – conditions on the

values of the covariates (Glynn and Laird, 1986). However, if the data are only MAR rather than MCAR, LD can result in bias since the relationship between the covariates and missingness is not accounted for. Even if the data are MCAR, LD has drawbacks in terms of the loss of information from dropping incomplete observations, and a resulting loss of precision in estimation. The greater the fraction of missing data, the greater the potential bias and loss of information from using LD.

PD attempts to overcome this disadvantage by using all of the available data. Regression with PD, as proposed by Glasser (1964), works by replacing the covariance σ_{jk} used to calculate estimated coefficients with the pairwise covariance $\sigma_{jk}^{(jk)}$ calculated using all observations for which x_j and x_k are present. For example, if we have a regression of y on k dependent variables, the estimated coefficients can be expressed in terms of variances and covariances like so

$$\beta = C^{-1}\Sigma'_{xy} \tag{2}$$

$$\beta_0 = \bar{y} - \bar{x}\beta$$

Where β is a vector of slope coefficients, C is a $k \times k$ matrix of variances and covariances between the independent variables, Σ_{xy} is a vector of covariances between the regressand and each of the regressors, β_0 is the estimated intercept, $\bar{y}$ is the mean of the dependent variable, and $\bar{x}$ is a vector containing the means of the independent variables. Under PD, we simply calculate each of the variances and covariances in C and $Cov(x_i, y)$ pairwise. Given that OLS yields unbiased estimates of the slope parameters in a frontier, this approach may be taken to estimate the frontier parameters.

Although PD makes use of all the available data, there are drawbacks. As pointed out by Little (1992), the covariance matrix of the X variables is not necessarily positive-definite, and indeterminate slope parameters can result when it is not. Haitovsky (1968) finds that this problem is most severe when the independent variables are highly correlated. There are also issues around the calculation of standard errors under pairwise deletion, discussed by Little (1992). However, the main disadvantage of both LD and PD in the context of efficiency analysis is the inability to obtain efficiency predictions for observations with missing data.

The ML approach to missing data, proposed by Rubin (1976), involves assuming a joint distribution for the data and integrating out the missing values to form a likelihood function. The distribution of the missing data mechanism can be taken into account, or may be ignored if the data are MAR and the parameters of the model and of the missing data mechanism have distinct spaces, in the sense that their joint parameter space is simply the product of their individual parameter spaces. Bayesian inference is based on the posterior distribution obtained by combining the same likelihood function with prior distributions for the parameters of the model and of the missing data mechanism.

The ML and Bayesian approaches are potentially promising, in that they could be used to obtain efficiency predictions for each observation. However, they are relatively demanding in that new models would have to be developed for each stochastic frontier specification. For this reason, we adopt the approach of imputation of missing values, which is more readily adapted to any frontier method.

3.2 Imputation of Missing Values

This section discusses the general approach of replacing missing values with imputed values. In the context of efficiency analysis, this approach is preferable for two main reasons: first, it enables us to derive efficiency scores for observations with missing values. Second, once imputed values have been obtained, any complete data method – whether econometric, such as SFA, or deterministic, such as Data Envelopment Analysis (DEA, see Charnes et al. (1978)) – may be applied. There are however many alternative methods of imputation, each of which has its own advantages and disadvantages. These are discussed below.

3.2.1 Single Imputation Methods

The simplest imputation methods involve replacing each missing value with only one imputation. Little and Rubin (2002) group single imputation methods into two categories: explicit modelling approaches, based on formal statistical models with explicit assumptions, and implicit modelling approaches based on algorithms, in which the statistical model and its assumptions are only implicit.

Implicit modelling approaches are mainly appropriate for longitudinal survey data of a sort not generally used in efficiency analyses, and therefore we do not go into detail on these approaches. They include *hot deck imputation*, which substitutes values from similar responding units, *substitution* in which non-responding units are replaced by similar units at the fieldwork stage, and *cold deck imputation* in which missing values are replaced by a value from some external source, such as data from a previous survey.

Explicit modelling approaches include *mean imputation* which substitutes means, e.g. sample means, means over time for a particular unit, or means across a class of units or observations, *regression imputation* in which replaces missing values of a given variable by predictions from a regression of that variable on a vector of independent variables, and *stochastic regression imputation* (Herzog and Rubin, 1983) which replaces missing values by regression imputation plus a residual which is a random draw from the estimated error distribution, and reflects the uncertainty in prediction from the regression.

The main disadvantages of single imputation methods, aside from the issue of choosing between imputation methods, are the bias introduced by errors in the independent variables, and the difficulty of deriving estimates standard errors that take into account the added uncertainty as a result of the inclusion of imputed values. Little and Rubin (2002) note three approaches to doing so under single imputation methods: applying explicit variance formula, which could be very difficult under certain imputation methods, modifying imputations such that valid standard errors may be derived, which may

compromise the quality of the imputations, and resampling methods such as bootstrapping and jackknifing, which rely on large datasets. A more general technique, which may be used in a variety of applications, is multiple imputation, discussed below.

3.2.2 Multiple Imputation and its statistical properties

Multiple imputation (MI) – see Rubin (1978a; 1978b; 1986; 1996) – is a method in which we replace each missing value with a vector of $D \geq 2$ imputed values, thereby creating D imputed datasets. MI was developed as a method for dealing with missing data in a wide variety of contexts – not limited to linear regression, but also for estimating means, standard deviations, and skewness of distributions, and for the estimation of nonlinear models. Indeed, many differing MI methods have been proposed appropriate to a range of modelling contexts – linear regression for imputing values of continuous variables – see Schenker and Taylor (1996) – truncated regression for values of variables restricted to a certain range (see Raghunathan et al. (2001), logistic regression for values of binary variables (see Rubin (1987) and Raghunathan et al. (2001)). Raghunathan et al. (2001) and van Buuren (2007) also suggest the use of Poisson regression models to impute values for count variables, ordered logistic regression to impute values of ordinal variables, and multinomial logistic regression to impute values of nominal variables.

MI methods have therefore been used in many different contexts – discrete choice modelling (Steimetz and Brownstone, 2005; Raghunathan and Siscovick, 1996), Poisson regression (Smith and Cummings, 2004; Böhning et al., 2002), and as we discuss later in more detail, stochastic frontier analysis. As such, MI is applicable in many different settings, and its validity, as we discuss below, is determined by the appropriateness of the MI method used to the problem at hand.

As Little and Rubin (2002) note, any single imputation method that involves taking draws from a predictive distribution – e.g. stochastic regression imputation – can be used to impute multiple datasets. Rubin (1987) recommends drawing from a Bayesian posterior predictive distribution, and MI is usually motivated from a Bayesian perspective. Nonetheless resulting inferences can be shown to have good sampling properties (Little and Rubin, 2002). In discussing the validity of MI from a frequentist perspective, Rubin (1987) follows Neyman (1934) in distinguishing two versions of frequentist validity: *randomisation validity*, and *confidence validity*. The former is achieved when, for an interval estimate, the nominal interval coverage is equal to the actual interval coverage, and for hypothesis tests the nominal rejection equals the actual rejection rate, while the latter is achieved when nominal interval coverage and nominal rejection rates are less than or equal to actual interval coverage and actual rejection rates, respectively.

As Rubin (1996) points out, when the sampling distribution of a complete data estimate $\hat{\theta}$ is asymptotically normal, randomisation validity is theoretically achievable. Result 4.1 in Rubin (1987)

shows that, if the complete data inference is randomisation-valid and the MI method is *proper* – as defined by the author – then MI leads to randomisation-valid inference as $D \to \infty$. An analogous result regarding confidence validity is given by Rubin (1996). MI is suitable for application to a wide class of models – see Schenker and Welsh (1988) and Kim (2004) for a discussion of the properties of MI estimators in the linear modelling context.

Regarding proper imputation, Rubin (1987; 1996) argues that:

"If imputations are drawn to approximate repetitions from a Bayesian posterior distribution … [of the missing data] … under the posited response mechanism and an appropriate model for the data, then in large samples the imputation method is proper."

The above discussion suggests that MI has desirable properties for large D if an appropriate imputation method. This raises the questions of appropriate choice of D in practice, and of the performance of MI when the imputation method is inappropriate. The former is addressed by Rubin and Schenker (1986), who show that in many cases, when the fraction of missing data is modest – 30% or less – interval estimates using $D = 2$ have coverages very close to their nominal coverages. Rubin (1987) shows that the large sample relative efficiency, in terms of standard errors of an estimate based on D imputations is approximately $(1 + \gamma/D)^{-1/2}$, where γ is the fraction of missing information defined by (5). That is, the standard error of an estimate based on D imputations is $\sqrt{1 + \gamma/D} - 1$ times larger than it would be if based on infinite imputations (von Hippel, 2005).

Regarding the appropriateness of the MICE and MVN methods in the context of SFA, note that the SF model differs from the basic linear regression model only in the skewness of the error term. As such, unbiased estimates of the frontier parameters may be obtained via least squares, with the exception of the intercept, which yields an unbiased estimate of $\beta_0 + E(u_i)$. We therefore argue that the use of techniques appropriate to linear regression may be applied to impute missing data for SF models without introducing serious biases.

3.2.3 Multiple Imputation Methods

In simple cases, when only one variable contains missing values, univariate imputation models may be used in which we regress the variable with missing values on a vector of covariates, and then simulate new parameters from their joint posterior distribution under an uninformative or improper prior and predict values D times.

In many cases, the presence of missing values is not restricted to a single variable. If the missing data pattern is monotone, then the missing values may be imputed sequentially by independent univariate imputation models, before multiple imputations are taken from a 'final' model estimated using earlier imputations. However, in cases with an arbitrary missing data pattern, an iterative approach is needed.

One approach to MI with a general pattern of missing data is multiple imputation by chained equations (MICE) as proposed by van Buuren et al. (1999), under which regression models are specified for each variable with missing values, and these are estimated in turn using available cases for the dependent variable and imputed values for the independent variables. Imputations are updated at each iteration until convergence is achieved.

An alternative to MICE is to assume that all variables in the imputation model jointly follow a multivariate normal (MVN) distribution. As described by Schafer (1997), imputed values can be obtained from the multivariate normal model using a Markov chain Monte Carlo (MCMC) algorithm. The EM algorithm may be used to obtain starting values of the parameters of the MVN distribution.

The main advantages of MICE over the MVN model are firstly that each regression equation can be specified according to the nature of the dependent variable, e.g. linear regression for a continuous variable, logistic regression for a binary variable, or multinomial logistic regression for an ordinal variable. Second, MICE it does not require the assumption of multivariate normality. On the other hand, MICE lacks the same theoretical justification as MVN MI, since the fitting of a series of conditional distributions may not be consistent with a proper joint distribution, i.e. the regression specifications may be incompatible – see Arnold et al. (1999; 2001) – which may cause issues in convergence. However, simulation evidence from van Buuren et al. (2006) indicates that even with strongly incompatible specifications, there is little impact on estimated parameters. Furthermore, Schafer (1997) and Little and Rubin (2002) suggest that the MVN model can provide reasonable estimates even when multivariate normality does not hold. Of course, in some cases it is possible to transform variables so that the assumption that their marginal distributions are normal is more appropriate, e.g. by taking logs of variables that only take on positive values.

A number of studies have compared the performance of MVN MI to MICE: van Buuren (2007) finds that MICE is preferred when the assumption of multivariate normality is not realistic, and Yu et al. (2007) finds that MICE also performs better than MVN MI when the data are highly skewed due to a large number of zero values. A simulation study by Lee and Carlin (2010) finds that MVN MI slightly outperforms MICE in terms of coverage, even when binary variables are included, and that both methods outperform LD. On the other hand, another simulation study by Kropko et al. (2014) finds that MICE is more accurate than MVN MI when binary variables are included and the data are MAR.

In a panel data setting, an additional consideration is the need to take into account clustering in the data when obtaining imputations. One approach would be to add indicator variables into the chained equations under MICE, or to the MVN model if we use MVN MI, although in the latter case this may be problematic given the assumption of multivariate normality. Another is to impute data separately for each DMU. However if T – the number of time periods – is small, this severely limits the number of variables that may be used in the imputation model. A third alternative, as suggested by Allison (2002),

is to reshape the data so that for each DMU there is only one observation, with values of the same variable at different points in time treated as distinct variables. This allows any given variable at any given point in time to be used as a predictors for any other variable at any other point in time. This method can work with many MI models, but requires that T be small relative to N, the number of DMUs.

Once we have obtained our imputations, we then apply standard complete data methods to each imputed dataset, obtaining D sets of results. The sampling variation of these results is then taken into account when calculating standard errors. Then, if β_{jd} is an estimated coefficient obtained by using dataset d, our estimate of the same coefficient under multiple imputation is

$$\beta_j = \frac{1}{D}\sum_{d=1}^{D}\beta_{jd} \tag{3}$$

The total variability associated with this estimate is

$$V_j = \frac{1}{D}\sum_{d=1}^{D}W_{jd} + \frac{1}{D-1}\sum_{d=1}^{D}(\beta_{jd}-\beta_j) \tag{4}$$

Where W_{jd} is the standard error of β_j obtained by estimating the model with imputed dataset d i.e. the overall standard error associated with β_j is the sum of an average within-imputation variance and a between-imputation component. The fraction of information about the true parameter value missing due to missing data is estimated by:

$$\gamma = \frac{1}{D-1}\frac{\frac{1}{D-1}\sum_{d=1}^{D}(\beta_{jd}-\beta_j)}{\frac{1}{D}\sum_{d=1}^{D}W_{jd} + \frac{1}{D-1}\sum_{d=1}^{D}(\beta_{jd}-\beta_j)} \tag{5}$$

For large sample sizes, Rubin and Schenker (1986) show that the distribution of the *t* statistic follows a Student's *t* distribution with degrees of freedom:

$$v = (D-1)\left[1 + \frac{D-1}{D(D+1)}\frac{\sum_{d=1}^{D}W_{jd}}{\sum_{d=1}^{D}(\beta_{jd}-\beta_j)}\right]^2 \tag{6}$$

while for small samples, Barnard and Rubin (1999) show that the expression for small samples is

$$v^* = \left[\frac{1}{v} + (1-\gamma)\left(\frac{v_{com}+1}{v_{com}+3}\right)\right] \tag{7}$$

where v_{com} is the degrees of freedom when there are no missing values.

3.2.4 The use of auxiliary variables in imputations

An important consideration is the choice of variables to include in the imputation model. Allison (2002) points out that it is important to include all variables in the analysis in the imputation model, so that the relationships between the variables are fully taken into account. However, there is no need to limit ourselves to variables used in the final analysis: in many cases, it may be advantageous to include additional *auxiliary* variables if these are correlated with the missing variables. Auxiliary variables could include, for example, anything that serves as a reasonable proxy for the variable(s) with missing values, for example lagged or lead values of those variables. Allison (2002) shows examples in which reduced standard errors result from the inclusion of appropriate auxiliary variables in the imputation model. The use of auxiliary variables is potentially very attractive in the frontier analysis application, because it is often the case that there exists a priori information from either economics or engineering as to what explains the missing variables.

In our empirical example there are indeed natural auxiliary variables and so we use these in the imputation phase of our MI procedures.

3.2.5 The inclusion of observations with imputed dependent variable in the analysis stage

There has been some discussion in the literature as to whether imputed values of the dependent variable should be included in the analysis stage. On one hand, when independent variables are complete and the dependent variable is MAR, Little (1992) states that the incomplete cases contribute no additional information, implying that LD should be used. In the more general case where there are also missing values among the independent variables, von Hippel (2007) argues that observations in which the dependent variable is missing should be used in the imputation model, since this may improve imputation of missing independent variables in other observations, but that they add nothing but noise to the subsequent analysis. Von Hippel (2007) proposes an approach of *Multiple Imputation, then Deletion* (MID) – under which missing values of the dependent and independent variables are jointly imputed as usual but observations with imputed values for the independent variable are excluded from the analysis step – and provides evidence from simulated data that MID yields more efficient estimates than MI when imputation and analysis models are the same.

However, when imputation and analysis models differ due to the presence of auxiliary variables in the former, standard MI may outperform MID, as acknowledged by von Hippel (2007), who shows that the advantage of MI over MID is greater the stronger the correlation between auxiliary and dependent variables, and the smaller the proportion of missing values. Furthermore, simulations by Sullivan et al. (2015) examines the relative performance of MI and MID in the presence of auxiliary variables under two assumptions: first, that missingness of the dependent variable is unrelated to the auxiliary variable, and second, that the auxiliary variable is associated with missingness of the dependent variable. In the

latter case, Sullivan et al. (2015) find serious biases in the parameter estimates under MID, and in the former case that MI tends to produce more efficient estimates.

We return to this issue in our Monte Carlo simulations in section 4 to make conclusions on a preferred approach for SFA.

3.3 Methods used in SFA efficiency analysis

In the preceding sub-sections we have surveyed missing data methods. We now consider existing development and implementations in efficiency analysis. Little attention has been given to the issue of missing values in the context of SFA, with LD being the approach taken in almost every case despite its disadvantages.

A small number of applications of SFA briefly mention having made use of MI to handle missing values, but contain little or no further details or discussion. Two studies of the technical and cost efficiency of German physician practices, by Heimeshoff et al. (2014) and Kwietniewski and Schreyögg (2018) state that they use MI as a sensitivity check, but do not contain any further discussion of MI and the imputation method(s) used, and do not report the results obtained using MI. Likewise, Bhandari et al. (2015) and Qushim et al. (2016) analyse the technical efficiency of US beef and lamb farms, respectively, and make brief references to having used MI. In the latter case, the authors state that a truncated regression model was used, but no further details are given. Smith and Forster (2013) apply MICE and 'hot deck' MI in a study of the technical efficiency of agent-managed versus owner-managed sugar plantations in early 19th century St. Vincent and the Grenadines.

Several other studies go into greater detail. A series of studies of the technical efficiency of German firms using the IAB establishment panel dataset use MI in a stochastic frontier setting. Jensen et al. (2010) use truncated regression imputation model to impute censored wage data. Kölling and Rässler (2004) and Jensen and Rässler (2006; 2007) each use MVN MI models with uninformative priors, and apply transformations to address non-normality of variables, e.g. log transformations to right-skewed variables, and logit transformations to variables bounded by zero and one. Kölling and Rässler (2004) find that using this MI method leads to considerably different findings regarding the relative efficiency of East and West German firms, compared to LD. Depending on whether MI or LD is used, Jensen and Rässler (2006; 2007) obtain differing findings on effects of exports and collective bargaining – respectively – on efficiency and productivity.

Shaik and Tokovenko (2013), who apply the MVN MI approach to impute missing values of capital and fertiliser inputs in a Food and Agriculture Organization of the United Nations dataset on country-level agricultural production. The authors note that missingness of the data is associated with countries' income levels. Under various different assumptions about the distribution of u_i, the authors find some significant differences in average efficiency predictions across income groups under the MI and LD.

These latter studies highlight the potential usefulness of MI in SFA, given that its use can lead to different conclusions regarding efficiency, for example when the data are MAR rather than MCAR.

4 Monte Carlo Simulations

In this section, we analyse the performance of MI and MID methods for dealing with missing data in a stochastic frontier setting via a set of Monte Carlo simulation exercises. There are already a number of such simulation studies analysing the performance of MI under different data generating processes (DGPs) in a variety of modelling contexts – see for example Bentler (2000), von Hippel (2007), and Rubright et al. (2014). Ours is the first, however, to examine the performance of MI in the context of SFA, and as such we focus discussion on results particularly important in this context, e.g. the estimated scale parameters and their impact on efficiency prediction. We consider four data generating processes (DGPs) with two differing missing data mechanisms (MCAR and MAR), and different proportions of observations with missing data. We compare the results with those obtained via LD. Below, we outline the DGPs used.

4.1 Simulation design

Below we discuss how we generate the missing data and our simulations. The rationale behind what may appear a rather convoluted DGPs is as follows: for the data to be MCAR, we need to ensure that missingness is uncorrelated with any of the variables in the model. This is relatively simple. However, to ensure that the data are MAR, missingness of a given variable may be related to values of other variables, but not to the value of the variable itself. Since data are rarely MCAR (Rubin, 1976), it is important to consider the case in which the data are MAR. In the case of two or more variables with missing data, simulating data to be MAR without being MNAR is relatively complex. Examples of similar simulation studies which both MCAR and MAR missing data mechanisms include those mentioned above. As discussed in section 3.2, the literature indicates that the difference between MI and LD will increase as the missing data process departs from MCAR.

Under both DGPs, the dependent variable, c_i, is given by

$$c_i = x_{1i} + x_{2i} + x_{3i} + z_i + 1 + v_i + u_i, \qquad u_i = |w_i| \tag{8}$$

Where the right-hand side variables are drawn from a multivariate normal distribution, such that

$$\begin{pmatrix} x_{1i} \\ x_{2i} \\ m_{yi} \\ x_{3i} \\ m_{zi} \\ z_i \\ v_i \\ w_i \end{pmatrix} \sim N_8(\mathbf{0}, \boldsymbol{\Sigma}) \tag{9}$$

Where N_8 denotes a eight-dimensional multivariate distribution, $\mathbf{0}$ is a 8×1 vector of zero means, m_{yi} and m_{zi} are variables which determine whether or not y_i and z_i, respectively, are missing – the precise mechanism is explained further below – and the 8×8 covariance matrix $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \boldsymbol{\sigma}\mathbf{P}\boldsymbol{\sigma}' \tag{10}$$

Where $\boldsymbol{\sigma}$ is an 8×1 vector of standard deviations, such that, with the exception that $\sigma_v = 0.5$, each standard deviation is one[2].

$$\boldsymbol{\sigma}' = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0.5 \quad 1) \tag{11}$$

And $\mathbf{P}$ is an 8×8 correlation matrix, given by

$$\mathrm{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0.75 & 0 & 0 \\ 0 & 1 & \rho_{m_y,x_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{m_y,x_2} & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_{m_z,x_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \rho_{m_z,x_3} & 1 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{12}$$

That is, every variable is uncorrelated with every other variable, with the following a few exceptions. The first exception is that

$$\rho_{x_1,z} = 0.75 \tag{13}$$

This is so that one of the right-hand side variables, x_1, is useful for imputation of missing values of z_i. Second, when the data are MAR

$$\rho_{m_y,x_2} = \rho_{m_z,x_3} = 0.75 \tag{14}$$

That is, when the data are MAR, the missingness of y_i is correlated with x_{2i}, and the missingness of z_i is correlated with x_{3i}. On the other hand, we have

$$\rho_{m_y,x_2} = \rho_{m_z,x_3} = 0 \tag{15}$$

for our DGPs in which the data are MCAR, by necessity. Note that the DGP is a stochastic frontier model where we have missing data on costs, c_i, and one of the independent variables, z_i.

We use two different proportions of missing data: one in which 10% of observations have missing values - in either y_i or z_i or both – and one in which 50% of observations have missing values. The observations with missing values are chosen as follows. First, draws from a uniform distribution are

[2] We have set σ_v, the standard deviation of the noise term v_i, to 0.5 so that we have a relatively large signal-to-noise ratio, which reduces incidences of 'wrong skew'. Since convergence of ML estimation of SF models can be slow in cases of wrong skew, this was a practical measure to speed up the Monte Carlo simulations.

taken, and both y_i and z_i are set to missing in observations in which the highest 2% or 10% of values of these draws are found. Then, the remaining 98% or 90% of observations are partitioned into two halves: one in which y_i may be missing, and the other in which z_i may be missing. Within these, values are set to missing in observations with the largest values of m_{yi} or m_{zi}, respectively, are found, until either 10% or 50% of observations, depending on the DGP in question, contain missing values. Both the imputation models and subsequent stochastic frontier models are estimated using Stata 13.1[3].

4.2 Simulation Results: The benefits of MI over LD

Table 1 contains results from replications in which 10% of observations have values MCAR. Table 2 shows results for 50% of observations with values MCAR. Table 3 shows results for 10% of observations with values MAR. Finally, Table 4 shows results for 50% of observations MAR. In each case, the mean and median parameter estimates are shown, along with the bias according to both mean and median estimate. The root mean squared error (RMSE) is given in each case to compare the precision of the various methods. The final three columns provide, for each parameter, comparisons of the magnitudes of the biases relative to those under LD, and the ratio of the RMSE to that obtained via LD.

[3] We used Stata's *MI* command, which enables the use of MI in conjunction with a wide range of commands for estimating specific models, including the *frontier* command, which is used to estimate SF models. The *MI* command imputes *M* datasets according to a use-specified imputation model, estimates the model using each dataset, and combines the estimates as described in Section 3.2.

Table 1: Monte Carlo simulation results, MCAR data, 10% of observations have missing values

		Mean	Median	Bias (Mean)	Bias (Median)	RMSE	Mean bias vs. LD	Median bias vs. LD	RMSE vs. LD
LD	β_1	1.000562	1.001272	0.000562	0.001272	0.128651	-	-	-
	β_2	1.000030	1.000945	0.000030	0.000945	0.084830	-	-	-
	β_3	0.999934	1.000312	-0.000066	0.000312	0.087222	-	-	-
	β_z	0.999222	0.997815	-0.000778	-0.002185	0.132681	-	-	-
	α	1.065605	1.006251	0.065605	0.006251	0.301456	-	-	-
	σ_v	0.456263	0.459472	-0.043738	-0.040528	0.178812	-	-	-
	σ_u	0.910936	0.984032	-0.089064	-0.015968	0.365637	-	-	-
MNV MI	β_1	1.003461	1.003940	0.003461	0.003940	0.125890	0.002899	0.002668	0.978533
	β_2	0.999477	0.999174	-0.000523	-0.000826	0.082111	0.000493	-0.000120	0.967949
	β_3	1.000251	1.001111	0.000251	0.001111	0.084432	0.000185	0.000799	0.968017
	β_z	0.996524	0.996473	-0.003476	-0.003527	0.129761	0.002698	0.001342	0.977993
	α	1.092834	1.040371	0.092834	0.040371	0.290124	0.027229	0.034119	0.962408
	σ_v	0.480124	0.487275	-0.019877	-0.012725	0.166049	-0.023861	-0.027802	0.928624
	σ_u	0.855391	0.944581	-0.144609	-0.055419	0.391661	0.055546	0.039451	1.071173
MVN MID	β_1	1.003834	1.003850	0.003834	0.003850	0.125585	0.003272	0.002578	0.976162
	β_2	0.999870	1.000462	-0.000130	0.000462	0.082423	0.000100	-0.000483	0.971629
	β_3	1.000440	1.000507	0.000440	0.000507	0.084403	0.000374	0.000195	0.967674
	β_z	0.996147	0.994207	-0.003853	-0.005793	0.129361	0.003075	0.003608	0.974982
	α	1.078817	1.022936	0.078817	0.022936	0.289404	0.013212	0.016685	0.960019
	σ_v	0.469089	0.475786	-0.030911	-0.024214	0.170723	-0.012827	-0.016314	0.954766
	σ_u	0.883453	0.963960	-0.116547	-0.036040	0.371967	0.027483	0.020073	1.017311
MICE	β_1	1.003183	1.005143	0.003183	0.005143	0.126275	0.002621	0.003872	0.981530
	β_2	0.999800	1.000896	-0.000201	0.000896	0.082471	0.000171	-0.000050	0.972195
	β_3	1.000570	1.001346	0.000570	0.001346	0.084467	0.000504	0.001034	0.968417
	β_z	0.996626	0.996035	-0.003374	-0.003965	0.130374	0.002596	0.001780	0.982612
	α	1.093898	1.040778	0.093898	0.040778	0.295603	0.028293	0.034526	0.980583
	σ_v	0.478504	0.484501	-0.021496	-0.015499	0.167719	-0.022242	-0.025029	0.937964
	σ_u	0.855109	0.946167	-0.144891	-0.053833	0.395013	0.055827	0.037865	1.080340
MIDCE	β_1	1.002678	1.004988	0.002678	0.004988	0.126313	0.002116	0.003716	0.981825
	β_2	0.999809	0.998883	-0.000191	-0.001117	0.082744	0.000161	0.000171	0.975414
	β_3	1.000592	0.999532	0.000592	-0.000468	0.084099	0.000526	0.000156	0.964189
	β_z	0.996933	0.994972	-0.003067	-0.005028	0.130446	0.002289	0.002843	0.983160
	α	1.080336	1.024864	0.080336	0.024864	0.294972	0.014731	0.018613	0.978491
	σ_v	0.468893	0.474533	-0.031107	-0.025467	0.170863	-0.012631	-0.015061	0.955547
	σ_u	0.879265	0.962012	-0.120735	-0.037988	0.381173	0.031672	0.022021	1.042490

Table 2: Monte Carlo simulation results, MCAR data, 50% of observations have missing values

		Mean	Median	Bias (Mean)	Bias (Median)	RMSE	Mean bias vs. LD	Median bias vs. LD	RMSE vs. LD
LD	β_1	1.003477	1.002050	0.003477	0.002050	0.180337	-	-	-
	β_2	1.001135	1.000545	0.001135	0.000545	0.120006	-	-	-
	β_3	1.002043	1.001060	0.002043	0.001060	0.121010	-	-	-
	β_z	0.996480	0.992628	-0.003520	-0.007372	0.188728	-	-	-
	α	1.097890	1.008190	0.097890	0.008190	0.388256	-	-	-
	σ_v	0.393723	0.421581	-0.106277	-0.078419	0.268511	-	-	-
	σ_u	0.871220	0.986872	-0.128780	-0.013128	0.473280	-	-	-
MNV MI	β_1	1.016073	1.016044	0.016073	0.016044	0.162654	0.012596	0.013994	0.901946
	β_2	1.002449	1.003444	0.002449	0.003444	0.104665	0.001314	0.002899	0.872169
	β_3	1.000594	1.000349	0.000594	0.000349	0.105482	-0.001449	-0.000711	0.871679
	β_z	0.978286	0.979661	-0.021714	-0.020339	0.170230	0.018194	0.012967	0.901989
	α	1.253462	1.217277	0.253462	0.217277	0.380296	0.155572	0.209088	0.979500
	σ_v	0.564749	0.570381	0.064749	0.070381	0.162085	-0.041528	-0.008038	0.603642
	σ_u	0.553174	0.668827	-0.446826	-0.331173	0.618859	0.318046	0.318044	1.307595
MVN MID	β_1	1.017466	1.017916	0.017466	0.017916	0.158501	0.013989	0.015866	0.878916
	β_2	1.001830	1.003742	0.001830	0.003742	0.103371	0.000695	0.003197	0.861387
	β_3	1.001820	1.000384	0.001820	0.000384	0.101822	-0.000223	-0.000676	0.841436
	β_z	0.976563	0.976278	-0.023437	-0.023722	0.166966	0.019918	0.016351	0.884693
	α	1.185671	1.122651	0.185671	0.122651	0.361007	0.087781	0.114461	0.929819
	σ_v	0.492229	0.518643	-0.007771	0.018643	0.199622	-0.098506	-0.059776	0.743440
	σ_u	0.675193	0.813673	-0.324807	-0.186327	0.551836	0.196027	0.173199	1.165983
MICE	β_1	1.011320	1.011225	0.011320	0.011225	0.161028	0.007843	0.009175	0.892927
	β_2	1.001229	1.002799	0.001229	0.002799	0.103888	0.000094	0.002254	0.865697
	β_3	1.002692	1.002755	0.002692	0.002755	0.105397	0.000649	0.001695	0.870974
	β_z	0.984719	0.980740	-0.015281	-0.019260	0.166128	0.011761	0.011888	0.880254
	α	1.249447	1.217927	0.249447	0.217927	0.377186	0.151557	0.209737	0.971489
	σ_v	0.556961	0.569283	0.056961	0.069283	0.164928	-0.049316	-0.009136	0.614233
	σ_u	0.563845	0.693745	-0.436155	-0.306255	0.609643	0.307375	0.293127	1.288125
MIDCE	β_1	1.011760	1.013097	0.011760	0.013097	0.158757	0.008283	0.011047	0.880335
	β_2	1.001692	1.002819	0.001692	0.002819	0.102093	0.000557	0.002274	0.850734
	β_3	1.003104	1.001404	0.003104	0.001404	0.102408	0.001061	0.000344	0.846276
	β_z	0.983612	0.980352	-0.016388	-0.019648	0.163683	0.012868	0.012277	0.867300
	α	1.183009	1.124238	0.183009	0.124238	0.359344	0.085119	0.116048	0.925533
	σ_v	0.486909	0.516395	-0.013091	0.016395	0.203206	-0.093186	-0.062024	0.756789
	σ_u	0.677240	0.813916	-0.322760	-0.186084	0.551272	0.193980	0.172956	1.164791

Table 3: Monte Carlo simulation results, MAR data, 10% of observations have missing values

		Mean	Median	Bias (Mean)	Bias (Median)	RMSE	Mean bias vs. LD	Median bias vs. LD	RMSE vs. LD
LD	β_1	0.994872	0.994412	-0.005128	-0.005588	0.131084	-	-	-
	β_2	0.996871	0.996737	-0.003130	-0.003263	0.087556	-	-	-
	β_3	1.000070	1.000528	0.000070	0.000528	0.088451	-	-	-
	β_z	1.006048	1.006423	0.006048	0.006423	0.132172	-	-	-
	α	1.052813	1.004158	0.052813	0.004158	0.290016	-	-	-
	σ_v	0.450624	0.456004	-0.049377	-0.043996	0.179419	-	-	-
	σ_u	0.930306	0.999667	-0.069694	-0.000333	0.351685	-	-	-
MNV MI	β_1	0.999718	0.999391	-0.000282	-0.000609	0.126146	-0.004846	-0.004979	0.962330
	β_2	0.996416	0.997594	-0.003584	-0.002406	0.085852	0.000454	-0.000857	0.980539
	β_3	1.000932	1.001391	0.000932	0.001391	0.085512	0.000862	0.000863	0.966773
	β_z	1.001963	1.006254	0.001963	0.006253	0.126941	-0.004085	-0.000170	0.960426
	α	1.088494	1.037311	0.088494	0.037311	0.285921	0.035681	0.033153	0.985882
	σ_v	0.476341	0.482943	-0.023659	-0.017057	0.165852	-0.025718	-0.026938	0.924383
	σ_u	0.866231	0.953437	-0.133770	-0.046563	0.386134	0.064075	0.046231	1.097952
MVN MID	β_1	0.999515	1.001451	-0.000486	0.001451	0.126920	-0.004643	-0.004137	0.968239
	β_2	0.996563	0.996388	-0.003437	-0.003612	0.085437	0.000308	0.000350	0.975802
	β_3	1.000967	1.000919	0.000967	0.000919	0.085259	0.000897	0.000391	0.963911
	β_z	1.001742	1.003584	0.001742	0.003584	0.127508	-0.004306	-0.002839	0.964715
	α	1.070557	1.023041	0.070557	0.023041	0.281799	0.017744	0.018883	0.971668
	σ_v	0.463449	0.469935	-0.036551	-0.030065	0.172508	-0.012825	-0.013931	0.961480
	σ_u	0.897357	0.971547	-0.102643	-0.028453	0.364727	0.032949	0.028120	1.037082
MICE	β_1	0.998660	0.999441	-0.001340	-0.000559	0.128650	-0.003788	-0.005029	0.981436
	β_2	0.997556	0.998789	-0.002444	-0.001211	0.085924	-0.000685	-0.002052	0.981363
	β_3	1.000949	0.999847	0.000949	-0.000153	0.084919	0.000879	-0.000375	0.960064
	β_z	1.002296	1.004450	0.002296	0.004450	0.128377	-0.003752	-0.001974	0.971291
	α	1.086656	1.030080	0.086656	0.030080	0.285446	0.033843	0.025922	0.984241
	σ_v	0.476632	0.481388	-0.023368	-0.018612	0.163798	-0.026008	-0.025383	0.912935
	σ_u	0.870600	0.954206	-0.129400	-0.045794	0.379456	0.059706	0.045461	1.078963
MIDCE	β_1	0.998531	0.999042	-0.001469	-0.000958	0.127969	-0.003659	-0.004630	0.976241
	β_2	0.996871	0.997598	-0.003129	-0.002402	0.084816	-0.000001	-0.000861	0.968710
	β_3	1.000850	1.001191	0.000850	0.001191	0.084818	0.000780	0.000663	0.958924
	β_z	1.002250	1.004563	0.002250	0.004563	0.127781	-0.003798	-0.001860	0.966775
	α	1.071989	1.019107	0.071989	0.019107	0.284716	0.019176	0.014950	0.981727
	σ_v	0.464059	0.467489	-0.035941	-0.032511	0.171403	-0.013436	-0.011485	0.955324
	σ_u	0.896319	0.976920	-0.103681	-0.023080	0.366844	0.033987	0.022747	1.043104

Table 4: Monte Carlo simulation results, MAR data, 50% of observations have missing values

		Mean	Median	Bias (Mean)	Bias (Median)	RMSE	Mean bias vs. LD	Median bias vs. LD	RMSE vs. LD
LD	β_1	0.987244	0.987526	-0.012756	-0.012474	0.187677	-	-	-
	β_2	0.997444	0.995768	-0.002556	-0.004232	0.130236	-	-	-
	β_3	0.998000	0.999247	-0.002000	-0.000753	0.129027	-	-	-
	β_z	1.010020	1.010474	0.010020	0.010474	0.187636	-	-	-
	α	1.086678	1.002620	0.086678	0.002620	0.378165	-	-	-
	σ_v	0.387920	0.409128	-0.112080	-0.090872	0.272122	-	-	-
	σ_u	0.891933	1.011283	-0.108067	0.011283	0.458337	-	-	-
MNV MI	β_1	1.007226	1.007329	0.007226	0.007329	0.164329	-0.005530	-0.005145	0.875594
	β_2	0.997105	0.996530	-0.002895	-0.003470	0.109292	0.000339	-0.000762	0.839186
	β_3	1.000554	1.000183	0.000554	0.000183	0.109596	-0.001446	-0.000571	0.849402
	β_z	0.987033	0.990970	-0.012967	-0.009030	0.168795	0.002947	-0.001445	0.899585
	α	1.239056	1.198541	0.239056	0.198541	0.369848	0.152378	0.195921	0.978008
	σ_v	0.561379	0.567512	0.061379	0.067512	0.158377	-0.050702	-0.023360	0.582005
	σ_u	0.570387	0.702199	-0.429613	-0.297801	0.605758	0.321547	0.286518	1.321644
MVN MID	β_1	1.008478	1.011768	0.008478	0.011768	0.160772	-0.004278	-0.000706	0.856644
	β_2	0.994668	0.996699	-0.005332	-0.003301	0.105512	0.002775	-0.000931	0.810165
	β_3	1.001150	1.000284	0.001150	0.000284	0.106825	-0.000850	-0.000470	0.827925
	β_z	0.985261	0.990912	-0.014739	-0.009088	0.165652	0.004719	-0.001386	0.882834
	α	1.170764	1.115120	0.170764	0.115120	0.353253	0.084086	0.112500	0.934125
	σ_v	0.484022	0.512783	-0.015978	0.012783	0.205965	-0.096102	-0.078089	0.756883
	σ_u	0.700531	0.840493	-0.299469	-0.159507	0.537015	0.191402	0.148225	1.171661
MICE	β_1	1.002561	1.000952	0.002561	0.000952	0.165101	-0.010195	-0.011522	0.879710
	β_2	0.995902	0.995176	-0.004098	-0.004824	0.108454	0.001542	0.000592	0.832753
	β_3	1.000637	1.000278	0.000637	0.000278	0.109381	-0.001363	-0.000475	0.847737
	β_z	0.992914	0.998356	-0.007086	-0.001644	0.168079	-0.002934	-0.008830	0.895773
	α	1.234964	1.196989	0.234964	0.196989	0.364267	0.148286	0.194369	0.963251
	σ_v	0.556353	0.565833	0.056353	0.065833	0.161681	-0.055728	-0.025039	0.594150
	σ_u	0.588384	0.725401	-0.411616	-0.274599	0.588253	0.303549	0.263317	1.283452
MIDCE	β_1	1.004409	1.003522	0.004409	0.003522	0.161451	-0.008347	-0.008952	0.860259
	β_2	0.995121	0.996682	-0.004879	-0.003318	0.105064	0.002323	-0.000914	0.806719
	β_3	1.001555	1.000838	0.001555	0.000838	0.105671	-0.000445	0.000085	0.818988
	β_z	0.991485	0.993785	-0.008515	-0.006215	0.163361	-0.001505	-0.004260	0.870626
	α	1.169637	1.114548	0.169637	0.114548	0.351688	0.082959	0.111928	0.929986
	σ_v	0.482901	0.509326	-0.017099	0.009326	0.205173	-0.094982	-0.081546	0.753972
	σ_u	0.706234	0.854157	-0.293766	-0.145843	0.530393	0.185700	0.134560	1.157212

When the data are MCAR as in Table 1 and Table 2, we see as expected that there seems to be no clear pattern in terms of biases, comparing MI and MID methods to LD. However, there are improvements in precision when using MI or MID in terms of reducing RMSEs – greater when the proportion of observations with missing values increases. In both cases, σ_u and α are exceptions: estimates of these parameters appear to be both more biased and less precise under MI and MID than under LD. In the former case, the bias is negative, while in the latter case, the bias is positive. A possible explanation for this result is that, since the MIV and MICE imputation models do not explicitly account for the skewness of the error term, the resulting imputed data have the effect of reducing the skewness of the errors in the data overall. This would explain the fact that the biases in the estimation of σ_u are negative, while the biases in the estimation of α are positive.

Looking at Table 3 and Table 4, we again see that when the data are MAR, the case for MI over LD becomes stronger. In these cases, the reductions in RMSEs relative to LD are greater. This tendency is clearer when the proportion of observations with missing values is larger, i.e. when comparing Table 2 and Table 4. In terms of biases, estimates of β_1, β_2, β_z, and σ_u tend to be less biased, while again biases in the estimates of α and σ_u appear greater under MI and MID methods than under LD.

Overall these simulation results indicate that MI methods, particularly MID methods (see section 4.3 below) exhibit greater precision in estimation of the parameters than LD. This increases as the missing data mechanism moves away from MCAR to MAR and when the proportion of missing values increases. One issue identified is that there is a bias in the estimation of the variance of the inefficiency error term, however broadly there is a clear benefit to using MI over LD in SFA.

4.3 Simulation Results: Choice between MI and MI then Deletion

A noticeable result in every case is that MID methods yield better results than simple MI. Biases with respect to a, σ_v, and σ_u tend to be significantly smaller in magnitude when using MID rather than MI, and there is also nearly a universal improvement in RMSEs, suggesting greater precision. This implies that MID should be preferred over MI in the context of SFA; that is, imputed values of the dependent variable should not be used in estimation of the SFA model. This seems to support arguments by Little (1992) and von Hippel (2007), and complement Monte Carlo simulation evidence from the latter – in a linear regression setting – that cases with missing values of the dependent variable do not contain any information about the model parameters, and that their inclusion in the estimation of the model adds nothing but noise. As a result, as shown by von Hippel (2007), MID results in more precise parameters, smaller standard errors, and improved coverage of confidence intervals.

The advantages of MID over MI seem particularly important in the context of SFA. Comparing MI and MID results when the proportion of observations with missing values is large, we see that the bias in σ_v is not only smaller – in terms of magnitude – under MID than MI, but that the bias under MI is always positive, which demonstrates the way in which use of imputed values of the dependent variable

in the estimation of the model adds noise. Under LD, on the other hand, we see significant negative biases in σ_v in these cases. The estimate of the noise variance, while not of direct interest in the context of linear regression, is of great importance in SFA since it directly influences the deconvolution of estimated residuals into noise and inefficiency components – see for example Wang and Schmidt (2009).

When we also consider the clear advantage of MID over MI in terms of the estimation of σ_u, particularly evident when the fraction of missing observations is large, MID emerges as clearly preferable to MI in terms of efficiency prediction. The smaller downward biases in the estimation of σ_u under MID compared to MI may also be explained with reference to von Hippel (2007), who shows that the advantage of MID over MI is particularly acute when imputations of the dependent variable are problematic. As discussed previously, imputation of the dependent variable under a model that does not account for the skewness of the overall error would appear to be contributing disproportionately to the observed downward bias in the estimation of σ_u under MI, such that MID mitigates this effect.

Thus we recommend Multiple Imputation and then Deletion (MID) methods for use in the SFA context.

5 Empirical Application to Highway Maintenance Costs in England

The previous section provided an introduction to SFA and its motivations as well as reviewing approaches to accounting for missing data. We now apply these methods to a dataset analysed using SFA on English highways authorities' road maintenance cost, using MI to deal with the presence of missing data. The purpose of this application is to illustrate the power of MI to address missing data and to evaluate the positive and negative aspects of various imputation approaches.

We describe the model and data used, before subsequent sections discuss the models used for MI, and compare parameter estimates and efficiency scores derived using several different approaches.

Our dataset consists of data on 87 English highways authorities that were members of the CQC network in 2016-17, covering the financial years 2009-10 to 2015-16. Data on total expenditure on highway maintenance was collected from each member individually according to definitions agreed by a working group of CQC members. Data on the lengths, conditions, and traffic densities of local authority maintained roads were constructed from publically available Department for Transport (DfT) sources, and data on the land area of each authority in km^2 was collected from the Standard Area Measurements (SAM) dataset published by the Office for National Statistics (ONS). We also collected data on regional median gross hourly wages in the civil engineering industry from the Annual Survey on Hours and Earnings (ASHE), also published by the ONS, and a national index of materials prices in road construction that were published by the former Department for Business, Innovation and Skills (BIS). Variable names, definitions and summary statistics are given in Table 5 below.

We also include year indicator variables to control for neutral technical change. Each independent variable is mean-centred so that the first-order coefficients on the scale and density variables may be interpreted as elasticities at the sample means.

Table 5: Description of variables and summary statistics (2009-10 to 2015-16)

Variable	Description	Mean*	St. Dev.*
$TOTEX$	Total expenditure on road highway maintenance (£)**	11,639,665	11,378,166
$LAND$	Local authority area size (km^2)	123,957.930	168,051.340
RL	Total length of local authority roads (road km)	2,520.708	2,440.156
$TRAFFIC$	Traffic density (vehicle km/ road km)	1121.228	336.012
RDC	Proportion of roads in need of maintenance (%)	12.374	5.436
$DRDC$	Change in RDC from previous year (%)	0.178	2.895
$WAGE$	Regional gross wage in civil engineering (£/hr)	12.690	1.021
$ROCOSM$	Road construction materials cost index (1995=100)	262.450	23.099

* Excluding missing values.
**Nominal prices.

We estimate a stochastic cost frontier. Our cost variable is $\ln TOTEX_{it}$, and our output variables are $\ln LAND_{it}$ and $\ln RL_{it}$, which capture the size of the authority's network. Our input price variables are $\ln WAGE_{it}$ and $\ln ROCOSM_{it}$, labour and materials prices, respectively. No suitable capital price was available. We impose linear homogeneity of degree one in input prices by normalising our total expenditure and wage variables by the materials price index. We include $\ln TRAFFIC_{it}$, a measure of traffic density, since increased traffic density leads to increased wear-and-tear damage to the network, increasing maintenance costs. Finally, we include RDC_{it}, an inverse measure of network condition, to capture the relationship between network quality and maintenance costs, and $DRDC_{it}$ (the change in the latter from the previous year) to capture the cost impact of improvements (or deterioration) in network condition. Finally, a set of year dummies (excluding the final year) are included to capture technical change over time.

The purpose of this paper is not to discuss in detail how the final model was established; rather it is focused on the issue of accounting for missing data. A discussion about the background to the empirical model (and the interpretation of equation (16) below), can be found in Wheat (2017), Stead et al. (2018), and Wheat et al. (2019) who have analysed this, or earlier versions of this dataset (drawing on complete cases only). The precise specification of the estimated cost frontier is:

$$\begin{aligned} \ln(TOTEX_{it}/\ln ROCOSM_{it}) &= \beta_0 + \beta_1 \ln LAND_{it} + \beta_2 \ln RL_{it} \\ &+\beta_3(\ln LAND_{it})^2 + \beta_4(\ln RL_{it})^2 + \beta_5 \ln LAND_{it} \ln RL_{it} + \beta_6 \ln TRAFFIC_{it} \\ &+\beta_7 RDC_{it} + \beta_8 DRDC_{it} + \beta_9 \ln(WAGE_{it}/\ln ROCOSM_{it}) + \beta_{10} YEAR1_{it} \end{aligned} \quad (16)$$

$$+\beta_{11}YEAR2_{it} + \beta_{12}YEAR3_{it} + \beta_{13}YEAR4_{it} + \beta_{14}YEAR5_{it} + \beta_{15}YEAR6_{it} + \beta_{16}YEAR7_{it} + v_{it} + u_{it}$$

The i and t subscripts denote authority i in year t. Note the change from (1), since we are using panel data. The error terms v_{it} and u_{it} are respectively assumed to follow a normal distribution with standard deviation σ_v and an exponential distribution with scale parameter σ_u. That is, we use a pooled cross-sectional normal-exponential stochastic frontier specification Aigner et al. (1977). Missing values are found in the dependent variable, due to non-reporting of expenditure, and also in the RDC and $DRDC$ variables. The missing data pattern is non-monotone, therefore necessitating methods compatible with a general pattern of missing data. The precise pattern of missingness is discussed, along with the imputation models, in the next section.

5.1 Imputation Models

Given one of our contributions to the literature is to compare different MI approaches for a stochastic frontier application, we utilise two different approaches – discussed in section 3.2.2 – to MI. Each approach is appropriate for a general pattern of missing data such as ours, namely, MVN MI and MICE. We have chosen these approaches based upon their wide usage in the MI literature – including their use in the limited set of existing SFA applications, as discussed in Section 3.3 – and their appropriateness to imputing values of continuous variables.

As discussed in the previous section, we have three variables containing missing values: $TOTEX$, RDC, and $DRDC$. Given that $TOTEX$ may only take on positive values, this enters the imputation models – as it enters the final analyses – in log form, as $\ln TOTEX$, so that the distributional assumptions of the imputation models are more appropriate, and so that the models do not produce negative imputations.

In the case of the RDC variable, the log transformation is not sufficient to ensure that imputations fall within the required range, between 0 and 100. Another consideration is that RDC is in fact a weighted index of three different variables:

$$RDC_{it} = w_{A_{it}}RDCA_{it} + w_{BC_{it}}RDCBC_{it} + w_{U_{it}}RDCU_{it} \quad (17)$$

where $RDCA$, $RDCBC$, and $RDCU$ are condition indices and $w_{A_{it}}$, $w_{BC_{it}}$, and $w_{U_{it}}$ weightings for an authority's A, B and C, and U roads, respectively. A Roads are the highest standard if roads, followed by B and then C roads. U roads are the classification of all other roads that are of lower standard than A, B and C. The weightings – shares in total network length – are known, and therefore the missingness of RDC is explained by the missingness of one or more of $RDCA$, $RDCBC$, or $RDCU$. Given that we expect conditions across different road classifications to be correlated, and that we do not want to

discard known values of two (one) components of RDC when one (two) are missing, there is therefore a gain from imputing missing values of the three components of RDC separately.

In order to ensure that imputed values of $RDCA$, $RDCBC$, and $RDCU$ fall between 0 and 100, and that they can be modelled as normally distributed random variables in the imputation models, we divide the variables by 100 and transform them via the quantile function for the standard normal distribution, i.e. we define the variables

$$\overline{RDCA}_{it} = \Phi^{-1}\left(\frac{RDCA_{it}}{100}\right), \overline{RDCBC}_{it} = \Phi^{-1}\left(\frac{RDCBC_{it}}{100}\right), \overline{RDCU}_{it} = \Phi^{-1}\left(\frac{RDCU_{it}}{100}\right) \quad (18)$$

to be used in the imputation models, where Φ^{-1} is the standard normal quantile function. When imputations have been obtained, these are then be retransformed via the standard normal cumulative density function in order to recover the original values and imputations between 0 and 100. The final RDC variable is therefore

$$RDC_{it} = 100\left[w_{A_{it}}\Phi(\overline{RDCA}_{it}) + w_{BC_{it}}\Phi(\overline{RDCBC}_{it}) + w_{U_{it}}\Phi(\overline{RDCU}_{it})\right] \quad (19)$$

where Φ is the standard normal cumulative density function. Where one or more of the components of RDC_{it} are imputed, there will be D different versions of RDC_{it}, which we consider as imputations.

Rather than separately imputing values of $DRDC_{it}$, which is by definition $RDC_{it} - RDC_{it-1}$, we then impute it, where necessary, using the imputed values of RDC_{it} and RDC_{it-1} to ensure consistency between the two variables. Of course, in order to calculate values for 2010, we therefore need RDC values from 2009, which may have to be imputed. We deal with this by including RDC values back to 2008 – the earliest data available on road conditions – in auxiliary variables in the imputation models.

Overall, our approach to imputation of the RDC and DRDC variables described above has the following properties:

- Imputations bounded by the appropriate range: Our imputations for the components of RDC (RDCA, RDCBC and RDCU) are bounded between 0-100 i.e. the limits of applicable bounds of the data.
- Complete data cases unaffected: Our transforms recover exactly the values of RDC and DRDC for those observations with no missing data.
- Relationships between explanatory variables maintained: The relationship between RDC and DRDC is maintained i.e. DRDC is the change in RDC from one period to another rather than an independent variable.

We consider that these three properties are important for any imputation strategy. In particular this transformation approach is a convenient way of modifying the distributional assumptions in the imputation model, such that the assumed marginal distribution of a particular variable is more appropriate, and the resulting imputations are improved. This is a common approach in the literature, see e.g. Raghunathan et al. (2001), Allison (2002), Schafer and Graham (2002). This means that the use of imputation methods such as MVN MI, that have been criticised for the possibly inappropriate distributional assumptions made, need not be inconsistent with the distribution of the missing variables.

In the case of the MVN MI model, we account for the panel structure of the data by the data reshaping approach suggested by Allison (2002) – and explained previously in Section 3.2.3 – i.e. pooling all the observations on a given DMU together into a single observation. Given that we have seven years' worth of data, this results in seven times fewer observations – 87 from 588 – and seven versions of each variable, each corresponding to a particular year, so that $RDCA_{it}$ becomes $RDCA2010_i, RDCA2011_i, \dots, RDCA2016_i$ and so on. Each of the year-specific versions of our variables with missing values enters the model as a dependent variable, with four exceptions: $RDCA2014_i$, $RDCA2015_i$, $RDCBC2014_i$ and $RDCBC2015_i$, which have no missing values; these variables are therefore included as independent variables in the imputation model.

In the case of the MICE model, we specify equations for $\ln TOTEX_{it}$, $\overline{RDCA}_{it}$, $\overline{RDCBC}_{it}$, $\overline{RDCU}_{it}$, $\overline{RDCA}_{it-1}$, $\overline{RDCBC}_{it-1}$, $\overline{RDCU}_{it-1}$, $\overline{RDCA}_{it-2}$, $\overline{RDCBC}_{it-2}$, and $\overline{RDCU}_{it-2}$. In the equations relating to each of these variables, all the others are included as independent variables, along with a set of fixed effects and all remaining variables from the final analysis model, with the exceptions of RDC_{it} and $DRDC_{it}$. The lagged transformed condition variables are included as auxiliary variables, and for imputing $DRDC_{it}$. This approach makes intuitive sense, since we know from the complete cases that past road condition is a good predictor of current road condition, and that within authorities, there is a strong relationship between road condition statistics across classifications.

The specifications for both MI models are shown in Table 6 and Table 7, respectively.

Table 6: MVN MI model specification

Dependent variables	Independent variables
$\ln TOTEX2010_i, \dots, \ln TOTEX2016_i$ $\overline{RDCA}2008_i, \dots, \overline{RDCA}2013_i,$ $\overline{RDCA}2016_i, \overline{RDCBC}2016_i$ $\overline{RDCBC}2008_i, \dots, \overline{RDCBC}2013_i$ $\overline{RDCU}2008_i, \dots, \overline{RDCU}2016_i$	$\ln LAND2010_i, \ln RL2010_i, \ln RL2016_i, (\ln LAND2010_i)^2,$ $\ln LAND2010 \ln RL2010_i, \ln LAND2016 \ln RL2016_i$ $(\ln RL2010_i)^2, \ln TRAFFIC2010_i - \ln TRAFFIC2016_i$ $\overline{RDCA}2014_i, \overline{RDCA}2015_i, \overline{RDCBC}2014_i, \overline{RDCBC}2015_i$ $\ln WAGE2010_i - \ln WAGE2016_i, PROP\_UA_i, PROP\_UB_i,$ $PROP\_UC_i, PROP\_UU_i, PROP\_RA_i, PROP\_RB_i, PROP\_RC_i$

Table 7: MICE model specification

Equation	Dependent variable	Independent variables
1	$\ln TOTEX_{it}$	$\ln TOTEX_{it}^{*(1)}$, $\ln LAND_{it}$, $\ln RL_{it}$, $(\ln LAND_{it})^2$, $(\ln RL_{it})^2$, $\ln LAND_{it} \ln RL_{it}$, $\ln TRAFFIC_{it}$, $\ln WAGE_{it}$, $\overline{RDCA}_{it}^{*(2)}$, $\overline{RDCBC}_{it}^{*(3)}$, $\overline{RDCU}_{it}^{*(4)}$, $\overline{RDCA}_{it-1}^{*(5)}$, $\overline{RDCBC}_{it-1}^{*(6)}$, $\overline{RDCU}_{it-1}^{*(7)}$, $\overline{RDCA}_{it-2}^{*(8)}$, $\overline{RDCBC}_{it-2}^{*(9)}$, $\overline{RDCU}_{it-2}^{*(10)}$, Authority indicator variables
2	$\overline{RDCA}_{it}$	
3	$\overline{RDCBC}_{it}$	
4	$\overline{RDCU}_{it}$	
5	$\overline{RDCA}_{it-1}$	
6	$\overline{RDCBC}_{it-1}$	
7	$\overline{RDCU}_{it-1}$	
8	$\overline{RDCA}_{it-2}$	
9	$\overline{RDCBC}_{it-2}$	
10	$\overline{RDCU}_{it-2}$	

*(n): Except equation n.

Given that we have data from 84 authorities over 7 years, we have a total of 588 observations. The numbers of missing values for each of the incomplete variables is shown in Table 8 below. The highest number of missing values in a particular variable is 98, or around 16.7% of observations[4], while the variable with the least number of missing observations is $\ln TOTEX_{it}$ with 82 missing values, or around 13.9% of observations. We chose $D = 10$ as the number of imputed datasets and a ridge prior distribution for the MVN MI model. Following discussed in Section 3.2.2, this is a relatively large number of imputations, and below Table 10, we discuss the relative efficiency of our parameter estimates, in light of the choice of $D = 10$.

Auxiliary variables are included in the imputation models in the form of disaggregated road condition indices (by different road types) which are highly correlated with one another and hence useful in predicting in cases where one is missing. Lagged and lead values referring to past and future road condition values – including values pre-dating the start of the sample period – are also included on the basis that, for a given year, they are useful for predicting present road condition. Preliminary regressions using LD – not reported here – confirmed that this is the case, and furthermore that the condition of other road classifications, and in some cases lagged values of these, were also significant predictors. The presence of these auxiliary variables therefore significantly improves the quality of the imputation models with respect to imputing missing values of our condition variables. In the MVN MI model, proportion variables describing the composition of an authority's road network are also included, since

[4] This is the number of missing values for $DRDC_{it}$, the change in road condition, which is ultimately driven by missingness of the components of the condition index. Note that the number of missing values in $DRDC_{it}$ cannot be deduced by the numbers of missing values of the components of the condition index (as shown in Table 8) because the pattern of missingness is non-monotone, and very general.

these were likewise found significant in predicting road condition variables; these proportion variables, being time-invariant, are excluded from the MICE model due to the inclusion of indicator variables.

Table 8: Complete and imputed values

Variable	Complete values	Imputed values	Total values
$\ln TOTEX$	506	82	588
$\overline{RDCA}_{it}$	576	12	588
$\overline{RDCBC}_{it}$	574	14	588
$\overline{RDCU_{it}}$	541	47	588
$\overline{RDCA}_{it-1}$	570	18	588
$\overline{RDCBC}_{it-1}$	571	17	588
$\overline{RDCU}_{it-1}$	536	52	588
$\overline{RDCA}_{it-2}$	568	20	588
$\overline{RDCBC}_{it-2}$	566	22	588
$\overline{RDCU}_{it-2}$	534	54	588

5.2 Imputation model results

Results from the imputation models consist of parameter estimates and, in the case of the MVN MI model, covariances between the dependent variables. We do not report the MVN MI model results, since given the reshaping of each variable into seven year-specific variables, there are far too many parameters and covariances to summarise in any meaningful way. Also the latter are difficult to interpret except in that a positive (negative) sign denotes a positive (negative) correlation. The MICE model, however, results in a far more manageable number of parameters.

The estimated parameters of the four key equations of the MICE imputation model are presented in Table 9. The equations relating to the lagged values of the road condition indicators are omitted, as are parameters relating to the authority indicator variables, of which there are many. However, the presence of the authority indicator variables should be taken into account when interpreting the estimated parameters, particularly those relating to variable such as land area, road length, and traffic: these variables show little to no variation over time, and their impacts are likely being picked up primarily by the authority indicator variables, several of which are found to be statistically significant.

Table 9 shows that, as expected, lagged values of our condition variables seem to be statistically significant in predicting current values. This is intuitively sensible, and coincides with our preliminary investigation discussed previously, which indicated that lagged values of road condition are useful in predicting present values. First order lags of $\overline{RDCA}$, $\overline{RDCBC}$, and $\overline{RDCU}$ all have positive estimated coefficients, significant at the 1% level in their respective equations. In each case, these seem to be the most important predictors in terms of magnitude and statistical significance, and hence the most useful

in prediction. Second order lags are less significant, and some have negative signs, while the conditions of other road classifications – and in some cases their lags – are also significant in some cases; in particular, $\overline{RDCA}$ seems to have value in predicting $\overline{RDCBC}$ and vice versa. In addition, land area, road length, and traffic density all seem to be significant predictors of $RDCBC_{it}$. $RDCU_{it}$ appears to be significant in predicting $lnTOTEX_{it}$, and vice versa. Much of this information would be lost under a simpler procedure, such as linear interpolation.

The transparency in this respect afforded by the MICE method is important from a stakeholder engagement perspective, as it gives confidence to stakeholders that missing data is imputed in a sensible manner.

It should be noted that, given the nonlinear transformations applied to the road condition variables, their associated coefficients cannot be interpreted as marginal effects. The marginal effect of, e.g. $RDCA_{it-1}$ on $RDCA_{it}$ is given by

$$\frac{\partial RDCA_{it}}{\partial RDCA_{it-1}} = \beta_{2,2} \frac{\phi(\overline{RDCA}_{it})}{\phi(\overline{RDCA}_{it-1})} \tag{20}$$

In the absence of a substantial change in $RDCA$ from the previous year, $\beta_{2,2}$ is approximately equal to the marginal effect of $RDCA_{it-1}$ on $RDCA_{it}$, since $\phi(\overline{RDCA}_{it}) \approx \phi(\overline{RDCA}_{it-1})$. On the other hand, the marginal effect of $RDCBC_{it}$ on $RDCA_{it}$ is given by

$$\frac{\partial RDCA_{it}}{\partial RDCBC_{it}} = \beta_{2,5} \frac{\phi(\overline{RDCA}_{it})}{\phi(\overline{RDBC}_{it})} \tag{21}$$

and, given that $RDCBC_{it}$ tends to be larger than $RDCA_{it}$, i.e. A roads tend to be in better condition than B and C roads, the marginal effect will tend to be less than $\beta_{2,5}$. It follows that the marginal effect of $RDCA_{it}$ on $RDCBC_{it}$ will tend to be less than $\beta_{3,2}$, and so on.

Table 9: MICE imputation model equations (authority indicator variables omitted)

	$lnTOTEX_{it}$		$\overline{RDCA}_{it}$		$\overline{RDCBC}_{it}$		$\overline{RDCU}_{it}$	
$lnTOTEX_{it}$	-		0.041 (0.02)	**	-0.013 (0.02)		-0.08 (0.031)	**
$\overline{RDCA}_{it}$	0.203 (0.109)	*	-	-	0.533 (0.038)	***	0.003 (0.069)	
$\overline{RDCA}_{it-1}$	0.148 (0.112)		0.276 (0.044)	***	-0.095 (0.045)	**	0.082 (0.069)	
$\overline{RDCA}_{it-2}$	0.025 (0.123)		-0.091 (0.048)	*	-0.054 (0.047)		0.086 (0.071)	
$\overline{RDCBC}_{it}$	-0.087 (0.109)		0.542 (0.039)	***			0.081 (0.07)	
$\overline{RDCBC}_{it-1}$	-0.135 (0.119)		-0.141 (0.05)	***	0.379 (0.046)	***	0.011 (0.077)	
$\overline{RDCBC}_{it-2}$	0.043 (0.119)		0.038 (0.047)		-0.065 (0.046)		-0.008 (0.075)	
$\overline{RDCU}_{it}$	-0.212 (0.079)	***	-0.008 (0.031)		0.04 (0.03)			
$\overline{RDCU}_{it-1}$	0.190 (0.08)	**	-0.006 (0.033)		0.062 (0.032)	*	0.477 (0.044)	***
$\overline{RDCU}_{it-2}$	0.052 (0.078)		0.028 (0.032)		-0.108 (0.031)	***	-0.036 (0.047)	
$\ln LAND_{it}$	-0.698 (3.623)		1.629 (1.499)		4.757 (1.469)	***	-3.892 (2.238)	*
$\ln RL_{it}$	2.923 (4.154)		-2.251 (1.71)		-4.327 (1.684)	***	3.853 (2.547)	
$(\ln LAND_{it})^2$	-0.874 (1.614)		1.191 (0.627)	*	-1.366 (0.62)	**	0.028 (0.973)	
$(\ln RL_{it})^2$	-1.913 (4.823)		2.187 (1.901)		-6.606 (1.86)	***	0.393 (2.932)	
$\ln LAND_{it} \ln RL_{it}$	3.460 (5.779)		-3.244 (2.245)		8.401 (2.195)	***	-2.058 (3.516)	
$\ln TRAFFIC_{it}$	-0.882 (0.811)		-0.239 (0.32)		-1.085 (0.313)	***	-0.411 (0.48)	
$\ln WAGE_{it}$	0.339 (0.186)	*	0.175 (0.072)	**	0.116 (0.072)		-0.025 (0.11)	
R^2	0.947		0.840		0.853		0.811	

* 90% confidence level, ** 95% confidence level, *** 99% confidence level

5.3 Frontier Parameter Estimates

Parameter estimates for the cost frontier are shown in Table 10, corresponding to three alternative approaches: LD, MVN MID, and MIDCE. Parameter estimates and standard errors shown are derived using Rubin's rules as described in section 3.2.2. Each β parameter corresponds to an independent

variable as shown in (16), i.e. β_0 is a constant, β_1 corresponds to $\ln LAND_{it}$, β_2 corresponds to $\ln RL_{it}$, and so on.

We only present parameter estimates from the Multiple Imputation then Deletion methods, given that the Monte Carlo Simulation results (section 4) clearly showed that this approach minimised bias relative including observations with missing dependent variable data in the SFA model estimation stage. Table 11 confirms this finding as it shows the variance parameter estimates across all five potential estimation approaches which the estimates from the models which include all observations do appear to be out of line with the other estimates. Full parameter estimates are given in Appendix A[5].

Table 10: Estimated cost frontier parameters

Parameter	Variable	LD		MIDCE		MVN MID	
β_0		-0.673 (0.088)	***	-0.717 (0.078)	***	-0.534 (0.068)	***
β_1	$\ln LAND$	0.666 (0.110)	***	0.763 (0.102)	***	0.740 (0.100)	***
β_2	$\ln RL$	0.122 (0.133)		0.007 (0.126)		0.164 (0.122)	***
β_3	$(\ln LAND)^2$	0.542 (0.106)	***	0.653 (0.096)	***	0.646 (0.093)	***
β_4	$(\ln RL)^2$	1.279 (0.240)	***	1.472 (0.222)	***	1.433 (0.215)	***
β_5	$\ln LAND \ln RL$	-1.616 (0.314)	***	-1.897 (0.287)	***	-1.900 (0.28)	***
β_6	$\ln TRAFFIC$	0.586 (0.099)	***	0.629 (0.093)	***	0.632 (0.092)	***
β_7	RDC	0.018 (0.004)	***	0.021 (0.003)	***	0.007 (0.001)	***
β_8	$DRDC$	-0.014 (0.007)	*	-0.017 (0.006)	***	-0.004 (0.001)	***
β_9	$\ln WAGE$	1.800 (0.327)	***	1.655 (0.278)	***	1.461 (0.278)	***
β_{10}	$YEAR2$	0.148 (0.078)	*	0.171 (0.071)	**	0.157 (0.07)	***
β_{11}	$YEAR3$	0.247 (0.089)	***	0.267 (0.078)	***	0.243 (0.076)	***
β_{12}	$YEAR4$	0.226 (0.094)	**	0.222 (0.084)	***	0.202 (0.082)	***

[5] We also note that standard errors tend to be largest under the MI approaches, smallest under the MID approaches, and intermediate under LD. This suggests that MID can help to increase the precision of parameter estimates – by increasing the number of observations – relative to LD. However, the additional uncertainty when observations in which the dependent variable is imputed outweighs this advantage, increasing the imprecision of parameter estimates. This is yet another reason to prefer MID methods over MI, over and above any bias issues.

β_{12}	$YEAR5$	0.294 (0.086)	***	0.242 (0.078)	***	0.224 (0.077)	***
β_{13}	$YEAR6$	0.271 (0.086)	***	0.262 (0.078)	***	0.240 (0.076)	***
β_{14}	$YEAR7$	0.250 (0.098)	**	0.24 (0.087)	***	0.204 (0.085)	***
σ_v		0.306 (0.021)		0.309 (0.019)		0.288 (0.019)	
σ_u		0.259 (0.034)		0.257 (0.03)		0.279 (0.029)	
N		425		506		506	

* 90% confidence level, ** 95% confidence level, *** 99% confidence level

Table 11: Estimated variance parameters across all methods

Parameter	Variable	LD	MIDCE	MVN MID	MICE	MVN MI
σ_v		0.306 (0.021)	0.309 (0.019)	0.288 (0.019)	0.336 (0.031)	0.326 (0.029)
σ_u		0.259 (0.034)	0.257 (0.03)	0.279 (0.029)	0.260 (0.038)	0.369 (0.037)
N		425	506	506	588	588

Turning back to Table 10, the parameter estimates appear broadly similar across the three models. There are however some differences in the statistical significance of some parameter estimates, most notably the fact that β_8 which relates to the $DRDC$ variable – which is constructed from variables containing missing values – appears insignificant under LD, but significant or at least weakly significant under all other MI based approaches. This highlights one of the advantages of MI approaches when important auxiliary variables are included in the imputation models.

Regarding the relative efficiency of our parameter estimates under the two MID approaches , our largest fraction of missing information for any parameter – as defined by (5) – and the corresponding relative efficiency of the parameter estimate using the expression derived by Rubin (1987). This puts a lower bound on the relative efficiency of our parameter estimates, given our choice of $D = 10$. Table 12shows that, given the highest fraction of missing information from each model, the relative efficiency of parameter estimates using $D = 10$ are close to 100%. The gains from adding additional imputations are marginal.

Table 12: Largest fractions of missing information and corresponding relative efficiencies

	Largest fraction of missing information: $\boldsymbol{\gamma}$	Relative large sample efficiency: $(\mathbf{1}+\boldsymbol{\gamma}/\boldsymbol{D})^{-\mathbf{1}/\mathbf{2}},\ \boldsymbol{D}=\mathbf{10}$
MIDCE	0.0416	99.8%
MVN MID	0.0701	99.7%

Returns to scale (RTS) at sample mean are an important property of a cost function and so we predict them by noting

$$RTS_{it} = \frac{1}{\beta_1 + \beta_2 + 2(\beta_3 \ln LAND_{it} + \beta_4 \ln RL_{it}) + \beta_5(\ln LAND_{it} + \ln RL_{it})} \tag{22}$$

It should be noted that, since the independent variables are mean-centred – the means being calculated by pairwise deletion for all applications – comparable estimates of RTS at the sample mean are obtained by simply taking the inverse of $\beta_1 + \beta_2$. Table 9 shows that estimated RTS at sample means across the 5 different approaches are all above 1, implying in each case increasing RTS at sample means.

Table 13: Returns to scale at sample means

	LD	MIDCE	MVN MID
RTS	1.270	1.298	1.106

The five models also give similar estimates for RTS on an observation-by-observation basis. As shown in Table 14, observation-specific estimates of RTS across the 5 models are highly correlated. In terms of the scale characteristics of the model, there is therefore very little difference between the estimates obtained using the 5 different approaches.

Table 14: Correlation matrix of returns to scale estimates

	LD	MIDCE	MVN MID
LD	1.0000		
MIDCE	0.994	1.0000	
MVN MID	0.948	0.927	1.0000

Estimated elasticities with respect to traffic density are also similar in each of the 5 models, falling between 0.586 and 0.632, with the exception of a slightly lower estimate of 0.434 from the MVN MI model. These estimates imply that for a 1% increase in traffic density, there is an increase in cost of around 0.6%.

Perhaps most importantly, each model also gives very similar estimates of the coefficients relating to RDC_{it} and $DRDC_{it}$, our independent variables with missing values. These estimates suggest that for a 1-point increase in the proportion of roads that should be considered for maintenance, there is around a 0.006% increase in costs, but that a 1-point improvement from the previous year is associated with an approximately 0.004% increase in cost. The models therefore imply that improving network condition is costly, but that maintaining the network at a given level costs more the worse the condition of the network. Thus there is a saving from investing in improved road condition following a period of extra expenditure. This result is the same as in previous analyses of this dataset, such as in Wheat et al. (2019).

5.4 Efficiency Predictions

We now turn our attention to the prediction and comparison of efficiency scores from each model. As mentioned previously, efficiency prediction in SFA is typically based on the distribution of $u_i|\varepsilon_i$, and specifically the mean of this distribution, $E(u_i|\varepsilon_i)$, as proposed by Jondrow et al. (1982). In the case of a normal-exponential stochastic cost frontier such as we have estimated, $E(u_i|\varepsilon_i)$ is

$$E(u_i|\varepsilon_i) = \sigma_v \left[\frac{\phi(-\varepsilon_i/\sigma_v + \sigma_v/\sigma_u)}{\Phi(\varepsilon_i/\sigma_v + \sigma_v/\sigma_u)} + \varepsilon_i/\sigma_v + \sigma_v/\sigma_u \right] \tag{23}$$

In the case of MI models, since each imputed dataset yields different estimates of the parameters σ_v and σ_u and of the residual ε_i, it is not immediately clear how to proceed to predict efficiency using this framework. White et al. (2011) suggest obtaining separate predictions from each imputed dataset, then combining the predictions using Rubin's rules in the same manner as parameter estimates, obtaining average and standard errors of predictions. Another approach which seems more consistent with the parameter estimates presented, is to use the combined estimates of the σ_v and σ_u parameters – as shown in Table 10– together with a combined estimate of ε_i. We report the latter approach, although we note there is little difference between the two when we implemented the alternative.

In the following sub-sections, we compare efficiency predictions obtained under MIDCE, MVN MID and LD approaches. We first compare prediction for complete cases, i.e. those with no missing data, and then compare predictions in the cases where there are missing data.

5.4.1 Comparison of Complete Cases

We now compare efficiency predictions for complete cases. From a practitioner's point of view, we would ideally like to see that prediction is not unduly sensitive to the method used in these particular cases. Given the absence of missing data in these observations, variation in efficiency predictions between the various methods is purely driven by differences in the estimated parameters.

The differences in parameter estimates between the various methods will affect efficiency predictions in two ways. First, given that in the normal-exponential model the relationship between ε_i and $E(u_i|\varepsilon_i)$ is monotonic, any differences in the ranking of efficiency predictions is driven entirely by the differences in the estimated frontier parameters. Second, and more importantly, the average efficiency score is driven by the parameter σ_u, while the degree of shrinkage of each prediction toward this mean is determined by σ_v. These latter parameters therefore determine the differences in the mean efficiency score, and the spread of efficiency scores, from each model.

Figure 1: Comparison of efficiency predictions for complete cases

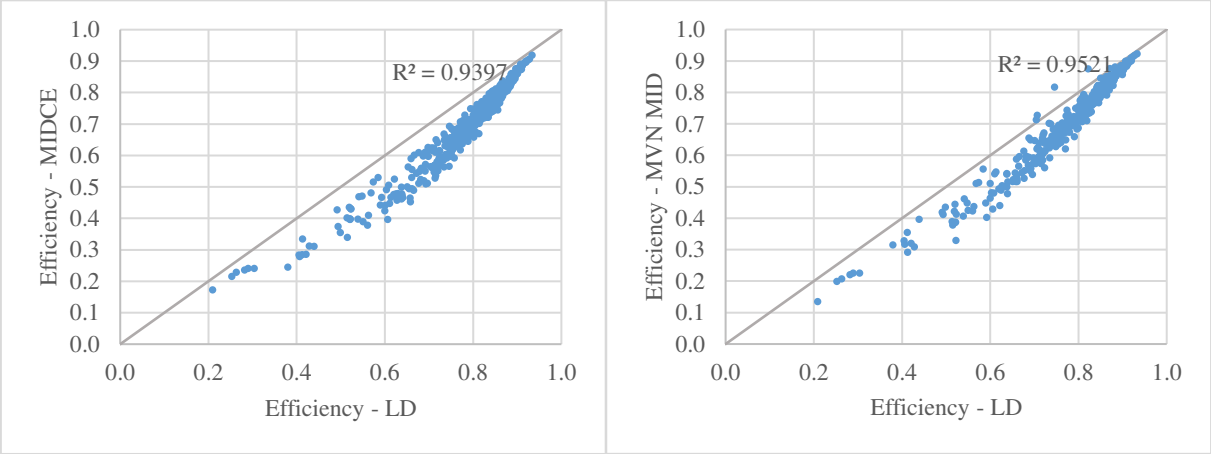


Figure 1 compares efficiency predictions from the MID models to those obtained via LD for all complete cases. We can see that for the two MID models, the efficiency predictions are very highly correlated. However, across the two MID models, the predicted efficiencies for each complete observation is lower than that obtained via LD. This may be explained by the effect of the observations with missing values on the spread of efficiency predictions. Correlations between the efficiency predictions for complete cases obtained via MI methods are shown in Table 15 below; as can be seen, the predictions obtained under the different methods are all highly correlated.

Table 15: Correlation of Efficiency Predictions (Complete Cases Only)

	LD	MIDCE	MVN MID
LD	1.0000		
MIDCE	0.9694	1.0000	
MVN MID	0.9758	0.9771	1.0000

We find that there is a higher concentration of observations with missing values of the dependent variables among the lower efficiency scores and rankings. This is particularly the case with the efficiency predictions from the MI models. Accordingly, the estimated scales of the inefficiency

distributions are greater in these models, so that the mean efficiency predictions are bound to be lower than that from the LD model. We also see from Table 11 that the signal to noise ratio σ_u/σ_v is greater in the MI and MID models than in the LD. The fact that $E(u_i|\varepsilon_i)$ is a shrinkage predictor of u_i, and the degree of shrinkage is inversely related to σ_u/σ_v, explains why the scatter plots in Figure 1 bend towards the 45-degree line at the tails.

5.4.2 Comparison of Incomplete Cases

We now turn our attention to prediction for those observations with missing values. In these cases, in addition to the variation in scores as a result of differing parameter estimates between the various approaches, there is the additional variation resulting from differing imputations of the missing values. Our interest here is twofold: first, we are interested in the sensitivity of efficiency predictions for these cases to the imputation model used. Second, we are interested in the comparison of efficiency predictions for these observations to those out-of-sample predictions under LD (using some simple method of single imputation).

Figure 2 compares efficiency predictions for those observations with missing values of independent variables only. For the types of observations considered above, the sensitivity of efficiency predictions to the imputation method used is due solely to differences in the imputed values of the missing independent variables, both indirectly, through differences in the resulting parameter estimates, and directly, through the impact of the imputed values on the estimated residuals. Figure 2 shows a very high correlation between the predictions from the different models, and that most observations lie close the 45-degree line, so that the magnitudes of the efficiency predictions from the different MID models are very similar. There is one noticeably outlying observation, however, for which the MVN MID models predict 42.4% efficiency and the MIDCE models predict 80.9% efficiency.

Figure 2: Comparison of efficiency predictions when independent variables are missing

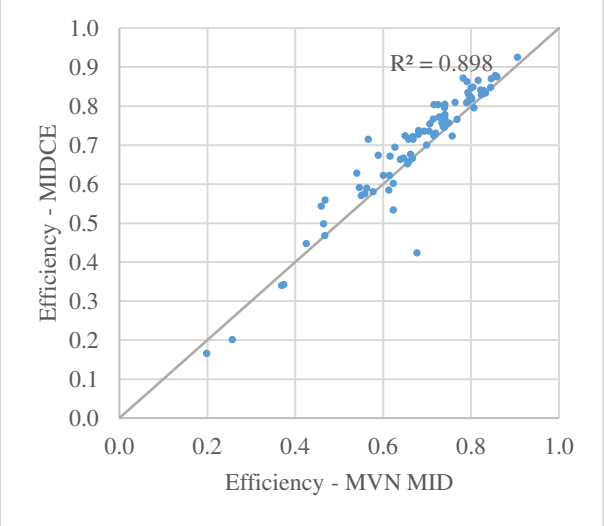
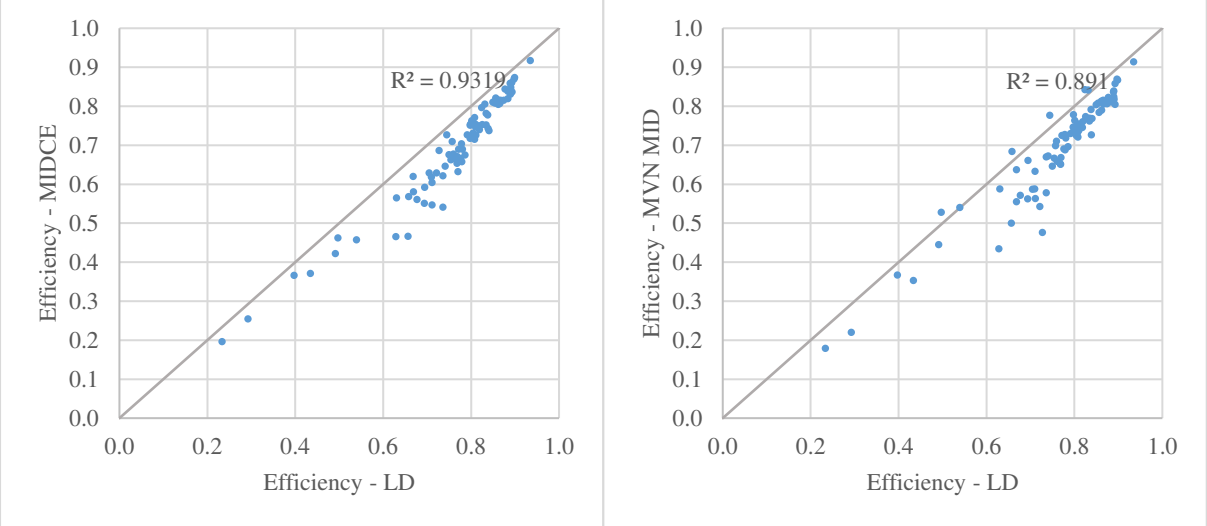


Figure 3 below compares efficiency predictions obtained under both of our MID models to those using out-of-sample prediction under LD. In the latter case, we used a simple linear interpolation (and extrapolation) method to fill in missing values of the independent variables. This is a simple approach but one that would seem an obvious candidate (as discussed in Section 2). Clearly interpolation only uses temporally adjacent values of the same variable (the year before and after RDC values) to impute missing values and not the broad set of variables that are included in the MVN MID and MIDCE approaches. As such we would expect the multiple imputation approaches to give superior imputations and thus better predictions of efficiency.

We find that the correlations between the efficiency predictions under MID and LD (and linear interpolation) are relatively close although there are some large exceptions. Comparing LD and MIDCE, the R-squared is only slightly lower than that for complete cases which is shown in Figure 1, although in terms of magnitude, there are some relatively large differences. Thus there is a difference in using imputations based on a full set of potential variables rather than relying on simple linear interpolation, and as discussed in the previous paragraph we prefer those from the MI techniques, particularly when accuracy for each observation is important as it is in efficiency analysis.

Figure 3: Efficiency prediction under MI and linear interpolation



6 Conclusions

The aim of this paper is to make the case for the use of Multiple Imputation (MI) methods in Stochastic Frontier Analysis (SFA). The issue of missing data is important as depending on the missing data mechanism, the statistical literature indicates that parameter estimates from econometric models maybe both imprecise and biased by only analysing complete cases. A further issue in SFA is the desire to predict efficiency for specific observations and missing data is clearly a barrier to this. In our empirical example nearly 20% of observations have a missing value related to one attribute (road condition) i.e. for 20% of observations no efficiency prediction can be computed.

MI offers a robust alternative to the standard listwise deletion (LD) approach. We demonstrate this by reviewing the statistical properties of MI techniques applied to a general set of models and undertake Monte Carlo simulation to verify that the results hold in the SFA case. We then provide an empirical example, which provides a comprehensive illustration of the feasibility and intuitive results of MI techniques applied to SFA.

We find the following results:

1) In our simulation exercise, MI methods have been shown to increase precision of the estimated model parameters relative to LD, although there does seem to be a greater bias in the inefficiency error variance relative to LD (discussed below). The benefit in MI over LD is clear as the proportion of missing data increases and as the missing data mechanism departs from Missing Completely at Random (MCAR). The later circumstance is important, because missing data tends not to be MCAR and Missing at Random (MAR) is a more realistic assumption. Our simulation results provide a clear rationale for use of MI methods in this case.

   As mentioned above, we find a greater bias in the variance of the inefficiency error component when undertaking MI methods compared to LD. We return to this issue in point 5 below. Despite this, we prefer MI over LD for two reasons. First, maximum likelihood estimation of the SFA model is known to have small sample bias particularly for the variance parameters (Olson et al., 1980). Thus, LD techniques still have bias, as we find in our simulations. Second, when comparing MI to LD, the increase in the root mean squared error (RMSE) –which encompasses both imprecision and bias – for the variance of the inefficiency is small, whereas all other parameters are estimated with substantially lower RMSEs. Thus overall, MI represents an improvement over LD.
2) We have shown how MI methods can exploit auxiliary data (data not used in the SFA model) to enhance the precision of imputations. The MICE method is particularly attractive since it produces a set of imputation equations which expose the influence of each variable to external scrutiny. In our highway maintenance example, the results are clear that the immediate past value of road condition is of great value in predicting the current value as we would expect, but there are other factors such as the condition of other road types which have predictive power on the value of the missing road condition variable. In our empirical example, we also demonstrate methods of taking into account the panel structure of our data, and transformations to restrict the imputed values to a required range, in order to improve the quality of imputations. Through our simulation exercise, and confirmed in our empirical application, we have shown that there is a clear benefit of using MI approaches with deletion (MID approaches) of observations with imputed dependent variables. Whilst using observations with missing dependent variable in the imputation stage is useful, these observations should not be taken forward to the SFA model estimation stage. This is because doing so results in biased estimates

of the variance parameters in the SFA model and thus biased predictions of efficiency. In addition, MID is shown to yield more precise estimates of model parameters in terms of RMSEs.

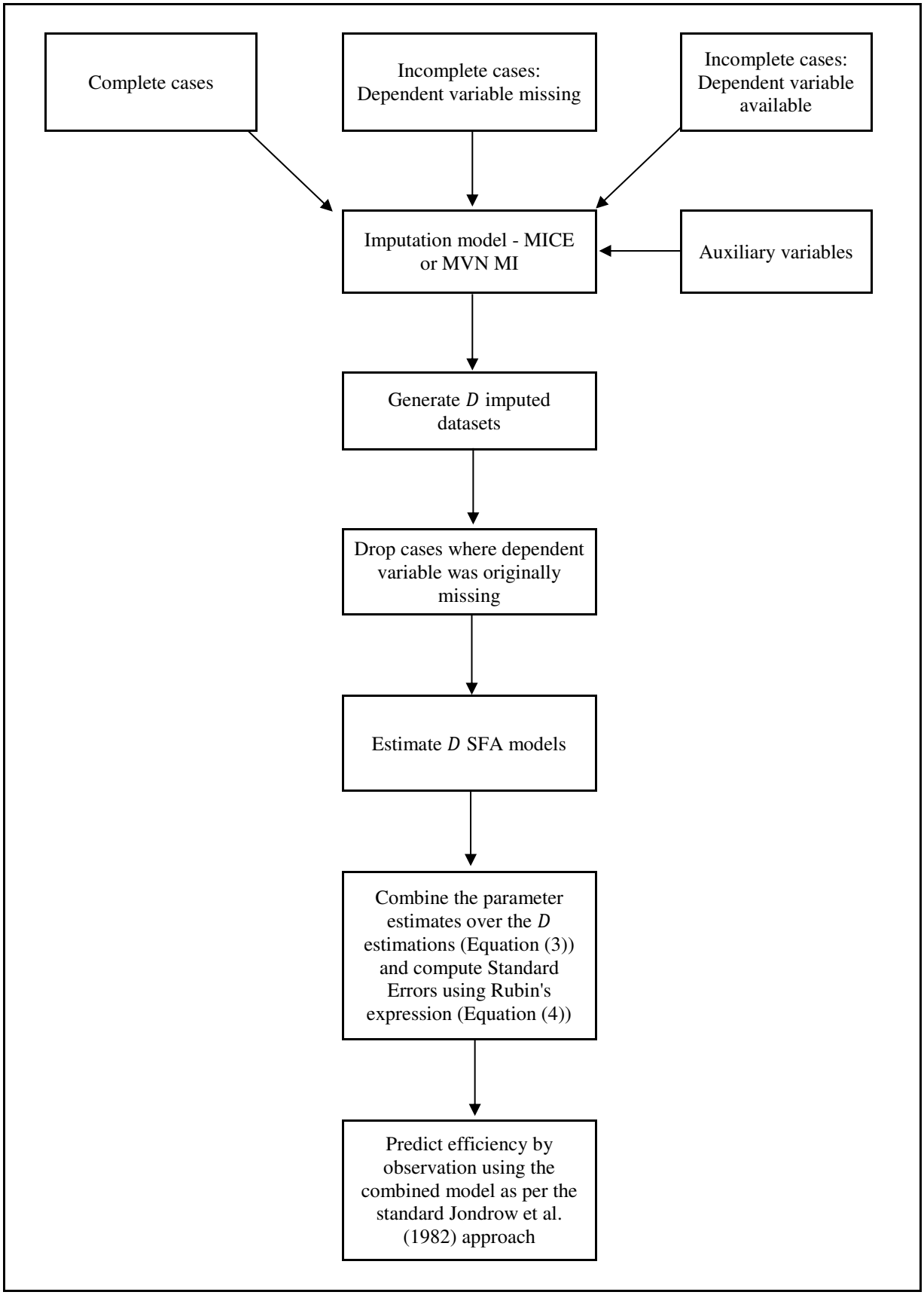
3) Turning to the efficiency predictions, LD, MVN MID and MIDCE are shown to yield very similar efficiency predictions for complete observations. This is in itself a reassuring finding, as it confirms that the MI approaches are consistent with results of more simplistic, and more widely used approaches to deal with missing data. Furthermore, given the reduction in standard errors for the parameter estimates when applying MI relative to LD, it is likely that the residuals computed from the MI approaches are more reflective of the realised (but unobserved) values of the error than from the LD models. The key message here is that the use of MI can improve efficiency predictions for complete observations due to greater precision in the estimated frontier as we implicitly use all available information in estimating the frontier.

   For observations with missing values among the independent variables, MVN MI and MICE are likewise shown to lead to very similar efficiency predictions. A key benefit of MICE is that the analyst can develop and test explicit relations between variables when modelling the missing data. This in turn provides some assurance that missing variable imputations are intuitive which would seem to be an important criterion in choosing between the two approaches, particularly when there is often strong economic and/or engineering expectation of certain linkages between missing and observed data (such as the road condition data in our example). The MICE models are therefore preferred in our setting.

4) Overall we summarise our findings on the recommended approach to using MI methods in SFA in Figure 4. It sets out the stages involved in implementing MI approaches and highlights the use of auxiliary data, the recommendation to use MID rather than MI, and our preferred approach to compute efficiency from the combined parameter estimates across the set of D estimated SFA models.

5) In terms of future research opportunities, we consider that the skewness of the assumed SFA error distribution may lend itself to none symmetric error distributions in the imputation models. Adopting such distribution might eliminate the greater bias in MI approaches relative to LD for the inefficiency variance estimate. A possible avenue for future research would therefore be to see if the use of alternative imputation models, e.g. which allow for skewness in the imputation model, could result in reduced bias in the estimation of the constant and the parameter(s) of the one-sided distribution. Another possible avenue for future research would be to examine the performance of MI under alternative SFA specifications.

Figure 4 Recommended approach for applying Multiple Imputation in SFA



Acknowledgements: The authors acknowledge funding from the CQC Efficiency Network (see http://www.nhtnetwork.org/cqc-efficiency-network/home/).

7 References

Aigner, D., Lovell, C.A.K. and Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics.* **6**(1), pp.21-37.

Allison, P.D. 2002. Missing data: Sage University Papers Series on quantitative applications in the social sciences, 07-136. *Thousand Oaks, CA.*

Arnold, B.C., Castillo, E. and Sarabia, J.M. 1999. *Conditional Specification of Statistical Models.* Springer Science & Business Media.

Arnold, B.C., Castillo, E. and Sarabia, J.M. 2001. Conditionally Specified Distributions: An Introduction (with comments and a rejoinder by the authors). pp.249-274.

Barnard, J. and Rubin, D.B. 1999. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika.* **86**(4), pp.948-955.

Battese, G.E. and Coelli, T.J. 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics.* **38**(3), pp.387-399.

Bentler, P.M. 2000. Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization AU - Gold, Michael Steven. *Structural Equation Modeling: A Multidisciplinary Journal.* **7**(3), pp.319-355.

Bhandari, B.D., Gillespie, J. and Scaglia, G. 2015. *Efficiency of U.S. Grass-Fed Beef Farms.* Southern Agricultural Economics Association.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L. and Kirchner, U. 2002. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* **162**(2), pp.195-209.

Charnes, A., Cooper, W.W. and Rhodes, E. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research.* **2**(6), pp.429-444.

Glasser, M. 1964. Linear Regression Analysis with Missing Observations among the Independent Variables. *Journal of the American Statistical Association.* **59**(307), pp.834-844.

Glynn, R. and Laird, N. 1986. *Regression estimates and missing data: complete case analysis*. Unpublished.

Haitovsky, Y. 1968. Missing Data in Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological).* **30**(1), pp.67-82.

Heimeshoff, M., Schreyögg, J. and Kwietniewski, L. 2014. Cost and technical efficiency of physician practices: a stochastic frontier approach using panel data. *Health Care Management Science.* **17**(2), pp.150-161.

Herzog, T.N. and Rubin, D.B. 1983. Using multiple imputations to handle nonresponse in sample surveys. In: Madow, W.G., Olkin, I. and Rubin, D.B. eds. *Incomplete data in sample surveys.* Academic Press, pp.209-245.

Jensen, U., Gartner, H. and Rässler, S. 2010. Estimating German overqualification with stochastic earnings frontiers. *AStA Advances in Statistical Analysis.* **94**(1), pp.33-51.

Jensen, U. and Rässler, S. 2006. Stochastic production frontiers with multiply imputed German establishment data. *Zeitschrift für ArbeitsmarktForschung - Journal for Labour Market Research.* **39**(2), pp.277-295.

Jensen, U. and Rässler, S. 2007. *The Effects of Collective Bargaining on Firm Performance: New Evidence Based on Stochastic Production Frontiers and Multiplu Imputed German Establishment Data.* Instituts für Arbeitsmarkt- und Berufsforschung, No. 2003, 3, Nuremburg, Germany.

Jondrow, J., Knox Lovell, C.A., Materov, I.S. and Schmidt, P. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics.* **19**(2), pp.233-238.

Kim, J.K. 2004. Finite Sample Properties of Multiple Imputation Estimators. *The Annals of Statistics.* **32**(2), pp.766-783.

Kölling, A. and Rässler, S. 2004. *Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models.* Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Kropko, J., Goodrich, B., Gelman, A. and Hill, J. 2014. Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches. *Political Analysis.* **22**(4), pp.497-519.

Kwietniewski, L. and Schreyögg, J. 2018. Efficiency of physician specialist groups. *Health Care Management Science.* **21**(3), pp.409-425.

Lee, K.J. and Carlin, J.B. 2010. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology.* **171**(5), pp.624-632.

Little, R.J.A. 1992. Regression with Missing X's: A Review. *Journal of the American Statistical Association.* **87**(420), pp.1227-1237.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data.* Wiley.

Meeusen, W. and van Den Broeck, J. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review.* **18**(2), pp.435-444.

Neyman, J. 1934. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society.* **97**(4), pp.558-625.

Olson, J.A., Schmidt, P. and Waldman, D.M. 1980. A Monte Carlo study of estimators of stochastic frontier production functions. *Journal of Econometrics.* **13**(1), pp.67-82.

Qushim, B., Gillespie, J. and McMillin, K. 2016. Meat goat enterprise efficiency analysis in the Southeastern United States. *Journal of Agricultural and Applied Economics.* **48**(1), pp.52-72.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology.* **27**(1), pp.85-96.

Raghunathan, T.E. and Siscovick, D.S. 1996. A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* **45**(3), pp.335-352.

Rubin, D.B. 1976. Inference and Missing Data. *Biometrika.* **63**(3), pp.581-592.

Rubin, D.B. 1978a. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics.* **6**(1), pp.34-58.

Rubin, D.B. 1978b. Multiple imputation in sample surveys. In: American Statistical Association ed. *Proceedings of the Section on Survey Research Methods, 1978.* pp.20-34.

Rubin, D.B. 1986. Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics.* **4**(1), pp.87-94.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Rubin, D.B. 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association.* **91**(434), pp.473-489.

Rubin, D.B. and Schenker, N. 1986. Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association.* **81**(394), pp.366-374.

Rubright, J.D., Nandakumar, R. and Glutting, J.J. 2014. A simulation study of missing data with multiple missing X's. *Practical Assessment, Research & Evaluation.* **19**(10), pp.2-8.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* CRC press.

Schafer, J.L. and Graham, J.W. 2002. Missing data: our view of the state of the art. *Psychological methods.* **7**(2), pp.147-177.

Schenker, N. and Taylor, J.M.G. 1996. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis.* **22**(4), pp.425-446.

Schenker, N. and Welsh, A.H. 1988. Asymptotic Results for Multiple Imputation. *The Annals of Statistics.* **16**(4), pp.1550-1566.

Shaik, S. and Tokovenko, O. 2013. Evaluating the Importance of Multiple Imputations of Missing Data on Stochastic Frontier Analysis Efficiency Measures. In: *Agricultural and Applied Economics Association and Canadian Agricultural Economics Society Joint Annual Meeting*, *4th-6th August 2013, Washington, DC, USA*.

Shephard, R.W. 1953. *Cost and Production Functions.* First ed. Princeton University Press.
Smith, K.M. and Cummings, P. 2004. Passenger seating position and the risk of passenger death or injury in traffic crashes. *Accident Analysis & Prevention.* **36**(2), pp.257-260.
Smith, S.D. and Forster, M. 2013. *'The curse of the Caribbean'? Agency's impact on the efficiency of sugar estates in St. Vincent and the Grenadines, 1814-1829.* Oxford University Economic and Social History Series, No. 112, University of Oxford, Department of Economics.
Stead, A.D., Wheat, P. and Greene, W.H. 2018. Estimating efficiency in the presence of extreme outliers: A logistic-half normal stochastic frontier model with application to highway maintenance costs in England. In: Greene, W.H., Khalaf, L., Makdissi, P., Sickles, R.C., Veall, M. and Voia, M. eds. *Productivity and Inequality.* Springer, pp.1-19.
Steimetz, S.S.C. and Brownstone, D. 2005. Estimating commuters' "value of time" with noisy data: a multiple imputation approach. *Transportation Research Part B: Methodological.* **39**(10), pp.865-889.
Sullivan, T.R., Salter, A.B., Ryan, P. and Lee, K.J. 2015. Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data. *American Journal of Epidemiology.* **182**(6), pp.528-534.
van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research.* **16**(3), pp.219-242.
van Buuren, S., Boshuizen, H.C. and Knook, D.L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine.* **18**(6), pp.681-694.
van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation.* **76**(12), pp.1049-1064.
von Hippel, P.T. 2005. How Many Imputations Are Needed? A Comment on Hershberger and Fisher (2003). *Structural Equation Modeling: A Multidisciplinary Journal.* **12**(2), pp.334-335.
von Hippel, P.T. 2007. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology.* **37**, pp.83-117.
Wang, W.S. and Schmidt, P. 2009. On the distribution of estimated technical efficiency in stochastic frontier models. *Journal of Econometrics.* **148**(1), pp.36-45.
Wheat, P. 2017. Scale, quality and efficiency in road maintenance: Evidence for English local authorities. *Transport Policy.* **59**, pp.46-53.
Wheat, P., Stead, A.D. and Greene, W.H. 2019. Robust stochastic frontier analysis: a Student's t-half normal model with application to highway maintenance costs in England. *Journal of Productivity Analysis.* **51**(1), pp.21-38.
White, I.R., Royston, P. and Wood, A.M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine.* **30**(4), pp.377-399.
Yu, L.M., Burton, A. and Rivero-Arias, O. 2007. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research.* **16**(3), pp.243-258.

Appendix – Full parameter estimates

Table 16: Estimated cost frontier parameters including MICE and MVN MI

Parameter	Variable	LD		MIDCE		MVN MID		MICE		MVN MI	
β_0		-0.673 (0.088)	***	-0.717 (0.078)	***	-0.534 (0.068)	***	-0.650 (0.091)	***	-0.669 (0.079)	***
β_1	$\ln LAND$	0.666 (0.110)	***	0.763 (0.102)	***	0.740 (0.100)	***	0.667 (0.125)	***	0.562 (0.118)	***
β_2	$\ln RL$	0.122 (0.133)		0.007 (0.126)		0.164 (0.122)	***	0.124 (0.163)		0.349 (0.141)	**
β_3	$(\ln LAND)^2$	0.542 (0.106)	***	0.653 (0.096)	***	0.646 (0.093)	***	0.538 (0.193)	**	0.588 (0.115)	***
β_4	$(\ln RL)^2$	1.279 (0.240)	***	1.472 (0.222)	***	1.433 (0.215)	***	1.206 (0.416)	***	1.332 (0.263)	***
β_5	$\ln LAND \ln RL$	-1.616 (0.314)	***	-1.897 (0.287)	***	-1.900 (0.28)	***	-1.550 (0.571)	**	-1.756 (0.342)	***
β_6	$\ln TRAFFIC$	0.586 (0.099)	***	0.629 (0.093)	***	0.632 (0.092)	***	0.587 (0.094)	***	0.434 (0.118)	***
β_7	RDC	0.018 (0.004)	***	0.021 (0.003)	***	0.007 (0.001)	***	0.020 (0.004)	***	0.006 (0.001)	***
β_8	$DRDC$	-0.014 (0.007)	*	-0.017 (0.006)	***	-0.004 (0.001)	***	-0.020 (0.006)	***	-0.004 (0.002)	*
β_9	$\ln WAGE$	1.800 (0.327)	***	1.655 (0.278)	***	1.461 (0.278)	***	1.569 (0.292)	***	1.526 (0.338)	***
β_{10}	$YEAR2$	0.148 (0.078)	*	0.171 (0.071)	**	0.157 (0.07)	***	0.133 (0.07)	*	0.201 (0.080)	**
β_{11}	$YEAR3$	0.247 (0.089)	***	0.267 (0.078)	***	0.243 (0.076)	***	0.236 (0.078)	***	0.306 (0.087)	***
β_{12}	$YEAR4$	0.226 (0.094)	**	0.222 (0.084)	***	0.202 (0.082)	***	0.185 (0.087)	*	0.268 (0.096)	***
β_{12}	$YEAR5$	0.294 (0.086)	***	0.242 (0.078)	***	0.224 (0.077)	***	0.205 (0.081)	**	0.261 (0.088)	***
β_{13}	$YEAR6$	0.271 (0.086)	***	0.262 (0.078)	***	0.240 (0.076)	***	0.208 (0.08)	***	0.282 (0.089)	***
β_{14}	$YEAR7$	0.250 (0.098)	**	0.24 (0.087)	***	0.204 (0.085)	***	0.191 (0.089)	**	0.244 (0.102)	**
σ_v		0.306 (0.021)		0.309 (0.019)		0.288 (0.019)		0.336 (0.031)		0.326 (0.029)	
σ_u		0.259 (0.034)		0.257 (0.03)		0.279 (0.029)		0.260 (0.038)		0.369 (0.037)	
N		425		506		506		588		588	

* 90% confidence level, ** 95% confidence level, *** 99% confidence level