International Journal of Food and Agricultural Economics
ISSN 2147-8988
Vol. 2 No. 2 pp. 29-48

IMPUTATION METHODS AND APPROACHES: AN ANALYSIS OF PROTEIN SOURCES IN THE MEXICAN DIET

Jose Antonio Lopez
Department of Agricultural Sciences, Texas A&M University-Commerce, P.O. Box 3011, Commerce, TX 75429-3011, USA, Email: Jose.Lopez@tamuc.edu

Abstract

Several imputation approaches using a large sample and different levels of censoring are compared and contrasted following a multiple imputation methodology. The study not only discusses these imputation approaches, but also quantifies differences in price variability before and after price imputation, evaluates the performance of each method, and estimates and compares parameters and elasticities from a complete demand system. The study's findings reveal that small variability among the mean prices from the various imputation approaches may result in relatively larger variability among the underlying parameter estimates of interest and the ultimately desired measures. This suggests that selection bias may be avoided or reduced by validating the imputation approaches and choosing the imputation method based on an analysis of the ultimately desired measures.

Keywords: *Censored prices, elasticities, imputation methods, multiple imputation, protein demand.*

1. Introduction

Survey design, implementation, and institutional constrains often lead to a frequently encountered problem with consumer survey data, the existence of censored observations. With home-scan data and household surveys being more accessible to researchers, this problem of increasing importance is becoming more and more common; usually taking place in high proportions (e.g., Taylor, Phaneuf, & Piggott, 2008; Dong, Gould, & Kaiser, 2004; Gould, Lee, Dong, & Villarreal, 2002; Golan, Perloff, & Shen, 2001; Sabates, Gould, & Villarreal, 2001; Dong & Gould, 2000; Heien, Jarvis, & Perali, 1989; Cox & Wohlgenant, 1986) in dependent variables, independent variables, or both. It occurs when the value of an observation is partially known (also called item nonresponse). This happens when the value of a variable of interest (e.g., the dependent variable) is unknown; but information on related variables (e.g., the independent variables) is known.

When there is item nonresponse only on the dependent variable, applied economists use parametric models (e.g., the probit and tobit models, or their multinomial versions). However, when dealing with item nonresponse on an independent variable, a surprisingly high number of studies such as Golan et al. (2001) and Dong et al. (2004) use very simple techniques (e.g., simple regional or quarterly averages) or omit missing observations. These simple techniques are often satisfactory but it is critical to compare and assess their performance with other imputation methods and approaches to determine if selection bias can be avoided or minimized. There are several ways in which a missing value can be substituted with a replacement value: deductive imputation, cell mean imputation, hot-deck imputation, and cold-deck imputation. Unfortunately, some of these methods may be time

consuming (e.g., deductive imputation) or perhaps unfeasible (e.g., cold-deck imputation) when the data sample is large and/or the data is limited.[1]

In deductive imputation the researcher deduces the missing value by using logic and the relationships among the variables. For instance, if the geographical location of a household is missing, it can be recovered by using other variables such as the consecutive order of household interviews and the time period when the household was interviewed. If the previously and the subsequently interviewed households were interviewed during the same week and they both belong to the same city, then the logical imputation for the missing geographical location would be to use the same city from the two other households.

Cell mean imputation consists of grouping the observations (e.g., households) into classes (e.g., strata and state) and using the non-missing values of the variable of interest (e.g., non-missing prices) to impute the missing values of this or another variable of interest (e.g., missing prices). Cell mean imputation has been employed by Golan et al. (2001, p. 545) and Dong et al. (2004, p. 1099). Clearly, the more specific the classes are (e.g., strata and county), the more likely the researcher is to obtain an estimate that is closer to the true value. Cell mean imputation is appropriate if the missing values are missing completely at random. The disadvantage of this method is that the variance in the imputed variable decreases.[2] To avoid losing variability in the variable of interest, the researcher may alternatively use the mean and standard deviation from the non-missing values of the variable of interest and generate values for imputation from a normal distribution with this mean and this standard deviation.

Lohr (1999, p. 275) explains that the term *hot deck* dates back to the time computer programs and datasets were punched on cards. The card reader used to warm the data cards, so the term *hot deck* was used to refer to the data cards being analyzed. Similar to cell mean imputation, after the observations have been grouped into classes, hot deck imputation uses a non-missing value of the variable of interest to impute the missing values of this or another variable of interest. The non-missing value may be the previous non-missing value in the class, a non-missing value chosen at random in the class, or the nearest non-missing value in the cell, where the distance may be defined according to some criteria that is based on another variable.

Contrary to hot deck, cold deck imputation uses a dataset other than the dataset being analyzed to impute the missing value. These datasets may be from a previous survey or from another source. Cold deck imputation is common in time series datasets. The researcher sometimes pulls data from different sources to complete a time series for a particular variable of interest on which little information is available.

The various imputation methods can be compared following a multiple imputation methodology (see Lohr, 1999; Rubin, 1996; Rubin, 1987). In multiple imputations, a missing value is imputed more than once by using different imputation methods. Each imputation method generates a new dataset with non-missing observations. Each dataset is then analyzed as if no imputation had been done. "[T]he different results give the analyst a measure of the additional variance due to imputation" (Lohr 1999, p. 277). Typically, the same model is used to analyze each imputed dataset.

This research paper compares and contrast several price imputation approaches, under large samples and different levels of censoring following a multiple imputation methodology. The general objective is to discuss and compare imputation approaches by using a complete demand system model on the imputed datasets. In particular, an Almost Ideal Demand System (AIDS) that incorporates the restrictions of adding-up, homogeneity, and symmetry is employed. The paper not only discusses various imputation approaches, but also quantifies the differences in price variability before and after imputation, evaluates the performance of each method under different levels of missing data, and estimates and compares the ultimately desired parameter estimates (i.e., the parameter estimates obtained

from the complete demand system) under each imputation procedure. There are few studies in the existing literature that have explored this critical issue using a large cross-sectional sample and different levels of censoring.

To accomplish the general objective of the study, data on prices of several important protein sources in the Mexican diet (meat, dairy, eggs, tubers, vegetables, legumes, and fruits) are used from the 2008 survey of Mexican household incomes and expenditures (*Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)).* ENIGH 2008 is a recent and reliable source of information, and it is published by a Mexican governmental institution (*Instituto Nacional de Estadística, Geografía e Informática (INEGI)).* In ENIGH 2008, a total of 29,468 households were interviewed. A comparison of price imputation methods under different levels of censoring is ideal with this survey because the sample of households is large.

This study's findings reveal that small variability among the price imputation approaches may lead to large variability among the underlying parameter estimates of interest and the ultimately desired measures (e.g., measures of price responsiveness). This means researchers have to be very careful when choosing a price imputation procedure if they want to avoid or reduce selection bias. However, results seem to indicate that it is possible that a simple cell mean price imputation that uses two levels of urbanization within Mexico's 31 states and the Federal District (e.g., Golan et al., 2001; Dong et al., 2004) and results in a considerable lost of price variability, may lead to parameter estimates that are satisfactory under lager levels of censoring. Similarly, simply excluding the censored observations may result in a significant sample size reduction, but it may also lead to parameter estimates that are satisfactory under large levels of censoring.

2. Methods and Procedures

2.1. Imputer's Models

The cell mean imputation method is also referred to as a zero-order missing price procedure (Cox & Wohlgenant 1986, p. 913). Researches such as Golan et al. (2001, p. 545) and Dong et al. (2004, p. 1099) have employed this method. For instance, to impute prices for Mexican households that did not make meat purchases, Golan et al. (2001, p. 545) "assume[d] that those households face the average price level for that product in that particular location: a rural or urban area in a particular state or federal district." Similarly, "[f]or [Mexican] households not purchasing a particular commodity, [Dong et al. (2004, p. 1099)] replace[d] unobserved unit values with the average unit value obtained by purchasing households in the same area, represented by state of residence and degree of urbanization."

Other researchers such as Zheng and Henneberry (2009, p. 878) have used Cox and Wohlgenant's (1986, p. 913) first-order missing price procedure. Using the non-missing prices of commodity i, this method first computes the regional mean prices (mp_i) and then calculates the corresponding deviations from the regional mean prices (dmp_i). That is,

$$dpm_i = p_i - mp_i. \tag{1}$$

Subsequently, this method regresses dmp_i as a function household characteristics, which are proxies for household preferences for unobserved household characteristics. That is,

$$dpm_i = z_i'\beta_i + e_i, \tag{2}$$

where z_i' is a $(1 \times K)$ vector of household characteristics, β_i is a $(K \times 1)$ vector of parameters, and e_i is random error. Interested in analyzing unit values, Cox and Wohlgenant (1986, p. 913) assume that the deviations from mean prices reflect quality differences that are induced by household characteristics and nonsystematic supply-related factors. Substituting equation (2) into (1) and solving for p_i gives the price/quality functions. The OLS parameter estimates obtained from equation (2) are used to predict the values of the missing prices. The quality-adjusted missing price estimates or imputed prices are obtained from

$$\tilde{p}_i = \widehat{dmp}_i + mp_i \tag{3}$$

where $\tilde{p}_i$ is an estimate of p_i for the corresponding missing prices.

A similar but simpler approach consists of regressing the non-missing prices as a function of household characteristics, and then using the resulting parameters estimates to predict the missing prices. This method is referred to as a simple regression imputation approach (e.g., Lopez, Malaga, Chidmi, Belasco, & Surles, 2012).

Other imputation methods are also based on algorithms. These include the expectation-maximization (EM) algorithm and the Markov Chain Monte Carlo (MCMC) algorithm. The EM algorithm finds the MLE of the vector of parameters by iterating two steps until the iterations converge. The expectation step (E-step) computes the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates.[3] The maximization step (M-step) estimates the parameters that maximize the complete-data log likelihood from the E-step. For multivariate normal data, the observed-data log likelihood being maximized can be expressed as

$$log\ L(\theta|X_{obs}) = \sum_{g=1}^{G} log\ L_g(\theta|X_{obs}) \tag{4}$$

where G is the number of groups with distinct missing patterns, log $L(\boldsymbol{\theta}|X_{obs})$ is the observed-data log likelihood from the g^{th} group, and

$$log\ L_g(\theta|X_{obs}) = -\frac{n_g}{2} log\left|\Sigma_g\right| - \frac{1}{2}\sum_{hg}\left(x_{hg} - \mu_g\right)'\Sigma_g^{-1}\left(x_{hg} - \mu_g\right) \tag{5}$$

where n_g is the number of observations in the g^{th} group, the summation is over the household observations in the g^{th} group, x_{hg} is a vector of observed values corresponding to observed variables, $\mu_g$is the corresponding mean vector, and $\boldsymbol{\Sigma}_g$ is the associated covariance matrix. Schafer (1997, pp. 163-181) and SAS Institute Inc. (2004, pp. 2536-2537) provide a description of the EM algorithm for multivariate normal data.

The MCMC algorithm has applications in Bayesian inference. The entire joint posterior distribution of the unknown numbers can be simulated to obtain posterior parameter estimates of interest. The Bayesian approach to missing data consists of a data augmentation procedure that is implemented in two steps. The imputation step (I-step) draws values for X_{mis} from a conditional predictive distribution of X_{mis} given X_{obs}. That is, with a current estimate of $\theta^{(t)}$ at the t^{th} iteration,

$$X_{mis}^{(t+1)} \sim Pr(X_{mis}|X_{obs}, \theta^{(t)}). \tag{6}$$

The posterior step (P-step) draws values for θ from a conditional distribution of θ given X_{obs}. That is,

$$\theta^{(t+1)} \sim Pr(\theta|X_{obs}, X_{mis}^{(t+1)}). \tag{7}$$

The steps in equations (6) and (7) are iterated creating a Markov chain

$$\left(X_{mis}^{(1)}, \theta^{(1)}\right), \left(X_{mis}^{(2)}, \theta^{(2)}\right), \ldots, \tag{8}$$

which converges in distribution to $Pr(X_{mis}, \theta | X_{obs})$. It is assumed that this distribution is stationary (SAS Institute Inc. 2004, p. 2548). Schafer (1997) and SAS Institute Inc. (2004, pp. 2547-2552) provide a description of the MCMC technique. In this paper, the MCMC method uses a multiple chain to produce five (if a 30% censored level) or ten imputations (if a 70% censored level).[4] Five imputations will produce a relative efficiency of 0.9434 when there is a 30% censoring level while ten imputations will produce a relative efficiency of 0.9346 when there is a 70% censoring level (SAS Institute Inc. 2004, p. 2562). In addition, two-hundred iterations were performed for the first and subsequent imputations. The EM algorithm is used to derive the set of initial parameter values for the MCMC and a noninformative prior was used in the P-step (see Schafer 1997, p. 154; & SAS Institute Inc. 2004, p. 2550). The imputations were combined by computing the average of the *m* complete-data estimates (see SAS Institute Inc., p. 2561).

2.2. Analyst's Model

In practice, the model that is used to impute the data (i.e., the imputer's model) is not the same as the model used to analyze the imputed data (i.e., the analyst's model). This paper uses the Almost Ideal Demand System (AIDS) to analyze the imputed datasets that were obtained from different imputation methods and different levels of censoring. The results provide a measure of the additional variance obtained due to imputation (Lohr 1999, p. 277).

The Almost Ideal Demand System (AIDS) was developed by Deaton and Muelbauer (1980) as an arbitrary first order approximation of any demand system. The functional form is consistent with household-budget data and it is not difficult to estimate. In the AIDS model, the Marshallian demand function for commodity *i* in share form is specified as

$$w_{ih} = \alpha_i + \sum_j \gamma_{ij} \log(p_{jh}) + \beta_i \log(m_h/P_h) + \varepsilon_{ih} \tag{9}$$

where w_{ih} is the budget share for commodity i and household h; p_{jh} is the price of commodity j and household h, m_h is total household expenditure on the commodities being analyzed; α_i, β_i and γ_{ij} are parameters, and ε_i is a random term of disturbances, and P_h is a price index.

In a nonlinear approximation, the price index P_h is defined as

$$\log(P_h) = \alpha_0 + \sum_k \alpha_k \log(p_{kh}) + \frac{1}{2} \sum_k \sum_j \gamma_{kj} \log(p_{kh}) \log(p_{jh}) \tag{10}$$

In the linear approximation of the AIDS model (LA/AIDS) suggested by Stone (1954), equation (5) is estimated by

$$\log(P_h^*) = \sum_k w_{kh} \log(p_{kh}) \tag{11}$$

The demand theory properties of adding-up, homogeneity and symmetry can be imposed on the system of equations by restricting parameters in the model as follows:

Adding-up: $$\sum_i \alpha_i = 1, \sum_j \gamma_{ij} = 0, \text{ and } \sum_i \beta_i = 0; \tag{12}$$

Homogeneity: $$\sum_i \gamma_{ij} = 0; \tag{13}$$

Symmetry: $$\gamma_{ij} = \gamma_{ji}. \tag{14}$$

In the AIDS model, the Marshallian (uncompensated) and the Hicksian (compensated) price elasticities as well as the expenditure elasticities can be computed from the estimated coefficients as follows:

Marshallian Price Elasticity: $$e_{ij} = -\delta_{ij} + \gamma_{ij}/w_i - \beta_i/w_i \left(\alpha_j + \sum_k \gamma_{kj} \ln p_k \right) \quad (15)$$

Hicksian Price Elasticity: $$e^c_{ij} = e_{ij} + w_j e_i \quad (16)$$

Expenditure Elasticity: $$e_i = 1 + \beta_i/w_i \quad (17)$$

where δ is the Kronecker delta equal to one if $i = j$ and equal to zero otherwise.

One equation is omitted in the estimation of this system, but the parameters of that equation will be recovered by making use of the theoretical classical properties. Usually the equation excluded is the one holding the smallest budget share.

2.3. Data

Mexican data on household income and weekly expenditures was obtained from *Encuesta Nacional de Ingresos y Gastos de los Hogares* (2008), which is a nation-wide survey encompassing Mexico's 31 states plus one Federal District (a territory which belongs to all states). ENIGH is a cross-sectional data sample published since 1977 (e.g., see Heien et al., 1989) by a Mexican governmental institution (*Instituto Nacional de Estadística, Geografía e Informática* (INEGI)). ENIGH collects data by giving direct interviews and recording household expenditures on groceries and several other items for one week.

Seven food sources of protein were analyzed in this study. These are meat (which includes beef, pork, processed meat, chicken, processed poultry meat, seafood, and other meats), dairy (which includes milk, cheese, and other milk derived products), eggs, tubers (which includes raw, fresh, and processed tubers), vegetables (which includes fresh and processed vegetables and pod vegetables), legumes (fresh and processed), and fruits (fresh and processed). More specific information about the food products included in each category can be obtained from ENIGH (2008).

In this study, a subsample of 3,572 households containing non-missing prices and quantities of several important protein sources in the Mexican diet is used. To accomplish the objectives of the study, prices from this dataset were randomly censored at levels of 30% (2,500 non-missing price observations and 1,072 censored price observations) and 70% (1,072 non-missing price observations and 2,500 censored price observations). In this study, all prices were censored for the same instances; therefore, this study considers only one missing data pattern (i.e., only one group of observations can be formed with the resulting dataset).[5] Quantities, on the other hand, were not censored.

3. Results and Discussion

Several approaches to data imputation were explored in this study: excluding censored observations (ECO), cell mean imputation (CM), Cox and Wohlgenant's first-order missing price procedure (CW), simple regression imputation (SR), the EM algorithm, and the MCMC algorithm. The ECO approach discarded the censored observations and focused only on the non-censored observations. The CM method considered only one cell; therefore, it replaced censored prices with the simple average of the non-censored prices. The CW and SR methods are regression imputation approaches.[6] Table 1 summarizes the eighteen variables

that were used as proxies for household preferences in the CW method (see equation (2)) and in the SR method. Three of these variables were excluded to avoid the multicollinearity (*urban*, *university*, and *C*). This means that the baseline consists of urban households from the central region whose decision makers have completed university education or graduate school education.[7] The EM and the MCMC algorithms compute the maximum likelihood estimate (MLE) for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the data with the missing values, assuming a multivariate normal distribution. The numeric variables included in the EM and MCMC algorithms were the seven protein prices and fifteen variables from Table 1 (*urban*, *university*, and *C* excluded).[8]

Table 2 reports the means and standard errors of the various protein categories under no censoring, a 30% censoring level, and a 70% censoring level. The column titled No Censoring reports the means and standard errors of the means for the entire 3,572 households with no censored observations. This column can be used to validate the various imputation approaches that were explored.

Table 1. Proxy Variables for Household Preferences

Variable	Description
p00_11	Number of household members who are less than 12 years old.
p12_64	Number of household members who are or are between 12 and 64 years old.
p65_more	Number of household members who are or are older than 65 years old.
inc	Household income.
rural	This variable takes the value of 1 for household locations with a population of 14,999 people or less and 0 if otherwise.
urban	This variable takes the value of 1 for household locations with a population of 15,000 people or more and 0 if otherwise.
element	This variable takes the value of 1 if the household decision maker has elementary school education or less and 0 if otherwise.
highsch	This variable takes the value of 1 if the household decision maker has high school education or if he/she is a high school graduate and 0 if otherwise.
college	This variable takes the value of 1 if the household decision maker has some college, college or incomplete university education and 0 if otherwise.
university	This variable takes the value of 1 if the household decision maker has completed university or has some graduate school education and 0 if otherwise.
NE	This variable takes the value of 1 if the household is located in the Northeast region of Mexico and 0 if otherwise.
NW	This variable takes the value of 1 if the household is located in the Northwest region of Mexico and 0 if otherwise.
CW	This variable takes the value of 1 if the household is located in the Central-West region of Mexico and 0 if otherwise.
C	This variable takes the value of 1 if the household is located in the Central region of Mexico and 0 if otherwise.
SE	This variable takes the value of 1 if the household is located in the Southeast region of Mexico and 0 if otherwise.
d_car	This variable takes the value of 1 if the household has a 4-wheel vehicle and 0 if otherwise.
d_refri	This variable takes the value of 1 if the household has a refrigerator at home and 0 if otherwise.
supermkt	This variable takes the value of 1 if the household purchased the protein product or commodity from a supermarket and 0 if somewhere else.

Table 2. Observed and Imputed Prices (n = 3,572)

	No Censoring		30 % Censoring Level											
	Observed Prices		Excluding Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
p_i	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	47.0064	0.4462	47.0064	0.3071	47.0651	0.3141	46.9953	0.3124	46.9953	0.3124	46.9948	0.3123
p_2	23.7807	0.4708	23.9239	0.5504	23.9239	0.3785	23.7270	0.3893	23.8325	0.3874	23.8325	0.3874	23.8344	0.3874
p_3	18.7620	0.1311	18.8758	0.1769	18.8758	0.1216	18.8716	0.1252	18.8804	0.1242	18.8804	0.1242	18.8810	0.1242
p_4	15.5820	0.5964	16.0031	0.7511	16.0031	0.5165	16.0858	0.5219	16.0884	0.5180	16.0884	0.5180	16.0860	0.5180
p_5	13.3280	0.1362	13.1985	0.1662	13.1985	0.1143	13.2242	0.1189	13.2155	0.1173	13.2155	0.1173	13.2162	0.1173
p_6	18.6618	0.2500	18.4720	0.2282	18.4720	0.1571	18.4876	0.1615	18.5022	0.1591	18.5022	0.1591	18.5021	0.1591
p_7	10.3969	0.1455	10.4638	0.1685	10.4638	0.1159	10.4885	0.1184	10.4776	0.1177	10.4776	0.1177	10.4770	0.1177
	No Censoring		70 % Censoring Level											
	Observed Prices		Exculdign Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
p_i	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	45.2598	0.6193	45.2598	0.1938	45.3959	0.2255	45.3696	0.2156	45.3696	0.2156	45.3730	0.2156
p_2	23.7807	0.4708	23.4655	0.8953	23.4655	0.2794	23.9321	0.3333	23.6935	0.3108	23.6935	0.3108	23.6877	0.3107
p_3	18.7620	0.1311	18.5115	0.1558	18.5115	0.0487	18.5492	0.0568	18.4960	0.0547	18.4960	0.0547	18.4973	0.0547
p_4	15.5820	0.5964	14.6550	0.9537	14.6550	0.2977	14.5172	0.3249	14.5298	0.3079	14.5298	0.3079	14.5285	0.3079
p_5	13.3280	0.1362	13.6131	0.2372	13.6131	0.0740	13.6248	0.0844	13.6234	0.0834	13.6234	0.0834	13.6229	0.0834
p_6	18.6618	0.2500	19.0796	0.6189	19.0796	0.1937	18.8062	0.2198	19.0082	0.2119	19.0082	0.2119	19.0097	0.2119
p_7	10.3969	0.1455	10.2498	0.2817	10.2498	0.0879	10.1045	0.0972	10.2020	0.0926	10.2020	0.0926	10.2015	0.0926

Note: p_i, i = 1, …, 7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits. Average exchange rate in 2008 is US $1 = 11.14 Pesos (Banco de México).
Source: ENIGH 2008 Database, computed by author.

Table 3. Root Mean Square Error (RMSE) and Root Mean Square Percent Error (RMSPE) for Imputed Prices

	30% Censoring									
	CM		CW		SR		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
p_1	15.9498	0.5609	15.0249	0.5277	15.1083	0.5325	15.1083	0.5325	15.1139	0.5328
p_2	23.6157	0.9100	22.4628	0.8696	22.4946	0.8713	22.4946	0.8713	22.5092	0.8724
p_3	4.6705	0.5376	4.4238	0.5624	4.4348	0.5711	4.4348	0.5711	4.4406	0.5716
p_4	22.1532	0.8809	21.8287	0.9245	22.0666	0.9111	22.0666	0.9111	22.0679	0.9113
p_5	6.0702	0.5903	5.7229	0.5693	5.8029	0.5502	5.8029	0.5502	5.8044	0.5520
p_6	9.4277	0.7907	9.2105	0.6643	9.2567	0.6841	9.2567	0.6841	9.2574	0.6825
p_7	6.2683	0.7147	6.2678	0.7862	6.2593	0.7504	6.2593	0.7504	6.2635	0.7500
Overall	38.5966	1.9215	37.1921	1.8945	37.4087	1.8796	37.4087	1.8796	37.4223	1.8802
	70% Censoring									
	CM		CW		SR		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
p_1	15.4196	0.5142	15.2015	0.5040	15.1526	0.5082	15.1525	0.5082	15.1572	0.5083
p_2	22.6790	0.9412	21.8412	0.9802	21.7891	0.9366	21.7891	0.9366	21.7817	0.9365
p_3	9.1615	0.6595	8.9020	0.6733	8.9764	0.6827	8.9764	0.6827	8.9763	0.6818
p_4	29.3571	0.9543	29.4960	1.0555	29.5222	1.0570	29.5222	1.0570	29.5311	1.0588
p_5	6.5642	0.5079	6.3488	0.5125	6.4298	0.5079	6.4298	0.5079	6.4293	0.5077
p_6	9.8939	0.8132	10.2613	0.6832	10.3302	0.7581	10.3301	0.7581	10.3298	0.7570
p_7	9.2151	0.7564	9.1513	0.7763	9.1047	0.7508	9.1047	0.7508	9.1035	0.7508
Overall	43.8608	1.9968	43.4365	2.0284	43.4447	2.0285	43.4447	2.0285	43.4483	2.0287

The ECO approach (i.e., the columns titled Excluding Cen. Obs.) reports the mean and standard error, for 2,500 households when there is a 30% censoring level and for 1,072 households when there is a 70% censoring level. Since this strategy discards incompletely recorded units and focuses only on the completely recorded units, this strategy is sometimes referred to as a complete case analysis (Rubin 1996, p. 474; & Little and Rubin 2002, p. 41). However, excluding observations "can lead to serious biases… and it is not very efficient, especially when drawing inferences for subpopulations" (Little and Rubin 2002, p. 19). The last five approaches first impute the censored observations and then report the mean and standard error of the means for the imputed datasets.

At the 30% censoring level, these approaches result in mean values with small variability but standard errors with relatively larger variability. For instance, compared to the dataset with no missing price observations (i.e., the No Censoring column), the mean prices from the different methods ranged from being 1.02% lower (i.e., the legumes mean price estimate from the ECO approach or from the CM method) to 3.25% higher (i.e., the tubers mean price estimate from the SR method). On the other hand, the standard errors of the means ranged from being 37.16% lower (i.e., the standard error estimate of the legumes mean price from the CM method) to 34.94% higher (i.e., the standard error estimate of the eggs mean price from the ECO approach).

At the 70% censoring level, variability increases in both means and standard error of means. Compared to the dataset with no missing price observations, the mean prices from the different approaches ranges from being 6.83% lower (i.e., the tubers mean price estimate from CW method) to 2.24% higher (i.e., the legumes mean price estimate from the ECO approach or from the CM method). In addition, the variability in the standard errors of the means is larger at a 70% censoring level than at a 30% censoring level. For instance, the

estimate of the standard error of the eggs mean price obtained from CM method is 62.84% lower than the same estimate obtained from the dataset with no missing price observations. Likewise, the standard error of the legumes mean price obtained from ECO approach is 147.59% higher than the same standard error obtained from the dataset with no missing observations.

A simple comparison of the mean prices obtained from the dataset with no censored prices with the mean prices obtained from the various imputation approaches is inappropriate because positive errors would cancel out with negative errors. Hence, to appropriately evaluate which method generated the best imputations, the root mean square error (RMSE) and the root mean square percent error (RMSPE) were computed.[9] Table 3 reveals that at the 30% censoring level, the EM and the SR methods generated the best estimates (RMSPE = 1.8696) while the worst estimates were generated by the CM method (RMSPE = 1.9215). At the 70% censoring level, the CM method (RMSPE = 1.9968) generated the best estimates while the worst estimates were generated by the MCMC method (RMSPE = 2.0287). Notice that the imputation method that provides the best estimates for each price varies across prices when considering the RMSPE disaggregated as opposed to considering the overall measure.

Table 4 reports the parameter estimates from full AIDS models, equations (4) and (5), estimated under various approaches to price imputation for a 30% censoring level. From a total of 41 parameters estimated, at least 32 are statistically different from zero at the 0.05 significance level for each approach. Compared to the parameter estimates obtained from the dataset with no censored prices, the parameter estimates from the different approaches are on average 31% higher or lower. The difference ranged from being 615.72% lower (i.e., $\hat{\gamma}_{27}$ from the CW method) to 172.65% higher (i.e., $\hat{\gamma}_{35}$ from the ECO approach). These differences are remarkably higher under a 70% censoring level.[10]

Tables 5, 7, and 9 report estimates for the Marshallian own-price elasticities, the Hicksian own-price elasticities, and the expenditure elasticities respectively. Differences are also observed between the different censoring approaches and the elasticity estimates obtained from the dataset with no censored observations. Compared to the no-cenrored Marshallian own-price elasticity estimates, the elasticity estimates from the different approaches are on average 6.32% higher or lower (Table 5). Compared to the no-censored Hicksian own-price elasticity estimates, the Hicksian elasticity estimates from the different approaches are on average 7.25% higher or lower (Table 7). Similarly, compared to the no-censored expenditure elasticity estimates, the elasticity estimates from the different approaches are on average 3.03% higher or lower (Table 9). These elasticity estimates range from 48.22% lower ($\hat{e}_{44}$, EM method, and 70% censoring level) to 10.12% higher ($\hat{e}_{66}$, ECO approach, and 70% censoring level), from 50.54% lower ($\hat{e}^c_{44}$, EM method, and 70% censoring level) to 10.57% higher ($\hat{e}^c_{66}$, ECO approach, and 70% censoring level), and from 10.21% higher ($\hat{e}_6$, CM method, and 70% censoring level) to 12% higher ($\hat{e}_2$, CM method, and 70% censoring level) respectively. Consistent with the results from the AIDS parameter estimates, the ECO approach provided the closest estimates to the no-censored elasticity estimates.

Interestingly, even when there was small variability in the imputed mean prices (Table 2), considerable larger variability was found in the ultimately desired elasticity measures (Tables 5, 7, and 9). This suggests that setting aside a portion of the dataset with non-missing observations for validation purposes may provide insight into choosing the most appropriate imputation method and avoiding or reducing selection bias.

Table 4. AIDS Parameter Estimates Under 0% and 30% Censoring Levels

			30% Censoring									
	No Censoring		Excluding Cen. Obs.		Cell Mean		Cox and Wohlgenant		EM Algorithm		MCMC Algorithm	
Par.	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err
γ_{11}	0.0269 ***	0.0062	0.0283 ***	0.0075	0.0324 ***	0.0074	0.0267 ***	0.0071	0.0303 ***	0.0073	0.0305 ***	0.0073
γ_{12}	0.0203 ***	0.0031	0.0203 ***	0.0037	0.0146 ***	0.0038	0.0219 ***	0.0036	0.0228 ***	0.0037	0.0227 ***	0.0037
γ_{13}	-0.0261 ***	0.0021	-0.0292 ***	0.0025	-0.0292 ***	0.0025	-0.0298 ***	0.0024	-0.0310 ***	0.0024	-0.0310 ***	0.0024
γ_{14}	-0.0037 ***	0.0012	-0.0047 ***	0.0014	-0.0046 ***	0.0016	-0.0059 ***	0.0016	-0.0064 ***	0.0016	-0.0064 ***	0.0016
γ_{15}	-0.0059 **	0.0032	-0.0012	0.0039	-0.0001	0.0038	0.0016	0.0037	0.0003	0.0038	0.0003	0.0038
γ_{16}	-0.0148 ***	0.0022	-0.0172 ***	0.0027	-0.0169 ***	0.0026	-0.0162 ***	0.0026	-0.0182 ***	0.0026	-0.0182 ***	0.0026
γ_{17}	0.0033 *	0.0025	0.0038 *	0.0030	0.0038	0.0030	0.0016	0.0030	0.0022	0.0030	0.0022	0.0030
γ_{22}	-0.0199 ***	0.0029	-0.0238 ***	0.0034	-0.0160 ***	0.0038	-0.0214 ***	0.0035	-0.0225 ***	0.0035	-0.0224 ***	0.0035
γ_{23}	-0.0037 ***	0.0012	-0.0034 ***	0.0014	-0.0041 ***	0.0014	-0.0046 ***	0.0013	-0.0041 ***	0.0013	-0.0041 ***	0.0013
γ_{24}	-0.0019 ***	0.0007	-0.0017 ***	0.0008	-0.0003	0.0009	-0.0013 *	0.0009	-0.0016 **	0.0009	-0.0017 **	0.0009
γ_{25}	0.0082 ***	0.0019	0.0104 ***	0.0022	0.0085 ***	0.0023	0.0091 ***	0.0022	0.0089 ***	0.0022	0.0089 ***	0.0022
γ_{26}	-0.0031 ***	0.0012	-0.0020 *	0.0015	-0.0026 **	0.0014	-0.0030 ***	0.0014	-0.0029 ***	0.0014	-0.0030 ***	0.0014
γ_{27}	0.0002	0.0015	0.0003	0.0017	0.0000	0.0018	-0.0008	0.0018	-0.0005	0.0018	-0.0005	0.0018
γ_{33}	0.0264 ***	0.0023	0.0248 ***	0.0027	0.0260 ***	0.0026	0.0278 ***	0.0026	0.0272 ***	0.0026	0.0272 ***	0.0026
γ_{34}	0.0039 ***	0.0009	0.0042 ***	0.0011	0.0024 ***	0.0011	0.0028 ***	0.0011	0.0028 ***	0.0011	0.0029 ***	0.0011
γ_{35}	0.0010	0.0019	0.0017	0.0022	0.0027	0.0022	0.0023	0.0022	0.0027	0.0022	0.0027	0.0022
γ_{36}	0.0027 **	0.0015	0.0045 ***	0.0018	0.0047 ***	0.0017	0.0037 ***	0.0017	0.0047 ***	0.0017	0.0047 ***	0.0017
γ_{37}	-0.0042 ***	0.0014	-0.0026 *	0.0016	-0.0027 **	0.0016	-0.0022 *	0.0016	-0.0024 *	0.0016	-0.0024 *	0.0016
γ_{44}	0.0072 ***	0.0007	0.0067 ***	0.0008	0.0087 ***	0.0009	0.0106 ***	0.0009	0.0101 ***	0.0009	0.0101 ***	0.0009
γ_{45}	-0.0032 ***	0.0011	-0.0025 ***	0.0012	-0.0041 ***	0.0013	-0.0040 ***	0.0013	-0.0036 ***	0.0013	-0.0037 ***	0.0013
γ_{46}	-0.0012 *	0.0008	-0.0008	0.0010	-0.0013	0.0010	-0.0019 **	0.0010	-0.0008	0.0010	-0.0008	0.0010
γ_{47}	-0.0010 *	0.0008	-0.0011	0.0009	-0.0008	0.0010	-0.0004	0.0010	-0.0005	0.0010	-0.0005	0.0010
γ_{55}	0.0181 ***	0.0033	0.0129 ***	0.0039	0.0140 ***	0.0039	0.0122 ***	0.0039	0.0129 ***	0.0039	0.0129 ***	0.0039
γ_{56}	-0.0058 ***	0.0018	-0.0078 ***	0.0021	-0.0076 ***	0.0021	-0.0074 ***	0.0020	-0.0070 ***	0.0021	-0.0069 ***	0.0021

Table 4. ***Continued***

			30% Censoring									
	No Censoring		Excluding Cen. Obs.		Cell Mean		Cox and Wohlgenant		EM Algorithm		MCMC Algorithm	
Par.	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err	Estimate	Approx Std Err
γ_{57}	-0.0125 ***	0.0018	-0.0134 ***	0.0022	-0.0134 ***	0.0022	-0.0138 ***	0.0022	-0.0142 ***	0.0022	-0.0141 ***	0.0022
γ_{66}	0.0250 ***	0.0019	0.0267 ***	0.0023	0.0270 ***	0.0022	0.0276 ***	0.0022	0.0268 ***	0.0022	0.0268 ***	0.0022
γ_{67}	-0.0028 ***	0.0013	-0.0034 ***	0.0016	-0.0033 ***	0.0016	-0.0029 **	0.0016	-0.0026 **	0.0016	-0.0026 **	0.0016
γ_{77}	0.0170 ***	0.0020	0.0163 ***	0.0023	0.0163 ***	0.0024	0.0185 ***	0.0024	0.0179 ***	0.0024	0.0179 ***	0.0024
α_1	0.2673 ***	0.0097	0.2689 ***	0.0116	0.2754 ***	0.0109	0.2676 ***	0.0104	0.2662 ***	0.0107	0.2661 ***	0.0107
α_2	0.1377 ***	0.0064	0.1372 ***	0.0076	0.1247 ***	0.0075	0.1326 ***	0.0071	0.1292 ***	0.0072	0.1292 ***	0.0072
α_3	0.1506 ***	0.0033	0.1545 ***	0.0040	0.1536 ***	0.0036	0.1545 ***	0.0035	0.1557 ***	0.0035	0.1557 ***	0.0036
α_4	0.0641 ***	0.0019	0.0648 ***	0.0023	0.0687 ***	0.0024	0.0701 ***	0.0024	0.0714 ***	0.0024	0.0714 ***	0.0024
α_5	0.1896 ***	0.0055	0.1823 ***	0.0065	0.1835 ***	0.0061	0.1807 ***	0.0059	0.1814 ***	0.0060	0.1815 ***	0.0060
α_6	0.1301 ***	0.0035	0.1345 ***	0.0043	0.1321 ***	0.0039	0.1307 ***	0.0038	0.1333 ***	0.0038	0.1333 ***	0.0038
α_7	0.0606 ***	0.0044	0.0579 ***	0.0051	0.0620 ***	0.0050	0.0638 ***	0.0049	0.0628 ***	0.0050	0.0628 ***	0.0050
β_1	0.0447 ***	0.0046	0.0452 ***	0.0055	0.0324 ***	0.0048	0.0406 ***	0.0047	0.0395 ***	0.0047	0.0394 ***	0.0047
β_2	0.0312 ***	0.0037	0.0312 ***	0.0044	0.0523 ***	0.0042	0.0386 ***	0.0040	0.0408 ***	0.0040	0.0409 ***	0.0040
β_3	-0.0345 ***	0.0015	-0.0341 ***	0.0018	-0.0352 ***	0.0015	-0.0349 ***	0.0015	-0.0349 ***	0.0015	-0.0350 ***	0.0015
β_4	-0.0133 ***	0.0009	-0.0133 ***	0.0011	-0.0141 ***	0.0010	-0.0132 ***	0.0010	-0.0137 ***	0.0010	-0.0137 ***	0.0010
β_5	-0.0133 ***	0.0026	-0.0133 ***	0.0031	-0.0173 ***	0.0026	-0.0160 ***	0.0026	-0.0162 ***	0.0026	-0.0162 ***	0.0026
β_6	-0.0335 ***	0.0016	-0.0350 ***	0.0020	-0.0348 ***	0.0016	-0.0339 ***	0.0016	-0.0339 ***	0.0016	-0.0339 ***	0.0016
	R-sqr		R-sqr		R-sqr		R-sqr		R-sqr		R-sqr	
w_1	0.0384		0.0416		0.0221		0.0335		0.0339		0.0339	
w_2	0.0381		0.0451		0.0499		0.0416		0.0446		0.0445	
w_3	0.1780		0.1822		0.1636		0.1837		0.1870		0.1871	
w_4	0.0872		0.0902		0.0760		0.0785		0.0778		0.0781	
w_5	0.0265		0.0234		0.0245		0.0259		0.0245		0.0247	
w_6	0.1430		0.1516		0.1489		0.1438		0.1434		0.1435	

Note: Significance levels of 0.05, 0.10, and 0.20 are indicated by triple asterisks (***), double asterisks (**), and an asterisk (*) respectively.

Tables 6, 8, and 10 reveals that at the 30% censoring level, the CM method generated the best estimates for the Marshallian (RMSPE = 396.0614) and the Hicksian price elasticities (RMSPE = 1767.3065) while the worst estimates were generated by the CW method (RMSPE = 971.9238) and the MCMC method (RMSPE = 2236.7423) respectively. In case of the expenditure elastic estimates (Table 10), the best estimates were generated by the SR and EM methods (31.6429) while the worst estimates were generated by the CM method (RMSPE = 39.3499).

The methods that provided the best estimates for the elasticities at the 30% censoring level are not necessarily the same at the 70% censoring level. In the latter, the CM method (RMSPE = 994.2068) generated the best estimates for the Marshallian price elasticities while the worst estimates were generated by the MCMC method (RMSPE = 2212.905). In the case of the Hicksian price elasticities at 70% censoring level, the CM method generated the best estimates (RMSPE = 1564.8632) while the EM method generated the worst estimates (RMSPE = 4890.0503). Last, the CM and CW methods provided the best estimates (RMSPE = 68.8568) for the expenditure elasticities while the MCMC method provided the worst estimates (RMSPE = 76.2207). Consistent with the results for the price analysis (Table 3), the results for the elasticity analysis (Tables 6, 8, and 10) were mixed when analyzing specific elasticity estimates (as opposed to overall estimates) at either 30% censoring level or 70% censoring level.

Interestingly, at the 30% censoring level, the EM method generated the best mean price estimates (RMSPE = 1.8796, Table 3), but it was the CM method which generated the best elasticity estimates overall (RMSPE = 1877.5699). This suggests that the imputation method that would be chosen should be selected based on an analysis from the ultimately desired measures. On the other hand, at the 70% censoring level, the CM method generated the best mean price estimates (RMSPE = 1.9968, Table 3) as well as the best overall elasticity estimates (RMSPE = 1852.7305). This suggests that under large level of censoring, simple techniques such as the CM method may perform satisfactory or even provide better estimates than sophisticated techniques such as the EM and MCMC methods.

Table 5. Marshallian Own-Price Elasticity Estimates Under 0%, 30%, and 70% Censoring Levels

	No	30% Censoring					70% Censoring				
	Censoring	ECO	CM	CW	EM	MCMC	ECO	CM	CW	EM	MCMC
e_{11}	-0.9300	-0.9267	-0.9120	-0.9288	-0.9189	-0.9184	-0.9412	-0.9035	-0.9472	-0.8995	-0.8999
e_{22}	-1.1009	-1.1216	-1.0772	-1.1050	-1.1102	-1.1097	-1.0532	-0.9946	-1.0067	-1.0437	-1.0444
e_{33}	-0.6560	-0.6783	-0.6487	-0.6292	-0.6360	-0.6353	-0.5835	-0.4990	-0.4441	-0.4615	-0.4624
e_{44}	-0.8196	-0.8312	-0.7960	-0.7527	-0.7661	-0.7648	-0.7816	-0.7479	-0.4537	-0.4244	-0.4240
e_{55}	-0.8924	-0.9227	-0.9138	-0.9256	-0.9211	-0.9211	-0.8313	-0.8230	-0.8574	-0.8379	-0.8377
e_{66}	-0.6477	-0.6289	-0.6098	-0.6043	-0.6172	-0.6170	-0.7132	-0.6976	-0.5750	-0.5815	-0.5810
e_{77}	-0.7998	-0.8063	-0.8089	-0.7862	-0.7917	-0.7920	-0.7797	-0.7851	-0.6921	-0.7873	-0.7873

Note: e_{ij}, $i = j = 1, 2, \ldots, 7$, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Table 6. Root Mean Square Error (RMSE) and Root Mean Square Percent Error (RMSPE) for After-Imputation Marshallian Own-Price Elasticity Estimates

	30% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_{11}	0.1113	0.3157	0.0736	0.1418	0.0911	0.2216	0.0928	0.2293
e_{22}	0.2086	0.1450	0.2392	0.1678	0.2663	0.1866	0.2679	0.1865
e_{33}	0.2644	8.3271	0.2972	13.0268	0.2854	11.6893	0.2869	11.8010
e_{44}	0.6481	5.5709	0.7454	4.8520	0.8006	5.6926	0.8078	5.7448
e_{55}	0.2019	3.1883	0.1551	2.3391	0.1671	2.6428	0.1664	2.6433
e_{66}	0.8954	20.7972	1.0144	25.9407	0.9566	20.6203	0.9566	20.6394
e_{77}	0.3860	8.4851	0.4097	8.9000	0.4028	7.6417	0.4018	7.2458
All e_{ij}, i = j	1.2399	24.8028	1.3884	30.8365	1.3809	25.6848	1.3854	25.6482
All e_{ij}, i, j = 1, 2, …, 7	267.0056	396.0614	267.0065	971.9238	267.0070	694.2498	267.0070	683.6587
	70% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_{11}	0.1070	0.3619	0.1070	0.3619	0.1138	0.4675	0.1124	0.4585
e_{22}	0.3186	0.2096	0.3186	0.2096	0.2513	0.1339	0.2447	0.1328
e_{33}	0.7800	21.5138	0.7800	21.5138	0.8629	24.6745	0.8663	25.9810
e_{44}	0.9381	9.8106	0.9381	9.8106	2.5436	31.5370	2.5802	32.1871
e_{55}	0.5910	12.7837	0.5910	12.7837	0.4668	10.1846	0.4746	10.4610
e_{66}	0.4019	38.0428	0.4019	38.0428	0.5798	26.6884	0.5920	29.7259
e_{77}	1.2067	42.6318	1.2067	42.6318	1.1940	38.0270	1.2246	37.1436
All e_{ij}, i = j	1.8890	63.1460	1.8890	63.1460	3.0447	62.1747	3.0913	63.9058
All e_{ij}, i, j = 1, 2, …, 7	267.0114	994.2068	267.0114	994.2068	267.0317	2167.6406	267.0326	2212.9051

Note: e_{ij}, $i = j$ = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Table 7. Hicksian Own-Price Elasticity Estimates Under 0%, 30%, and 70% Censoring Levels

	No	30% Censoring					70% Censoring				
	Censoring	ECO	CM	CW	EM	MCMC	ECO	CM	CW	EM	MCMC
e^c_{11}	-0.5254	-0.5196	-0.5257	-0.5335	-0.5240	-0.5235	-0.5425	-0.5478	-0.5717	-0.5261	-0.5265
e^c_{22}	-0.8705	-0.8930	-0.8134	-0.8602	-0.8627	-0.8621	-0.8186	-0.6961	-0.7331	-0.7702	-0.7709
e^c_{33}	-0.6136	-0.6351	-0.6096	-0.5891	-0.5963	-0.5956	-0.5426	-0.4617	-0.4045	-0.4228	-0.4236
e^c_{44}	-0.7932	-0.8049	-0.7676	-0.7232	-0.7367	-0.7354	-0.7548	-0.7183	-0.4230	-0.3924	-0.3919
e^c_{55}	-0.7371	-0.7684	-0.7676	-0.7766	-0.7732	-0.7733	-0.6741	-0.6756	-0.7078	-0.6891	-0.6889
e^c_{66}	-0.6103	-0.5918	-0.5753	-0.5683	-0.5811	-0.5809	-0.6748	-0.6651	-0.5424	-0.5477	-0.5472
e^c_{77}	-0.6964	-0.7028	-0.7072	-0.6810	-0.6873	-0.6875	-0.6764	-0.6861	-0.5938	-0.6876	-0.6876

Note: e^c_{ij}, $i = j$ = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Table 8. Root Mean Square Error (RMSE) and Root Mean Square Percent Error (RMSPE) for After-Imputation Hicksian Own-Price Elasticity Estimates

	30% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e^c_{11}	0.1074	0.4619	0.0839	0.2068	0.0935	0.3238	0.0948	0.3361
e^c_{22}	0.2699	0.2670	0.2767	0.2540	0.3000	0.2706	0.3016	0.2706
e^c_{33}	0.2547	37.2035	0.2883	46.5190	0.2763	39.2094	0.2777	40.2197
e^c_{44}	0.6437	3.5044	0.7419	3.6957	0.7969	3.9035	0.8041	3.9611
e^c_{55}	0.2031	2.6494	0.1591	1.9659	0.1700	2.2310	0.1693	2.2282
e^c_{66}	0.8900	13.4721	1.0104	21.1317	0.9523	34.9030	0.9522	25.1823
e^c_{77}	0.3783	383.0000	0.4021	75.2798	0.3948	251.0664	0.3939	258.2006
All e^c_{ij}, i = j	1.2410	385.0638	1.3874	91.0782	1.3791	256.5352	1.3836	262.5646
All e^c_{ij}, i, j = 1, 2, …, 7	267.0100	1767.3065	267.0103	1884.7531	267.0117	2204.6346	267.0118	2236.7423
	70% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e^c_{11}	0.1074	0.4619	0.0839	0.2068	0.0935	0.3238	0.0948	0.3361
e^c_{22}	0.2699	0.2670	0.2767	0.2540	0.3000	0.2706	0.3016	0.2706
e^c_{33}	0.2547	37.2035	0.2883	46.5190	0.2763	39.2094	0.2777	40.2197
e^c_{44}	0.6437	3.5044	0.7419	3.6957	0.7969	3.9035	0.8041	3.9611
e^c_{55}	0.2031	2.6494	0.1591	1.9659	0.1700	2.2310	0.1693	2.2282
e^c_{66}	0.8900	13.4721	1.0104	21.1317	0.9523	34.9030	0.9522	25.1823
e^c_{77}	0.3783	383.0000	0.4021	75.2798	0.3948	251.0664	0.3939	258.2006
All e^c_{ij}, i = j	1.2410	385.0638	1.3874	91.0782	1.3791	256.5352	1.3836	262.5646
All e^c_{ij}, i, j = 1, 2, …, 7	267.0100	1767.3065	267.0103	1884.7531	267.0117	2204.6346	267.0118	2236.7423

Note: e^c_{ij}, $i = j$ = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Table 9. Expenditure Elasticity Estimates Under 0%, 30%, and 70% Censoring Levels

	No	30% Censoring					70% Censoring				
	Censoring	ECO	CM	CW	EM	MCMC	ECO	CM	CW	EM	MCMC
e_1	1.1241	1.1249	1.0914	1.1144	1.1111	1.1110	1.1224	1.0548	1.0932	1.0890	1.0890
e_2	1.1568	1.1583	1.2472	1.1872	1.1974	1.1977	1.1546	1.2956	1.2403	1.2428	1.2428
e_3	0.5517	0.5591	0.5261	0.5347	0.5323	0.5322	0.5364	0.5280	0.5433	0.5380	0.5380
e_4	0.6646	0.6632	0.6688	0.6911	0.6825	0.6823	0.6693	0.6699	0.6994	0.7091	0.7092
e_5	0.9214	0.9207	0.8941	0.9031	0.9015	0.9013	0.9186	0.8926	0.8935	0.8972	0.8972
e_6	0.5273	0.5146	0.4985	0.5156	0.5156	0.5155	0.5631	0.4735	0.4798	0.4884	0.4885
e_7	1.2211	1.2285	1.1960	1.2165	1.2133	1.2137	1.2032	1.1814	1.1755	1.1767	1.1766

Note: e_i, i = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Table 10. Root Mean Square Error (RMSE) and Root Mean Square Percent Error (RMSPE) for After-Imputation Expenditure Elasticity Estimates

	30% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_1	0.1283	0.0873	0.1111	0.0731	0.1082	0.0717	0.1086	0.0718
e_2	0.8784	0.5167	0.4656	0.2852	0.5254	0.3203	0.5329	0.3225
e_3	0.3633	3.9599	0.3624	3.6817	0.3634	3.6695	0.3641	3.7370
e_4	1.0277	32.5670	0.8455	31.4511	1.0237	27.9184	1.0262	28.5841
e_5	0.2886	1.5184	0.1988	1.0176	0.2239	1.1391	0.2235	1.1350
e_6	1.1353	21.6681	1.2125	15.3382	1.1935	14.3849	1.1930	14.5627
e_7	0.4210	0.2067	0.4118	0.2048	0.4154	0.2072	0.4154	0.2072
All e_i, i = 1, 2, …, 7	1.8777	39.3499	1.6597	35.2016	1.7649	31.6429	1.7683	32.3191
	70% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_1	0.1483	0.1018	CW	0.1018	0.1064	0.0700	0.1074	0.0707
e_2	2.7710	1.0150	0.148317562	1.0150	1.5162	0.6035	1.5337	0.6093
e_3	0.6588	19.2315	2.77101464	19.2315	0.6289	21.6351	0.6221	20.9504
e_4	1.1759	19.5651	0.658753257	19.5651	1.1047	15.9901	1.1107	16.2046
e_5	0.3454	3.2614	1.175925238	3.2614	0.2855	2.3004	0.2866	2.3024
e_6	0.6699	63.0620	0.345369372	63.0620	0.6878	70.9856	0.6851	71.4305
e_7	0.9989	0.3374	0.669907679	0.3374	0.9735	0.3114	0.9990	0.3117
All e_i, i = 1, 2, …, 7	3.3291	68.8568	0.998908267	68.8568	2.3299	75.9505	2.3525	76.2207

Note: e_i, i = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

4. Concluding Remarks

Several studies often use simple techniques to account for censored prices in models where prices are independent variables. These simple techniques either omit the missing prices or use price imputation approaches such as deductive imputation, cell mean imputation, hot-deck imputation, cold-deck imputation, and regression imputation. This study compares and contrast several imputation approaches under two levels of censoring by following a multiple imputation methodology (e.g., analyzes the ultimately desired measures). The imputation approaches analyzed are: excluding censoring observations (ECO), cell mean imputation (CM), Cox and Wolhgenant's (1986) first-order missing price procedure (CW), simple regression imputation (SR), the EM algorithm, and the MCMC algorithm.

Differences in price variability before and after price imputation are quantified, the performance of each method under different levels of missing data are evaluated, and elasticity estimates for several important protein sources (meat, dairy, eggs, tubers, vegetables, legumes, and fruits) in the Mexican diet are estimated under the various imputation procedures. These elasticity estimates are relatively recent and contribute to a better understanding of the Mexican demand for protein sources. In addition, these estimates can be used to analyze current and/or future trends in protein consumption.

The study's findings reveal that even when there is small variability among the imputer's model estimates, there may be larger variability among the analyst's model estimates. Therefore, it is recommended that the imputation method that is selected is based on an analysis of the ultimately desired measures. These measures may suffer from selection bias if an imputation method is inappropriately chosen. In addition, evaluating the imputation methods using a simple comparison of the mean prices or elasticities is inappropriate because calculating means cancels out positive errors with negative errors; therefore, computing the RMSE and the RMSPE is recommended. This is critical because the method that provides the best estimate is not necessarily the same when evaluating the estimates using a simple comparison and when evaluating the estimates using the RMSPE. Unfortunately, the RMSE and RMSPE cannot be computed for the ECO approach and the ECO approach may also be unfeasible when the censoring occurs in each price at different instances (i.e., the complete-case data may have few observations).

This study also found that the imputation method or approach that provides the best estimates varies across the imputed variables (i.e., p_i, $i = 1, 2, \ldots, 7$) and across the ultimately desired measures (i.e., e_{ij}, e_i, e_{ij}^c, $i, j = 1, 2, \ldots, 7$). Furthermore, results are sensitive to the censoring levels. That is, the method that generates the best estimates at the 30% censoring level is not necessarily the same method that generates the best estimate at the 70% censoring level. In particular, at high levels of censoring, a simple method such as the CM may perform satisfactory or even better than sophisticated methods.

Provided that the results are sensitive to the imputation approach chosen, it is recommended that a portion of the dataset is set aside for validation purposes and that the imputation method that would be chosen be selected based on an analysis from the ultimately desired measures (e.g., following a multiple imputation methodology).

Further research may be conducted with datasets where prices are not missing at random or where prices are not censored at the same instances (e.g., with datasets that have many missing data patterns). It should be noted that in this study the EM algorithm was observed to provide similar results to the SR method because only one missing data pattern was considered. This is because the EM algorithm uses maximum likelihood estimation. It was also observed that the SR method performed similar to the CW method when a simple regression was estimated for each of the means considered in the CW method. Finally, the estimates from the cell mean method may improve if more cells are used in the analysis. The more specific the classes are, the more likely the research is to obtain an estimate that is closer to the true value.

References

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications.* Cambridge: Cambridge University Press.

Cox, T. L., & Wohlgenant, M. K. (1986). Prices and Quality Effects in Cross-Sectional Demand Analysis. *American Journal of Agricultural Economics*, 68, 908-919.

Deaton, A., & Muellbauer, J. (1980). An Almost Ideal Demand System. *The American Economic Review*, 70(3), 312-325.

Dong, D., & Gould, B. W. (2000). Quality Versus Quantity in Mexican Poultry and Pork Purchases. *Agribusiness*, 16, 333-355.

Dong, D., Gould, B. W., & Kaiser, H. M. (2004). Food Demand in Mexico: An Application of the Amemiya-Tobin Approach to the Estimation of a Censored Food System. *American Journal of Agricultural Economics*, 86, 1094-1107.

Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). (2008). Aguascalientes, México: Instituto Nacional de Estadística, Geografía e Informática (INEGI). Internet site: http://www.inegi.gob.mx.

Gould, B. W., Lee, Y., Dong, D., & Villarreal, H. J. (2002, July). *Household Size and Composition Impacts on Meat Demand in Mexico: A Censored Demand System Approach*. Paper presented at the American Agricultural Economics Association Annual Meeting, Long Beach, California.

Golan, A., Perloff, J. M., & Shen, E. Z. (2001). Estimating a Demand System with Nonegativity Constraints: Mexican Meat Demand. *The Review of Economics and Statistics*, 8, 541-550.

Heckman, J. J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, 42, 679-694.

Heien, D., Jarvis, L. S., & Perali, F. (1989). Food Consumption in Mexico: Demographic and Economic Effects. *Food Policy Journal*, 14, 167-179.

Little, R. J. A., & Rubin, D. B.. (2002). *Statistical Analysis with Missing Data*, 2nd Edition. New Jersey: John Wiley & Sons Inc.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* New York: Duxbury Press.

Lopez, J. A., Malaga, J. E., Chidmi, B., Belasco, E., & Surles, J. (2012). Mexican Meat Demand at the Table Cut Level: Estimating a Censored Demand System in a Complex Survey. *Journal of Food Distribution Research*, 43(2), 64-90.

Taylor, M., Phaneuf, D., & Piggott, N. (2008, June). *Does Food Safety Information Affect Consumers' Decision to Purchase Meat and Poultry? Evidence from U.S. Household Level Data*. Paper presented at the Western Agricultural Economics Association Annual Meeting, Big Sky, MT.

Pindyck, R. S. Rubinfeld, D. L. (1997). *Econometric Models and Economic Forecasts,* 4th Edition. Massachusetts: Irwin McGraw-Hill, Inc.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

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

Sabates, R., Gould, B. W., & Villarreal, H. (2001). Household Composition and Food Expenditures: A Cross-Country Comparison. *Food Policy*, 26, 571-586.

SAS Institute Inc. *SAS/STAT 9.1 User's Guide.* (2004). Gary, NC: SAS Institute Inc.

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

Stone, J. R. N. (1954). The Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand. *Economic Journal*, 64, 511-527.

Zheng, Z., & Henneberry, S. R. (2009). An Analysis of Food Demand in China: A Case Study of Urban Households in Jiangsu Province. *Review of Agricultural Economics*, 31, 873-893.

Footnotes

[1] When there is item non-response on both the dependent variables (e.g., quantities) and the independent variables (e.g., prices), researchers combine both of these techniques, first using price imputation models and subsequently using models such as the censored nonlinear quadratic almost ideal demand system (censored NQUAIDS), the censored QUAIS, Amemiya-Tobin approach extensions to demand systems estimations, double-hurdle models, etc.

[2] For example, using four strata and Mexico's 31 states plus the Federal District produces 128 different values for the missing values. Using two strata and 32 states/locations produces 64 different values.

[3] The initial estimates for the EM algorithm can be obtained from the non-censored observations.

[4] For a discussion of single versus multiple chains refer to Schafer (1997, pp. 137-138).

[5] In practice, each price usually has a different censoring level and a price could be censored at a different time than another price. When this is the case, the dataset that only contains the non-censored observations may have few observations. In addition, the dataset that contains the censored observations may have many missing data patterns, but not all possible

patterns may show up in the dataset. For example, with i variables, $p_1, p_2, \ldots, p_i$, up to 2^i groups of observations (possible missing patterns) can be formed.

[6] The CW method computed the mean for each Mexican state and the Federal District.

[7] The parameter estimates from the simple regression imputation approach (i.e., method 3) at a 70% censoring level and the parameter estimates from the first-order missing price procedure of Cox and Wohlgenant (1986) under both the 30% censoring level and the 70% censoring level are available upon request.

[8] The MLE of the means and variance-covariance matrix from the last iteration of the EM or MCMC algorithms under both the 30% censoring level and the 70% censoring level are available upon request.

[9] The root mean square imputation error and the root mean square percent error for price p_i are defined as $RMSE = \sqrt{\frac{1}{(H*l)}\sum_{h=1}^{H*l}\left(p_{ih}^{imputed} - p_{ih}^{actual}\right)^2}$ and $RMSPE = \sqrt{\frac{1}{(H*l)}\sum_{h=1}^{H*l}\left(\frac{p_{ih}^{imputed} - p_{ih}^{actual}}{p_{ih}^{actual}}\right)^2}$ respectively, where l equals 0.30 or 0.70 depending on the censoring level (see Pindyck and Rubinfeld 1997, pp. 384-386). Similar definitions are used for the Marshallian and Hicksian price elasticities as well as the expenditure elasticities.

[10] The AIDS parameter estimates under the various approaches to price imputation for a 70% censoring level are available upon request.