Missing Data In Spatial Regression

Frederick J. Boehmke
University of Iowa

Emily U. Schilling
Washington University

Jude C. Hays
University of Pittsburgh

July 15, 2015


1Corresponding author: frederick-boehmke@uiowa.edu or emily-schilling@uiowa.edu. We are grateful for funding provided by the University of Iowa Department of Political Science. Comments from Michael Kellermann, Walter Mebane, and Vera Troeger greatly appreciated.
Abstract

We discuss problems with missing data in the context of spatial regression. Even when data are MCAR or MAR based solely on exogenous factors, estimates from a spatial regression with listwise deletion can be biased, unlike for standard estimators with independent observations. Further, common solutions for missing data such as multiple imputation appear to exacerbate the problem rather than fix it. We propose a solution for addressing this problem by extending previous work including the EM approach developed in Hays, Schilling and Boehmke (2015). Our EM approach iterates between imputing the unobserved values and estimating the spatial regression model using these imputed values. We explore the performance of these alternatives in the face of varying amounts of missing data via Monte Carlo simulation and compare it to the performance of spatial regression with listwise deletion and a benchmark model with no missing data.
1 Introduction

Political science data are messy. Subsets of the data may be missing; survey respondents may not answer all of the questions, economic data may only be available for a subset of countries, and voting records are not always complete. Over time, political scientists have become more attuned to the consequences and solutions for missing data (King et al. 2001; Honaker and King 2010) and the use of multiple imputation has become the norm in filling out incomplete datasets. Multiple imputation (Rubin 2009) or maximum likelihood approaches (Allison 2001) work well for a wide variety of estimators and their properties are generally well understood, but they do not necessarily work equally well for all estimators. In this paper we focus on missing data in the context of spatial regression, which adds a number of complexities that change the consequences and the solutions needed to address missing data. As the use of spatial regression expands in political science this is a problem that may be confronted quite frequently and researchers should understand the consequences of the most common solution — listwise deletion — as well as various approaches for addressing missing data.

We discuss an EM method for estimating spatial regression models with missing data that accounts for the spatial interdependence while also generating accurate standard errors through the use of multiple imputation. As we show in our simulation, missing data can be especially problematic for spatial econometric models because the outcome for one observation depends on the outcomes for others. Listwise deletion does not just remove observations in the spatial context, rather it effectively eliminates part of the spatial lag for some observations. So while listwise deletion does not lead to bias in standard linear regression applications when data are missing completely at random and only under certain circumstances when they are missing at random, in the spatial regression context both can produce bias. Missing data thus undermines researchers’ ability to properly
estimate model parameters in the spatial context, including capturing interdependence, much more so than in standard applications.

Further, as our simulations show, standard approaches for imputing missing data (e.g. Rubin 2009; King et al. 2001) also do not work off the shelf in the context of spatially autocorrelated data. Typical multiple imputation approaches assume joint normality of the vector of observations on all variables and use this assumption to generate an estimate of the distribution of missing observations from which to sample. With spatially correlated data we do not have separate observations since the realization of the dependent variable for one observation depends on the realization of the dependent variable in other observations. Thus we can no longer assume that observations are independent and identically distributed. Ignoring this violation and applying multiple imputation as if observations are independent can, according to our simulation, lead to even more biased estimates than listwise deletion.

In light of shortcomings in existing methods and in order to deal with the issue of missing data in spatial regression, we adapt and improve Hays, Schilling and Boehmke’s (2015) imputation algorithm for right censored spatial duration data for the spatial regression context. The algorithm iterates between imputing the values for the unobserved outcomes and estimating the spatial regression model given these imputed values. In doing so, we build on LeSage and Pace’s (2004) approach for spatial regression with unobserved (rather than censored) values of the outcome variable. Specifically we follow Hays, Schilling and Boehmke (2015) by employing multiple rather than single mean imputation for missing cases. More importantly, we extend our procedure to account for missing values in both the dependent and independent variables. We then explore the performance of the estimator in the face of varying amounts of missing data and spatial interdependence via Monte Carlo. We compare these results to the estimates received from listwise deletion. Before moving on to our proposed approach, we first review general concepts.
2 Missing Data

The three types of missingness include missing completely at random (MCAR), missing at random (MAR), and nonignorable missingness (NI). Here we limit ourselves to the first two types of missingness (MCAR and MAR) since NI requires a substantially different approach and so we will concentrate on those two in our discussion.

First, let $D$ denote the data matrix. This matrix includes the dependent variable ($Y$) and all of the independent variables ($X$). With no missing data, $D$ is fully observed and standard practices can be used to analyze it. A fully observed data set is incredibly rare in political science making the use of standard methods potentially unsuitable. The missing elements of $D$ can be denoted by a matrix, $M$, that has the same dimensions as $D$ and indicates whether each element is missing or not ($m_{ij} = 1$ if $d_{ij}$ is observed and $m_{ij} = 0$ when $d_{ij}$ is missing). Thus when there is missingness, $D$ includes the observed ($D_{obs}$) and the missing portions ($D_{mis}$) of the data.

The distinction between MCAR and MAR rests on our ability to predict which values are missing. Under MCAR, the pattern of missing values cannot be predicted with information in $D$. To be more specific, $M$ is independent of $D$ or $P(M|D) = P(M)$. This type of missingness can occur through survey or researcher error that occurs with equal probability across observations, but tends not to fit most social science data. Missing at random (MAR) processes occur more frequently since missing values often result from choices made by participants or the units of study, leading the pattern of missingness to depend on observed characteristics of the observations. Thus the probability that an element is missing from $D$ may depend on $D_{obs}$ and at the same time be independent of $D_{mis}$. More

---

1We will follow the notation used by King et al. (2001).
formally, $M$ is independent of $D_{\text{mis}}$ or $P(M|D) = P(M|D_{\text{obs}})$. The relationship between the missing and observed data does not need to be causal, rather one looks to predict the missing variables through a (possibly extensive) collection of variables included in the imputation process (Rubin 2009; King et al. 2001).

When analyzing incomplete data, the inferences that can be drawn depend on the method being used as well as the missingness assumption. Regardless of the assumption, conclusions drawn from analyses using listwise deletion tend to be inefficient and may be biased under MAR holds (or NI). The inefficiency that comes with listwise deletion is dramatic even under those situations (MCAR) when there is no bias (King et al. 2001). This is why multiple imputation methods are preferred with incomplete data. The inferences are more efficient than listwise deletion and are not biased under MCAR or MAR (Rubin 2009).²

### 3 Missing Data in Spatial Regression Models

Introducing spatial autocorrelation complicates things since conditional independence no longer holds for spatially autocorrelated variables. This matters both for the consequences of listwise deletion and also for our ability to use standard multiple imputation techniques. To illustrate, start with the the data generating process for a spatial lag model, written in matrix notation as follows:

$$y = \rho Wy + X\beta + u$$

where $y$ represents the observed outcome and $W$ specifies the connections between observations. The parameter $\rho$ captures the degree of spatial dependence across observations.

²Under the assumption of NI, this is not always the case because multiple imputation can be biased.
From this structural form we can derive the reduced form model:

\[ y = (I - \rho W)^{-1}X\beta + (I - \rho W)^{-1}u, \]  
\[ = \Gamma X\beta + \Gamma u, \]  
\[ = \Gamma X\beta + v. \]  

where \( \Gamma = (I - \rho W)^{-1} \) and \( v = \Gamma u \). The reduced form equation shows how the outcome for each unit depends on both the observed and unobserved components of other units through the spatial weights matrix. Previous work shows that ignoring spatial dependence leads to biased coefficient estimates as in the linear regression context (Franzese and Hays 2007; LeSage and Pace 2010) when \( \rho \neq 0 \).

What happens when we have missing data? For now, consider the case in which we have valid data for all independent variables but are missing data for the dependent variable according to either MCAR or MAR. As noted earlier, in standard applications with i.i.d. data, employing listwise deletion under these assumptions would not cause bias, though it would not use all the information in the data set. In spatial regression, though, the included observations will typically depend on omitted observations through the spatial lag. Excluding these observations relegates their effects to the error term. Further, the spatial lag means that the omitted terms depend on the same independent variables included in the regression equation. If the independent variables have a spatial pattern of correlation that relates to the spatial pattern in \( W \), the error terms will then correlate with the independent variables, leading to bias. Any strategy for addressing missingness in spatial regression must eliminate this correlation.

Now consider the other leading contender for addressing missingness: how does spatial regression complicate the application of multiple imputation? Consider the multivariate normal approach to imputation as described in King et al. (2001). Because it as-
sumes that observations are i.i.d, the data generating process with one missing variable simplifies to a linear regression. In the case described above when just \( Y \) may be missing, this reduces to predicting missing values of \( Y \) with a linear combination of variables: 
\[
E[Y_{\text{mis}}|W] = W\gamma
\]
(where \( W \) represents all the variables used in the imputation procedure).

Under the data generating process described in Equation 4, however, this will produce biased estimates of \( Y_{\text{mis}} \). In short, one would want to impute \( Y_{\text{mis}} \) using \( X \), which means 
\[
E[Y|W] = X\gamma.
\]
Of course, this corresponds to running a linear regression model for spatially autocorrelated data, which produces a biased estimate of \( \gamma \) (i.e., \( E[\hat{\gamma}|X] \neq \beta \)) as Franzese and Hays (2007) show. Thus one imputes biased estimates of \( Y \) given \( X \), which infect the estimates from the spatial regression.

Previous work on addressing missing data in spatial regression has taken other approaches. One approach used in political science replaces the missing values of \( Y \) with meaningful values in order to retain the full complement of data. For example, Beck, Gleditsch and Beardsley (2006) replace missing values of their dyadic dependent variable trade capturing trade with reverse trade flows. This approach will may improve results relative to listwise deletion, but at the very least does not comport with current best practices through the use of imputation as well as ignoring the issue that single imputation of any sort will lead to excessively small standard errors.

Outside of Political Science, there has been some research in economics on missing data in spatial regression. Kelejian and Prucha (2010) address missing data as it relates to the “boundary problem” in spatial analysis. For units located at the edge of a sample, we observe their outcomes and exogenous variables, but we do not observe some of the units that influence these boundary outcomes through spatial spillovers. Consequently, we are missing information needed to calculate spatial lag variables for these units. Kelejian and Prucha show that instrumental variable estimation is consistent for the parameters of spatial lag models in this particular context.
LeSage and Pace (2004) focus on real estate markets, treating the prices of unsold houses as missing data. They suggest replacing the unobserved prices with their expected values conditional on the observed data. They partition the model into sold and unsold subsamples, as below, where the subscripts $S$ and $U$ reference the sold and unsold observation matrices and vectors respectively.

\[
\begin{pmatrix}
  y_S \\
  y_U
\end{pmatrix} = \begin{pmatrix}
  \Gamma_{SS}X_S\beta + \Gamma_{SU}X_U\beta \\
  \Gamma_{UU}X_U\beta + \Gamma_{US}X_S\beta
\end{pmatrix} + \begin{pmatrix}
  \Gamma_{SS} & \Gamma_{SU} \\
  \Gamma_{US} & \Gamma_{UU}
\end{pmatrix} \begin{pmatrix}
  u_S \\
  u_U
\end{pmatrix}
\]

Define the covariance matrix of the reduced form disturbances as $E[\nu\nu'] = \Omega$ and $\Pi = \Omega^{-1}$. The conditional expectation for the price of unsold properties is

\[
E(y_U|y_S) = \mu_U + \Omega_{US}\Pi_{SS}(y_S - \mu_S) = \mu_U + \Omega_{US}\Pi_{SS}(v_S)
\]

This leads to a standard EM approach that iterates between imputation and estimation.

The method suggested by Lesage and Pace is similar to ours, but it differs in three important ways. First, Lesage and Pace do not recommend or consider multiple imputation. Single imputation can lead to overconfidence in the precision of estimated parameters. Second, Lesage and Pace impute the reduced form disturbances for missing observations using information about the reduced form disturbances of non-missing observations only. This can be inefficient as it ignores relevant information such as the variance and covariance information contained in $\Omega_{UU}$. Third, it does allow for missing values in the independent variables, which limits its applicability.
4 An EM Approach for Imputation and Estimation

To develop an estimator that simultaneously addresses spatial interdependence and right censoring we build off of Hays, Schilling and Boehmke’s (2015) EM approach for estimating spatial duration models with right censoring, which in turn was inspired by Wei and Tanner’s (1991) approach for estimating nonspatial durations with censoring, and LeSage and Pace’s (2004) approach for completely missing outcomes. In the former the data on $Y$ were partially observed since they were known to be greater than an observed censoring point, so the goal was to impute values at least as large as the censoring point while accounting for the spatial interdependence in the data. Our approach here follows the same logic by alternating between taking draws of the missing outcome variable for missing cases and estimating the model using both the observed values and the imputed values.

These approaches were developed to fill in missing information about the dependent variable using the EM algorithm (see, e.g. Dempster, Laird and Rubin 1977; McLachlan and Krishnan 2007). Here we want to allow for both missing values of the outcome and the independent variables. Imputation for multiple missing variables typically assumes a multivariate normal distribution (with adaptations for possibly non-normally distributed variables), but the presence of spatial interdependence makes this difficult to employ here. For some patterns of spatial connectedness, the outcome variable for one observation will depend on the outcome for every other observation. This violates the independence of the multivariate normal approach and complicates accounting for that interdependence as one might be able to do with time series data. We therefore turn to imputation via the iterated chained equations (ICE) approach, which has been shown to work well in the past for independent observations. Applying this at iteration $k$ in the ICE means implementing a single instance of an EM estimator for missing values of $Y$ given the current values of $X^{(k)}$ and then updating $X^{(k)}$ using the aforementioned values
of $Y^{(k)}$. We then repeat this procedure until the parameter vector converges.

Given the chained equations approach, we describe our procedure in its two steps by starting with the imputation of missing $Y$ given $X$ and then describing how we impute missing $X$ given $Y$. For exposition, we assume that we have two independent variables, $X_1$ and $X_2$, where the former belongs in the spatial regression of interest and the latter does not, but provides information with which to impute values of $X_1$. We write $X = [X_1 X_2]$. We assume that we fully observe $X_2$ but not $X_1$.

**4.1 Imputing the Outcome Variable**

To set up the procedure it helps to follow the approach of LeSage and Pace (2004) in sorting the data based on whether an observation is fully observed, incorporate the fact that we work with the current draw of $X^{(k)}$ and the most recent values of $Y$ from the previous iteration:

$$y^{(k-1)} = \Gamma X^{(k)} \beta + \Gamma u,$$

$$\begin{bmatrix} y_{mis}^{(k-1)} \\ y_{obs}^{(k-1)} \end{bmatrix} = \begin{bmatrix} \Gamma_{mis,mis} & \Gamma_{mis,obs} \\ \Gamma_{obs,mis} & \Gamma_{obs,obs} \end{bmatrix} \begin{bmatrix} X_{mis}^{(k)} \\ X_{obs}^{(k)} \end{bmatrix} \beta + \begin{bmatrix} \Gamma_{mis,mis} & \Gamma_{mis,obs} \\ \Gamma_{obs,mis} & \Gamma_{obs,obs} \end{bmatrix} \begin{bmatrix} u_{mis} \\ u_{obs} \end{bmatrix}.$$  

This makes clear the dependence of observed outcomes on unobserved outcomes both through the exogenous variables and through the error term. Properly imputing missing values therefore requires accounting for the interdependence between observations during the estimation and imputation stages. In order to do this, we need to sample from the multivariate distribution of reduced-form residuals. In the case of a multivariate normal distribution, we can use the Geweke-Hajivassiliou-Keane (GHK) sampler (Geweke 1989; Hajivassiliou and McFadden 1998; Keane 1994), which translates the correlated reduced-form errors into a linear combination of i.i.d. normal errors. This translation occurs via
the Cholesky decomposition which, as an upper triangular matrix, allows us to iteratively solve for both the correlated reduced-form errors as well as their i.i.d. components. For missing cases we take i.i.d. draws from an independent normal distribution whereas for observed cases we analytically solve for the precise values that then reproduce the observed values of $Y$ while preserving the estimated spatial correlation. This combination of random draws and calculated values is then multiplied by the Cholesky decomposition to produce a vector of errors with the estimated amount of spatial correlation. We then calculate the resulting values of $Y$, reestimate and repeat the process until the parameters stabilize.

The algorithm starts with the most recent spatial regression estimates. Using these results, we calculate the reduced form residuals. The data generating process follows from equation 4. The reduced form error $v$ represents a linear combination of i.i.d. errors. Using these reduced form residuals, we then calculate the Cholesky decomposition, $A^{-1}$, of the expected covariance matrix of these errors:

$$E[\hat{v}\hat{v}'] = E[(\Gamma\hat{u})(\Gamma\hat{u})'], \quad (9)$$
$$= \Gamma E[\hat{u}\hat{u}']\Gamma', \quad (10)$$
$$= \Gamma(\sigma^2 I)\Gamma', \quad (11)$$
$$= \sigma^2 \Gamma\Gamma'. \quad (12)$$

Where $\sigma^2$ represents the variance of the error terms, which follow a normal distribution with mean zero.

We can then write the reduced form errors as a linear combination of i.i.d. normal errors: $\hat{u} = A^{-1}\eta$. Since $A^{-1}$ is upper triangular, we can iteratively solve for the corresponding value of $\eta$ one observation at a time starting with the last observation. Given the way we divide the data into missing and then observed cases, the procedure calcu-
lates values for all the observed cases first. Assume without loss of generality that we have solved for the $k^{th}$ to last observation and that we observe $Y_k$. When we solve for the $(k - 1)^{th}$ to last observation, if $Y$ is observed this results in the values obtain by solving the linear described above. If the $(k - 1)^{th}$ to last observation is missing, we take a draw from the estimated distribution, i.e., normal with mean zero and standard deviation $\hat{\sigma}$, and enter it for $\eta_{k-1}$ (see our appendix for an example). We then treat this sampled value as fixed and repeat for the previous observation.\(^3\)

In order to create the imputed reduced form spatial errors, we multiply $\eta$ by $A^{-1}$. Using these imputed values for the errors, an imputed $y$ is calculated from the combination of $\hat{y}$ from the previous spatial lag model and the imputed spatial errors. This method returns the observed value for each of the nonmissing cases and imputed outcomes from the estimated data generating process for the missing cases. The imputed $y$ is then run in a spatial lag model and the parameter vector is saved. The algorithm continues until the results converge, which we assess by whether the log-likelihood changes by less than a small amount, e.g., 0.000001.\(^4\) This produces an estimate of $y^{(k)}$, which we use in the next step to generate an update of the exogenous variables.

### 4.2 Imputing Exogenous Variables

Calculating the next draw of the exogenous variable follows the standard ICE approach given the independence across observations. This can be written as the following linear

\(^3\)We discovered that sorting based on missingness matters for the quality of our estimates. Putting observed cases at the end means that only estimated errors from observed cases are used to solve for the corresponding entry of $\eta$. Failing to do sort the data in this way lead to a modest amount of bias, which we believe results from the inclusion of both analytically derived and imputed errors when calculating the value of $\eta$ for observed cases, which we suspect acts in a similar fashion to measurement error in spatial regression.

\(^4\)Tanner and Wong (1987) discuss the conditions under which the algorithm converges.
regression:

\[ X_1^{(k)} = \alpha_0 + \alpha_1 Y^{(k)} + \alpha_2 X_2 + \alpha_3 W Y^{(k)} + \epsilon. \]  

We then generate the next draw \( X_1^{(k+1)} \) with the predicted value that results from the estimates of the above equation.

### 4.3 Bringing it All Together

We alternate between these two steps until the parameter vector for imputing \( y \) converges. Once it does, we estimate the spatial regression model of interest on the final draw. Note that this includes only the variables thought to influence \( y \) and not the entire set of variables used in the imputation procedure. Here that means we regress \( y \) on \( X_1 \).

To account for estimation uncertainty in our procedure we multiply impute \( M \) draws of the errors for cases with missing values of \( Y \) (we have not implemented multiple imputation for missing \( X \) yet, but plan to do so next). After each step of the EM process, we take the average of the \( M \) parameter vectors as our current estimate to begin the next step. Once it converges, we calculate the final estimates using each of the multiply imputed data sets are combined using Rubin’s (2009) formula:

\[ \bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m, \]  

\[ Var(\bar{\theta}) = \frac{1}{M} \sum_{m=1}^{M} Var(\hat{\theta}_m) + \frac{M + 1}{M} \left( \frac{1}{M - 1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})^2 \right). \]
5 Monte Carlo

In order to evaluate our approach we conduct a series of Monte Carlo simulations. These allow us to study the EM algorithm’s properties, both statistically and computationally, and to compare the estimates that it produces to those that one would obtain with alternate procedures and to those from the original data before any missingness is applied as a benchmark comparison. To make these comparisons across a wide range of circumstances we vary both the amount of spatial interdependence and the amount of missing data.

Our data generating process proceeds as follows. We start with two hundred and twenty-five units spread out evenly across a fifteen by fifteen grid. We construct a spatial dependence matrix, $W$, based on queen contiguity. We then generate 225 independently and identically distributed observations of two independent variables, $X_1$ and $X_2$. Since we expect that spatial correlation in $X_1$ that overlaps with the spatial correlation in $Y$ will influence the presence and amount of bias, we generate these according to the following equations:

$$X_{1i} = \left[ \frac{\sqrt{(z_{1i} - 8)^2 + (z_{2i} - 8)^2}}{2.13} + \nu_{1i} \right] \frac{1}{\sqrt{2}},$$

$$X_{2i} = \left[ \frac{\sqrt{(z_{1i} - 8)^2 + (z_{2i} - 8)^2}}{2.13} + \nu_{2i} \right] \frac{1}{\sqrt{2}},$$

where $(z_{1i}, z_{2i})$ represents the location of observation $i$ in the ten by ten grid, which has (5.5, 5.5) as its center, and $\nu_1$ and $\nu_2$ represent i.i.d. standard normal random variables. We divide the first term by its standard deviation so that the two terms contribute equally and then normalize the entire variable to have mean zero and variance one. This produces two variables with spatial correlation near 0.7 that have a correlation of 0.5 with each other. We assume that the error term $u_i$ is i.i.d. standard normal, and generate $y$ using the following
data generating process:

\[ y = (I - \rho W)^{-1}(0 + X_1) + (I - \rho W)^{-1}u. \]  

(18)

With 225 observations our approach to generating \( W \) results in about around 3.2% connectivity between units. As is common, we row-standardized the spatial weights matrix by dividing each element by the sum of the elements in its row, which produces a matrix in which all of the elements represent proportions and each row sums to 1. Since most spatial relationships in political science are expected to be positive we run simulations with \( \rho \) varying from 0 to 0.75 by increments of 0.25. We hold the independent variables constant across all of the simulations.

To compare our estimator with previous approaches and since we have not fully worked out all the details of the version with missingness in multiple variables, we perform two sets of simulations. The first considers the case of missingness in only \( y \) and the second extends to the case with missingness in \( X_1 \) as well.

5.1 Missingness in the Dependent Variable

In this first set of simulations, we introduce missingness of three types and at two different rates. First, we generate data missing completely at random by setting 20 and then 40% of the values of \( y \) to missing. Second, we generate data missing at random by setting the probability of missingness in \( y_i \) according to \( 2\gamma \Lambda(X_{1i}) \), where \( \Lambda \) indicates the standard logistic cdf and \( \gamma \in [0, 0.5] \) is the desired rate of missingness. Note that on its own \( \Lambda(X_{1i}) \) gives an average probability of missingness of 0.5 since we standardized \( X_1 \), so we just scale from there, selecting \( \gamma = 0.2 \) and 0.4. Recall that neither approach would lead to bias in standard regression approaches. In the spatial context, however, we expect both to produce bias.
For each type and rate of missingness we generate 250 draws of $u$ from a standard normal distribution, calculate the value of $y$, then apply our missingness rule. We estimate five models: a spatial regression with listwise deletion, an independent multiple imputation approach that ignores the spatial dependence in the data, the LeSage and Pace EM approach that replaces missing values of $y$ with their expected values, our EM spatial regression model with multiple imputation of missing values, and a spatial regression model using the fully observed data. The latter serves as a benchmark against which to compare the others. For the naïve multiple imputation approach we include $y$, $X_1$, and $W X_1$ in the imputation procedure, with the latter included to help account for the spatial interdependence. We use $M = 5$ imputations and since we only have one missing variable, we use the built-in linear regression option in Stata’s mi imputation procedure.

For our EM estimator we set the maximum number of iterations for each draw to 200, the convergence criterion to 0.0000001, and use five imputations for each step of the EM process. At the higher rate of censoring, our EM estimator occasionally fails to produce estimates. In exploring individual draws we found this to be an issue usually resolved by changing the seed and rerunning the estimator, so we set our simulations to run the estimator up to three times with different seeds which greatly increased the convergence rate. We calculate standard errors with multiple imputations according to Equation 15.

[Figure 1 here.]

[Figure 2 here.]

Figures 1 and 2 present the average bias under MCAR and MAR based on $X$ for the intercept, slope coefficient, and spatial dependence parameters. Given the similarity in the results we discuss them in tandem. Note that while the benchmark estimates show little evidence of bias, there is a slight tendency towards underestimation for the spatial dependence parameter. This is consistent with other simulations of spatial estimators and should be kept in mind in evaluation the performance of the other estimators since it
represents the best case scenario in which no missingness occurs.

Start with the listwise deletion (LD) estimator, as it may be the approach adopted by most researchers. As expected, clear patterns of bias emerge in the spatial correlation parameter and in the slope coefficient. The former appears whenever the true correlation is not zero and increases with the amount of missing data. The latter increases both with the amount of missing data and the degree of spatial correlation, reaching nearly 20% with 40% missingness and $\lambda = 0.75$. Now consider the independent multiple imputation (MI) estimates. These show the greatest amount of bias for all three parameters, often by a factor of two or three. The bias inherent in the linear regression assumption used in the imputation procedure appears to exaggerate the relationship estimated by the spatial regression model.

Turning to the two EM type estimators, we see that both the LeSage and Pace (LP) EM approach with expected values (EM-EV) and the EM estimator with multiple imputation (EM-MI) produce estimates quite close to the true parameter value and that both greatly outperform the previous two approaches with non-zero spatial correlation. The EM-MI approach generally produces modestly greater deviations from the true values, but still quite small and in inconsistent directions.

Turning to a more focused comparison of the two EM approaches, Figure 3 compares their average standard errors to their sampling distributions’ standard deviation. While the EM-EV approach produces accurate estimates, its use of the expected value for the missing cases likely does not reflect the uncertainty inherent in these estimates. Unsurprisingly, then, Figure 3 indicates that the EM-EV approach produces standard errors substantially smaller than the standard deviation suggests they should be (these results come from the in MAR based on X simulations, but similar results occur for the MCAR simulations). In contrast, the EM-MI approach appears to provide much more accurate measures of uncertainty. These results should be treated as preliminary, however, since
we believe that further adjustments to the calculation of the standard errors need to be made in the EM-EV approach.

[Figure 3 here.]

Having assessed the bias of these estimators, we now move to a comparison of their performance on root mean squared error terms in Figures 4 and 5, which plot the square root of the sum of the variance and the squared bias of each parameter. The results mirror those for bias with the fully observed estimator producing the smallest RMSE, followed by the EM-EV, EM-MI, LD, and then the MI estimators. The EM-MI generally equals or outperforms the LD estimator in the presence of nonzero spatial autocorrelation. And while the EM-EV does the best (excepting the fully observed estimates), recall that it exhibits underestimated standard errors.

[Figure 4 here.]
[Figure 5 here.]

Overall, then, these results provide solid evidence in favor of our EM spatial model that accounts for missing data in the dependent variable. Even with relatively modest amount of spatial correlation, it reduces the bias that appears in listwise deletion with little to no cost in root mean square error terms and provides more accurate standard errors than the less biased EM-EV approach. Given this we now incorporate missingness in the independent variable in our next round of simulations.

5.2 Missingness in the Dependent and Independent Variable

With confidence in our imputation procedure for y, we now wrap its imputation into the more general ICE procedure that imputes missing values of \( X_1 \) as well, alternating between the two until convergence. Given the added complexity, we run the simulation with \( \lambda = 0.5 \) and set 30% of values of \( y \) missing according to MAR with probability \( 2(0.3)\Lambda(0.5X_{1i} + 0.5X_{2i}) \) and 15% of values of \( X_1 \) missing with probability \( 2(0.15)\Lambda(X_{2i}) \).
We estimate the same five different models here with the caveat that the two EM approaches refer to how we estimate $y$ in the context of the broader EM algorithm that iterates between updating $y$ and $X_1$. We employ single mean imputation in this process as in LeSage and Pace (2004) and then report the results of spatial regression of $y^{(K)}$ on $X_1^{(K)}$ for the EM expected value estimator and the results of running of EM with Imputation estimator for the spatial regression $y^{(K)}$ on $X_1^{(K)}$, with $K$ representing the final EM iteration. For the independent multiple imputation procedure we impute the two missing variables as bivariate normal and use the fully observed $X_2$ and the observed component of the spatial lag $WX_1$, i.e., treating missing cases of $X_1$ as zero.

[Figure 6 here.]

[Figure 7 here.]

We again report the results of our simulation in terms of bias and root mean squared error. Figure 6 report the average deviation of the estimates from the true parameter values. The results indicate that at the selected parameter values both of our EM approaches produces estimates closer to the truth than either listwise deletion or independent multiple. We see less improvement in the coefficient on $X_1$ compared to the case with just missing values of $y$, with bias nearing 25% for the EM approaches compare to over 20% for the LD and MI approaches. The root mean squared error results show a similar pattern.

6 Conclusion

We offer three contributions in this paper. First, we demonstrate the consequences of listwise deletion in the context of spatial regression. Unlike standard regression models, spatial regression with listwise deletion can produce biased parameter estimates even when the data are MCAR or MAR based on exogenous variables. Our Monte Carlo provides the
first detailed study of the existence and extent of such bias. Interestingly, we also find that with spatial data, multiple imputation appears to be worse than listwise deletion. Second, we provide an EM approach to address the issue of missingness in the dependent variable. Contrary to previous implementations (LeSage and Pace 2004) ours involves multiple rather than single imputation of the missing information, which produces more accurate standard errors. And third, we extend these methods to begin to address missingness in both the dependent variable and independent variables.

Clearly more work needs to be done. Most clearly we need to continue to refine our ICE procedure for missingness in multiple variables and extend it to more general settings. Second, we would like to illustrate our approach using real-world data through a series of replication studies. We also plan to consider alternate models of spatial dependence, including spatial error models and the spatial Durbin model.
References


URL: http://dx.doi.org/10.1111/j.1540-5907.2010.00447.x


Figure 1: Average Bias from Alternate Estimation Procedures Under MCAR, varying the Rate of Missingness and the Amount of Spatial Correlation

Notes: Results represent the average deviation from the true parameter value across 250 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations.
Figure 2: Average Bias from Alternate Estimation Procedures Under MAR (Based on $X$), varying the Rate of Missingness and the Amount of Spatial Correlation

**Notes:** Results represent the average deviation from the true parameter value across 250 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations.
Figure 3: Comparison of Standard Errors and Standard Deviations for estimates of Slope Coefficient from EM approaches under MAR based on $X$, varying the Amount of Spatial Correlation

Notes: Results represent the average estimated standard errors and the standard deviation of the estimates across 250 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations.
Figure 4: Comparison of Root Mean Standard Errors from Alternate Estimation Procedures Under MCAR, varying the Amount of Spatial Correlation

Notes: Results represent the average deviation from the true parameter value across 250 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations.

\[ RMSE^2(\hat{\theta}) = (\hat{\theta} - \theta_0)^2 + Var(\hat{\theta}). \]
Figure 5: Comparison of Root Mean Standard Errors from Alternate Estimation Procedures Under MAR (Based on $X$), varying the Amount of Spatial Correlation

**Spatial Lag Parameter**

**Intercept**

**Coefficient on X1**

Notes: Results represent the average deviation from the true parameter value across 250 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations.

$RMSE^2(\hat{\theta}) = (\hat{\theta} - \theta_0)^2 + Var(\hat{\theta})$. 

26
Figure 6: Average Bias from Alternate Estimation Procedures under MAR in Both $Y$ and $X$

Notes: Results represent the average deviation from the true parameter value across 200 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations. Spatial correlation parameter set to 0.5, missing rate set to 0.3 for $Y$ and 0.2 for $X_1$. 
Figure 7: Comparison of Root Mean Standard Errors from Alternate Estimation Procedures under MAR in Both Y and X

**Notes:** Results represent the average deviation from the true parameter value across 200 simulations, excluding cases for which the EM algorithm did not converge in three attempts with up to 200 iterations. EM algorithm performed with 5 imputations. 

\[ RMSE^2(\hat{\theta}) = (\hat{\theta} - \theta_0)^2 + \text{Var}(\hat{\theta}) \]. Spatial correlation parameter set to 0.5, missing rate set to 0.3 for Y and 0.2 for X1.