Article

Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database

Urban Affairs Review
1–24

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1078087417726394
journals.sagepub.com/home/uar

Cali Curley[1], Rachel M. Krause[2], Richard Feiock[3], and Christopher V. Hawkins[4]

Abstract

Studies of governments and local organizations using survey data have played a critical role in the development of urban studies and related disciplines. However, missing data pose a daunting challenge for this research. This article seeks to raise awareness about the treatment of missing data in urban studies research by comparing and evaluating three commonly used approaches to deal with missing data—listwise deletion, single imputation, and multiple imputation. Comparative analyses illustrate the relative performance of these approaches using the second-generation Integrated City Sustainability Database (ICSD). The results demonstrate the benefit of using an approach to missing data based on multiple imputation, using a theoretically informed and statistically supported set of predictor variables to develop a more complete sample that is free of issues raised by nonresponse in survey data. The results confirm the usefulness of the ICSD in the study of environmental and sustainability and other policy in U.S. cities. We conclude with a discussion of results and provide a set of recommendations for urban researcher scholars.

Keywords

imputation, sustainability, urban policy, missing data

[1]Indiana University–Purdue University Indianapolis, Indianapolis, IN, USA
[2]University of Kansas, Lawrence, KS, USA
[3]Florida State University, Tallahassee, FL, USA
[4]University of Central Florida, Orlando, FL, USA

Corresponding Author
Cali Curley, Indiana University–Purdue University Indianapolis, 801 W. Michigan St., BS4054, Indianapolis, IN 46202, USA.
Email: curleyc@iupui.edu

Introduction

This article seeks to raise awareness about how to treat missing data in urban studies research. Surveys provide a relatively efficient way to collect large amounts of consistently measured individual or organizational information needed to conduct comprehensive and accurate statistical analysis. This is particularly important if the aim of research is to produce generalizable findings and contribute to understanding a phenomenon by testing theory. However, missing data are a common and significant challenge in survey-based research. They often influence the selection of a statistical method of analysis, and, depending on their severity, can undermine the confidence of analysis. Nonetheless, the problems associated with missing data are among the least acknowledged issues when conducting and reporting analysis.

Missing survey data occur for three reasons: (1) noncoverage—the observation fell outside of the sample, (2) total nonresponse—the would-be respondent failed to respond to the survey, and (3) item nonresponse—the respondent skipped a particular survey item (Brick and Kalton 1996). Although data missing as a result of these different causes present distinct challenges for the researcher, listwise deletion, the default operation in most statistical software packages, is a common applied remedy for all three. This approach removes any observation from the analysis that has incomplete information, that is, is missing a value for any variable included in the model for any reason. Peng et al. (2006) examined 1,087 published studies in education and psychology, of which 48% contained missing data. Within that subset, they found that authors used listwise deletion 97% of the time.

This article demonstrates the impact that different remedies for missing data may have on research findings and offers a rationale for its appropriate treatment. We specifically discuss the classifications of missing data, the specific problems associated with each, and the common approaches that have been developed to address them. This is followed by an illustration of the treatment of missing data using three techniques—listwise deletion, single imputation, and multiple imputation—applied to data from the second-generation Integrated City Sustainability Database (ICSD) and comparison of their relative performance in analysis. We use the results of the analysis as the basis for a concluding discussion of the missing data techniques and provide a set of recommendations for researchers using survey data.

Overview of Missing Data

Three classifications of missing data are important to the following discussion: data Missing Completely at Random (MCAR), data Missing at Random (MAR), and data Missing Not at Random (MNAR). This taxonomy provides insight into which tool is appropriate for dealing with the missing data. Table 1 provides a brief overview of each. For data that are MCAR, the missing values are independent from values of observed or unobserved characteristics in the data set. Therefore, the missing value is not the result of a strategic choice on the part of the respondent, nor is it a function of other captured or uncaptured variables. This means that the observed pattern of missingness is not related to any other data, whether present or missing. For example, MCAR data might result if a survey respondent unintentionally failed to answer a question that the researcher is using as a variable in the analysis. It is difficult to ascertain whether data are truly MCAR; in this situation, the researcher must ask if there is any theoretical reason that the respondent may have wanted to avoid answering that question.

Little's (1988a) MCAR test can help inform the assessment as to whether data are MCAR or not. When encountering missing data, a researcher can calculate a chi-square test to examine patterns of missingness for a number of specified variables (the "mcartest" command in Stata). The null assumption is that the data are MCAR (fail to reject the null hypothesis by having a *p* value larger than .05). An application of this test is included in the discussion of listwise deletion below. This test is one of several tools that help determine whether the data associated with a variable are MCAR. Little's test should be used with a logit model—where the dependent variable is one if the value is missing and zero if not. The independent variables in this test would be the observed variables that explain whether the dependent variables value is zero or one (observed or missing).

If both tests suggest that the missing data are MCAR, then either listwise deletion or multiple imputations can be used without biasing the estimates. Because listwise deletion will impact the power of the analysis, multiple imputations may still be the preferred approach. However, if the overall number of cases lost is small, listwise deletion is still an appropriate method (King et al. 2001; Myers 2011). If, however, one of the tests fails, the missing data would need to be treated as either MAR or MNAR.

Data that are MAR can be predicted using observed variables. A common example is when an individual intentionally skips the question asking about his or her income in a survey but provides the researcher with their employment status, education level, and years of employment experience. In this context, the value of the missing data is dependent on the value of observed responses and thus, is characterized as being MAR.

Table 1. Overview of Types of Missing Data.

Missing Completely at Random (MCAR)	Missing at Random (MAR)	Missing Not at Random (MNAR, Nonignorable)
Missingness is independent from characteristics of either the observed data or the unobserved values in the data set.	Missingness is entirely explained by the observed data, that is, after observed values are accounted for, missingness is randomly distributed.	Missing observations are dependent upon unobserved values; missingness cannot be accounted for by controlling for observed data.

However, there is no available explanation for data that are MNAR. When data are MNAR, the researcher cannot approximate the missing values because the values of other relevant variables are also not observed. Consider the previous example. If the observed data did not include employment status, education level, or employment experience, it would be challenging to determine an expected value of the respondent's income. Moreover, a respondent's income itself often determines whether she or he provided a response. Therefore, if the researchers did not collect additional relevant explanatory variables, the missing data would be considered MNAR. Solutions to MAR data, such as multiple imputation, rely on the relationships between missing and observable data to determine the value of the missingness. Despite this, multiple imputation and maximum likelihood are often unbiased with MNAR data (Schafer and Graham 2002). Researchers may also consider for their MNAR data applying Heckman Selection Models to control for MNAR data (DeMaris 2014; Little 2016).

It is important to consider the reason why data are missing when determining their treatment in statistical analysis. As the different approaches—listwise deletion, single imputation, and multiple imputation—each make specific mathematical assumptions, misusing them may invalidate empirical results. Invalid assumptions and incorrect categorizations of missingness may (1) decrease the sample size and decrease the power to estimate models, (2) increase the potential for biased results, and (3) over or underestimate standard errors. These impacts are important. If a large number of observations is lost, the resulting analysis will lose statistical power, and variables that would have otherwise been statistically significant may no longer have enough variation to demonstrate their relationship to the dependent variable. If the subset of observations that were dropped due to missingness is systematically different from those that remain in the analysis, then both the sample and any subsequent estimates generated from it will be biased. Table 2 summarizes the advantages, disadvantages, concerns, and missingness assumptions of the different techniques explored in the next sections of this article.

Table 2. Techniques of Imputation.[a]

		Single Imputation		
Techniques	Listwise Deletion (Complete Case Analysis)	Mean Replacement (Mean Substitution)	Single Regression Replacement	Multiple Imputation
Technique Summary	Remove any entries with missing values; perform analysis without these observations	For variable “a” with missing values, take the mean of all included observations. Substitute the mean of “a” for missing values of “a”	Estimate the distribution of the missing variable(s), given covariates; take a random draw from this distribution for each value; perform analysis as usual[b]	Estimate the distribution (Bayesian posterior distribution) of the missing variable, given covariates; take random draws from this distribution to produce multiple versions (usually 3–10) of an imputed data set; perform analysis on each imputed data set and pool the results
Missingness Assumption	MCAR, occasionally MAR	MCAR	MCAR or MAR	MCAR or MAR
Advantages	Easiest, simplest	Preserves the mean of the data set; simple; allows use of all observations	Avoids bias in estimating; simpler than multiple imputation	Accounts for the extra uncertainty produced by imputing data; produces better estimates of missing values
Disadvantages	Loses valuable information; potentially contributes to bias	Artificially reduces standard deviation of data set, distorts relationships between variables	Misrepresents uncertainty of estimates; more complicated than listwise deletion or mean replacement	Requires complicated statistical methods or complicated software; harder to understand; takes extra steps

Impacts on Interpretation	Statistical analysis loses power; estimates could be biased if data are not missing completely at random	Estimate could be biased, standard errors will be artificially low; could produce results that are highly statistically significant, but inaccurate	Although theoretically unbiased, reduces confidence intervals of estimates	Because the method accounts for extra uncertainty, results can be interpreted as if data were not missing
References				
Method Exploration	Jones 1996, p. 223; Schafer and Graham 2002, p. 155	Downey and King 1998; Schafer and Graham 2002, p. 159	Donders et al. 2006, pp. 1088–89; Schneider 2001; van der Heijden et al. 2006[c]	Donders et al. 2006, p. 1089; King et al. 2001; Rubin 1987; Schafer 1997; Zhang 2003
Application	Park and Ha 2012, p. 394; Ryff and Keyes 1995, p. 722	Allen, Eby, and Lentz 2006, p. 572; Gallimore, Brown, and Werner 2011, pp. 186–87		Abayomi, Gelman, and Levy 2008; Fox and Swatt 2009; Miyama and Managi 2014

Note. MCAR = Missing Completely at Random; MAR = Missing at Random.

a. Additional missingness reference can be found in Schafer and Graham (2002, p. 151).

b. Single Imputation, defined more broadly, includes any method that replaces missing data with a single value. This would include mean replacement and hot-deck imputation; the latter is summarized by Andridge and Little (2010).

c. Applications of the single imputation technique are limited; these are primarily theoretical explorations of the technique.

Approaches to Handling Missing Data

Scholars use a variety of alternative techniques to accommodate missing data and minimize its negative effects. Three of the most widely used approaches identified by Little (1988b) are (1) examining the incomplete cases (Little 2016), (2) replacing values for missing data (Kong, Liu, and Wong 1994), and (3) providing statistical weights to complete cases (Brehm 1993; Little and Rubin 2014). Within the general category of data replacement, there are specific techniques that vary in complexity. Two commonly used techniques include single imputation via mean replacement and multiple imputation. This section proceeds through an examination of listwise deletion and these alternative techniques, and compares their performance using survey data in an application.

Listwise deletion is a commonly used approach to handle missing data and is a convenient choice in most software packages. Two conditions must be met for listwise deletion to be appropriate for dealing with missing data: the missing data are MCAR, and the sample remains large after the deletion of individual observations. Deleting observations for nonresponse is less consequential if the values are MCAR, because if missingness is completely random, the data deleted would also be random, and they would not cause the loss of important variation. As previously described, a statistical approach, referred to as Little's test, can be used to determine whether data are MCAR (Little 1988a). If the data are, instead, MAR or MNAR, they are inconsistent with the assumptions of listwise deletion, and its use may result in the sample mean being different from the population mean. It may also affect estimates in a manner like selection bias; if a set of respondents systematically chooses not to answer a question and those observations are then deleted from the sample, the observations that remain in the analysis may be meaningfully different from the larger population.

The second issue with listwise deletion is that it reduces the sample size and impacts the statistical power of the analysis. Smaller samples are more likely to generate false null results that might otherwise not be null with a larger sample. Consider a hypothetical survey sent to a population of 700 respondents that obtained a 50% response rate (n = 350). Of those respondents, 10% failed to answer a question contained in an analysis. If those missing data are MCAR, then, by dropping those incomplete cases through listwise deletion, the analysis is conducted on a random sample of 90% of those respondents (only losing 35 observations). If, however, 10% of the values are missing on four different variables (40% of the total data or 140 responses), the statistical power of a subsequent analysis is extensively reduced.

Single imputation is a general term that describes a family of missing data replacement techniques, including last value replacement, mean replacement, and single regression replacement. Last value replacement, which can be used with panel or time-series data, involves the replication of the most recent value in cases of missingness. Carrying the last known value forward yields a conservative estimate of the treatment effect when a posttest value is missing. For example, if a respondent was asked to rate their health on a scale of 1 to 10 and answered “8” the first time the survey was administered but failed to provide a response the second time it was administered, the researcher would replace the missing value with 8. A second version of value replacement, sometimes referred to as “hot-decking,” uses information from similar observations to replace missing data. It is built around a premise like that of propensity score matching; if observations can be matched with others that look similar across the known values for a set of variables, missing responses can be replaced by the value of its matched observation with observed responses. This technique is limited to data that are MCAR or MAR.

Mean replacement replaces missing observations with the mean value of that variable from observed responses in the sample. This preserves the overall mean of each variable but reduces the variation of the sample. By holding unobserved variables to the mean, it automatically sets the sum of squared differences for these observations to zero. This causes variance to be underestimated, and thus, it may not reflect the true relationship between the dependent and independent variables. When the degree of missingness is small and the sample size is large, this technique may be appropriate. The smaller the amount of missingness, the less impact this has on the overall variance estimate. However, in smaller samples, the effect of mean replacement on these relationships will be larger.

An advanced version of single imputation is the single regression replacement method. This approach uses relevant observed variables (i.e., “informing variables”) to predict the value of the missing response via a regression analysis. This technique works well for data that are MAR, because, by definition, the other variables that can inform the missing value are observed in the data. The variable with missing values that is being estimated serves as the dependent variable in a regression analysis, and the independent or “informing” variables are theoretically or statistically related to it. Once the coefficients of the informing variables are estimated, the missing value of the dependent variable can be calculated for each observation by substituting the associated values of each informing variable back into the estimated equation. This estimation technique allows the value of missing data to vary by observation based on responses to the informing variables.

Consider, for example, a scholar attempting to explain wages for a sample of survey respondents, but her data contains several missing responses to a question on professional competency. If she knows that age and education are correlated with the observed values for professional competency, she can use those variables in a regression to develop a best guess of professional competency for each respondent who failed to provide it. The imputed and observed values for professional competency can be used in a model to predict wages. In this illustration, it is not necessary to pick the “right” value for the missing data, but rather to provide a value that allows all of the other data to be used without hampering statistical inference (Rubin 1987, 1996).

In single regression replacement, the missing value is only measured once. This creates the potential for biasing the standard errors because there is no assessment of how likely the imputed value represents the true value. If the inherent uncertainty in the prediction of the missing values is not accounted for, subsequent analysis may be influenced by the predicted missing values more than the true observed data (creating the potential for included bias and over or underestimated standard errors).

Multiple imputation is an extension of the single imputation regression replacement method. As its name suggests, missing values are estimated multiple times. Analyzing multiply imputed data follows three steps: (1) the imputation of missing data, (2) the running of independent statistical analysis on the resulting individual data sets, and (3) the pooling of the results across the imputations.

The first step of multiple imputation is similar to the single regression replacement method described above; variables that are theoretically related or statistically correlated to the target variable are identified and used in a regression model to predict the values of the missing data. However, in multiple imputation, this replacement process is repeated numerous times to incorporate the uncertainty in the prediction process. Through this process, randomness is incorporated into the value of the error term, uncertainty in predicting the value of the missingness is included into the value of the missing data (Johnson and Young 2011; White, Daniel, and Royston 2010). Multiple imputation creates numerous data sets, each containing somewhat different estimates of the missing values. Rubin’s (1978) formula suggests three to 10 imputations are necessary to produce results that incorporate enough variation in the prediction process. Others argue the number of imputations should be similar to the percent of missing responses (Bodner 2008; Graham, Olchowski, and Gilreath 2007; Royston and White 2011). This ensures that the uncertainty inherent in the prediction of missing values is accounted for to appropriately increase the standard errors in the analysis.

A second key difference between single regression replacement and multiple imputation is how the data are analyzed as part of a theory-based model once missing values have been imputed. As described above, multiple imputation results in the creation of multiple data sets. Theory-based models that use multiple imputed data must, therefore, be estimated simultaneously with each set of data. Many statistical programs enable data to be specified as imputed, after which the simultaneous estimation is carried out automatically. For example, in Stata, multiple imputation data must be specified with the command *miset*, which clearly defines where one data set begins and ends. The analysis is then run with the phrase *mi estimate:* prior to specifying the model.

These designations instruct the statistical software to estimate the theory-based model across each of the imputed data sets. For example, if 20 rounds of imputation were used to generate values for the missing data, then 20 distinct data sets are created, and a theory-based model is estimated 20 times. Once the analysis is executed, the results are pooled together, and the pooled output is reported. The pooling process embeds the uncertainty from the imputation into the estimates of the standard errors, and the results can be interpreted as they would be for nonimputed data. There are several different pooling rules, but the specified defaults in statistical packages are usually appropriate. A detailed overview of pooling rules[1] for normally and nonnormally distributed parameters can be found in White, Royston, and Wood (2011) and Allison (2002), respectively.

In summary, multiple imputation works well when the missing data are MCAR or MAR and is particularly useful with MAR data. It helps to maintain the sample size and eliminate the potential selection that could result if cases with incomplete data were dropped. It also helps to reduce the likelihood of standard error bias. The three steps to analyzing imputed data are (1) imputing values for the missing data, (2) running theoretical analysis using the imputed data, and (3) pooling estimates into a single set of results. The first step involves imputing the missing values to generate an appropriate number of data sets. The number of imputations needed is dependent on the amount of missingness; the greater the percent of data that are missing, the larger the number of imputations that are needed. Each imputation results in the creation of another data set. The second step is analyzing the imputed data as part of the researcher’s theory-based model. This involves running the analysis simultaneously across each imputed data set. In most statistical software, this requires the researcher to specify the data as imputed. The final step is pooling the results. Pooling generates a single output that incorporates into the standard errors the potential uncertainty that is inherent in the imputation process.

Description and Illustration of ICSD Missing Data

The following sections illustrate the relative advantages and disadvantages of each approach by applying it to the ICSD (Feiock et al. 2014). We compare listwise deletion, single mean replacement, and multiple imputation techniques to demonstrate the value-added from using multiple imputation when the degree of missingness can have an impact on the outcome of analysis.

A recent article in this journal by Feiock and colleagues (2014) described the ICSD as a solution to the challenges associated with missing data in urban research. The ICSD combines the results of seven national surveys of city sustainability programs that were administered within an 18-month period in

2010–2011 into one comprehensive national data set. Table 3 presents basic information on the seven ICSD component surveys.[2] The process of survey harmonization yielded a large sample: 2,825 cities completed at least one of the seven surveys. However, the majority of cities did not answer all seven of the surveys, meaning that the ICSD contains a considerable amount of missing data.

The first generation of the ICSD uses a single regression replacement method to account for missing data (Feiock et al. 2014). The authors deal with missing observations within and across the surveys using a two-stage informed single regression imputation technique, which produced a single unified data set. The first stage imputed missing data within each completed survey, and the second used these data to impute across surveys, taking into consideration the different types of missingness. This process generates a single unique value for each missing observation in the original ICSD and results in a single complete data set for the ICSD. This “first-generation” ICSD enables more confident conclusions to be drawn from the results of empirical analyses of local politics, governance, and policy (Hawkins et al. 2016). It provides single imputed data for an extensive set of cities including both large and small cities that are already being widely used in urban research.[3]

The two-stage single imputation approach of the first-generation database is a significant improvement over listwise deletion, but further improvement is possible through the process of multiple imputation for the cities more than 50,000 in population. Cities above this population threshold were included in the sample frames for all seven surveys, making their overall levels of missing data lower and making them better candidates for multiple imputation. The second-generation ICSD described here complements the first-generation database by providing a multiple imputation version for this subset of ICSD cities.

The 683 U.S. cities, which per the 2010 Census, had populations more than 50,000, were included in the sample for each of the seven ICSD component surveys. Their response was particularly strong, with 90% of these cities responding to at least one survey. This virtually eliminates self-selection bias among this subsample and provides a unique opportunity to examine the sustainability policy, implementation, resources, obstacles, and motivations in medium and large U.S. cities. However, although they all

Table 3. Characteristics of the Surveys Comprising the Integrated City Sustainability Database.

Survey Name	Sampling Frame	Respondents	Response Rate (%)

ICMA Local Government Sustainability Policies and Programs Survey	8,569 local governments with a population of 10,000 or more residents	2,176	25.4
NLC Sustainability Survey	1,708 mayors in cities more than 10,000	442	26.6
EECBG Grantee Implementation Survey	970 municipal governments receiving EECBG awards, including all cities more than 30,000	747	77
Implementation of Energy Efficiency and Sustainability Programs	1,180 cities: all with populations more than 50,000 and a random sample of 500 cities with populations between 20,000 and 50,000	679	57.5
National Survey of Sustainability Management in U.S. Cities	601 cities with populations more than 50,000	263	44
Municipal Climate Protection Survey	664 cities with populations more than 50,000	329	49.5
Municipal Government Questionnaire	425 cities with populations more than 50,000 that have indicated explicit involvement in climate protection	255	60

Note. ICMA = International City/County Management Association; NLC = The National League of Cities; EECBG = Energy Efficiency and Conservation Block Grant.

shared a related scope, each survey used a somewhat different set of questions and response categories and ended up with a different set of responding cities. This is problematic in a multivariate context where models seek to draw information from across several surveys, because it can drastically reduce the sample size of available data. This reduction in sample size provides an important rationale for using a more advanced method of dealing with missing data, such as multiple imputations.

Figure 1 summarizes the process used to identify the theoretic and statistically relevant variables that were used as informing variables in the imputation process for the second-generation ICSD. The theoretical linkages were determined by developing two “general concepts”—one related to the “activity” and the other to “subject matter”—for every question contained within the seven surveys. For example, the question “Do any of your city’s efforts to encourage retrofits for energy efficiency include: Partnership or collaboration with nonprofit community organizations” is labeled with the activity concept of “Collaboration” and the subject matter concept of “Energy.” This develops sets of potentially theoretically related questions—called the concept list. A list of these concepts and how often they are attributed to variables in the ICSD surveys is presented in Table 4.

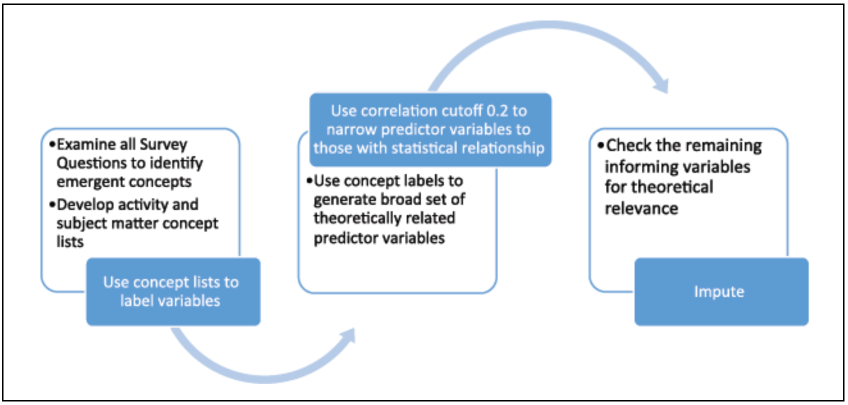

Figure 1. Process flow of informed multiple imputation.

Table 4. General Concept Description.

General Concept	Category	Description/Keywords	Count[a]
Climate	Subject matter	Climate change, climate protection, adaptation	71
Economic	Subject matter	Green business, green jobs, buy local programs, farmers' market	50
EECBG	Subject matter	Energy Efficiency Conservation Block Grant, ARRA, stimulus	109
Energy	Subject matter	Energy, energy efficiency, energy conservation	306
Environment	Subject matter	Land use, water, recycling, trees, community gardens, food	122
Social	Subject matter	Low-income, population, health, equity	32
Sustainability	Subject matter	Sustainability	172
Transportation	Subject matter	Vehicles, car-pooling, telework, condensed/flexible work days	69
Collaboration	Activity	Collaboration in general, partnership, cooperation	70
Community action	Activity	Any policy or programmatic action (loan program, tax credit, rebates, regulation, retrofit) that targets the community at large	114
Community planning	Activity	Inventory from community-wide emissions	7
Contracting	Activity	Contracting, outsourcing	29
General action	Activity	Any policy or programmatic action that does NOT specify target groups	93
General Planning	Activity	Planning, adopted planning goals, adopted policy	36
Government Action	Activity	Any policy or programmatic action targeting government operations	128

		(publicly owned building, purchase (credits), incentives, utility retrofit)	
Government Planning	Activity	Goal, inventory from city government operations	9
Infrastructure	Activity	Own operate, facility	46
Interdepartment	Activity	Coordinate within the city	46
Intergovernmental	Activity	Collaborate with other localities, state/federal government, cross-influence	59
Motivation	Activity	Why, What are the drivers of action?	45
Obstacle	Activity	Why not, Barriers	46
Performance measures	Activity	Measurement, resulting from efforts, indicators, evaluation	58
Priority	Activity	How important?	47
Public Engagement	Activity	Public education, info center, engage with . . .	31
Resources	Activity	Designated staff, money, funding	73

Note. EECBG = Energy Efficiency and Conservation Block Grant; ARRA = American Resource and Recovery Act.
a. Represents number of variables characterized as general concept.

The concept lists develop broad groupings of variables that have theoretic relationships and inform one another. In other words, these "informing variables" act almost as independent variables that may provide information to help predict missing values of a particular target variable. In some cases, the theoretically derived list of informing variables is too large to support convergence of the model determining the value of the missing responses and, therefore, statistical correlations are used to narrow the set. With the objective of identifying a small enough number of informing variables to enable statistical conversion, 0.2 was selected as the minimum correlation[4] between the variable being imputed and the potential informing variables. As a result, only variables that are theoretically and statistically relevant are retained as predictors, resulting in an average of 95 informing variables for each target variable in the data set.

A distribution of the nonmissing cases is used to determine the expectation of the distribution for missing responses. For example, if the nonmissing responses are normally distributed, the imputed responses will maintain a normal distribution. The distribution assigned is variable specific. A total of 20 imputations are generated for the results of the analysis that determines the value of a missing response. This process is repeated for all missing variables across the seven surveys. For the 683 cities with populations above 50,000, per the 2010 Census, complete data are generated for each of the 1,010 variables in the ICSD.

Compared with conducting analysis using either nonimputed or first-generation ICSD data, using the multiply imputed data generated from the process described above requires a few additional steps. The Stata code associated with these steps for several different types of analytic techniques has been included in the online appendix. As the code demonstrates, it is quite simple to analyze the imputed data. It primarily requires setting the data as multiply imputed and analyzing using *mi estimate:* prior to writing the code.

One complication with analyzing multiply imputed data is the generation of summary statistics. The goal of multiple imputations is to avoid generating a fixed point-estimate for the prediction of the missing value. Generating summary statistics of a single imputed data set, or each independently, would treat each data set as holding a true value for the missing observation. Therefore, traditional summary statistics are an inappropriate match for the technique because they do not account for the uncertainty inherent in the imputation. It may be more appropriate to report either a grand mean, which estimates the average of the multiply imputed data sets averages, and/or the descriptive statistics from the original, unimputed data.

A Comparison of Approaches Using the ICSD

We use the ICSD survey data in their raw and two imputed forms to demonstrate the relative performance of each of the three approaches to dealing with missing data: listwise deletion, mean replacement, and multiple imputation. For illustration purposes, we examine the factors that influence local action on sustainability in a generic empirical model that corresponds to those typical in the urban affairs literature.

Dependent Variable

The dependent variable is an additive index of the number of environmental sustainability-related policies and actions that cities reported having implemented in their jurisdictions. The additive index is a common dependent variable in quantitative studies of local sustainability (Bae and Feiock 2013; Krause 2012; Portney 2003). We select a dependent variable conducive to analysis using Ordinary Least Squares (OLS) regression. In all, 16 sustainability actions are included in this index and cluster in three primary areas: energy, transportation, and waste disposal.

Independent Variables

The independent variables reflect common operationalization of hypotheses in sustainability studies and relate to cities' motivations to engage in sustainability, obstacles hindering their action, and a series of control variables (Hawkins et al. 2016; Krause 2013; Krause, Feiock, and Hawkins 2016). A set of independent variables is drawn from the “Energy Efficiency and Conservation Block Grant (EECBG) Grantee Implementation Survey,” which asked respondents to identify the motivations for engaging in sustainability activities: achieving energy cost savings, the desire to build a sustainable community, and external public pressure. Two of the obstacle variables—lack of staff capacity and lack of information resources—likewise come from the EECBG Grantee Implementation Survey. The third obstacle—a lack of political will—is pulled from the Implementation of Energy Efficiency and Sustainability Programs Survey.[5]

Control variables include population density, per capita income, form of government, International Council for Local Environmental Initiatives (ICLEI) membership, percent of racial minority residents, and residents’ educational attainment. Each of these control variables has been used in previous studies regarding sustainability policy (Feiock et al. 2010; Krause 2010; Lubell, Feiock, and Handy 2009; Salon, Murphy, and Sciara 2014; Feiock and Bae 2011; Zahran et al. 2008; Betsill 2001). The data were collected from the U.S. Census Bureau, the International City/County Management Association, and ICLEI Local Governments for Sustainability, and thus, have near complete coverage.

Results

We apply OLS as our method of analysis and specify three identical models to estimate the impact of the different missingness treatments. The first model uses listwise deletion, the second uses single imputation mean replacement, and the third uses multiple imputations, which is the approach used in the second-generation ICSD.

Table 5, column two reports the results from the model using *listwise deletion*. Only 111 of the 683 cities with populations more than 50,000 remain in the model after listwise deletion removes all observations with incomplete data (a loss of 572). The results using this approach indicate that only one variable—ICLEI membership—has a statistically significant effect on the policy index. The drastic reduction in sample size and the resultant potential bias may contribute to the production of null findings related to motivations and obstacles to implementing policy.

The third column in Table 5 presents the results of the model using *mean replacement*. For each independent variable in the model, this technique simply replaces the missing observations with the mean value for that variable. This technique increases the size of the sample from 111 to 325. However, it still results in a total loss of 358 observations.[6] The results generated using mean replacement identify several additional statistically significant relationships compared with listwise deletion. Lack of political will, as well as the control variables of population density and education, are now statistically significant.

ICLEI membership remains significant, and the magnitude of its effect is larger. Perhaps the most meaningful change in the results is that, using mean replacement, lack of political will has a negative statistically significant relationship to the dependent variable. Cities lacking political will toward sustainability implement approximately one-half fewer

Table 5. Comparison of Missing Data Techniques—OLS of Additive Policy Index.

		Listwise Deletion		Mean Replacement		Multiple Imputation	
		Policy Index	*SE*	Policy Index	*SE*	Policy Index	*SE*
M	Reduced energy cost	–0.066	0.678	–0.402	0.385	–0.211	0.150
M	Sustainable communities	0.251	0.395	0.396	0.277	0.328**	0.136
M	Public pressure	0.446	0.354	0.394	0.255	0.145	0.129
O	Staff capacity	0.355	0.384	0.067	0.283	0.070	0.188
O	Lack of information	0.085	0.461	–0.165	0.320	–0.080	0.198
O	Lack of political will	–0.3027	0.377	–0.627**	0.297	–0.550***	0.177
C	Population per square mile	0.000	0.000	0.0001*	0.000	0.000	0.000
C	Per capita income	0.000	0.000	0.000	0.000	0.000	0.000
C	ICLEI member 2010	1.149**	0.582	1.814***	0.332	1.013***	0.255
C	Council manager	–3.372	2.763	–2.872	2.705	–0.384	0.497
C	Mayor council	–3.112	2.754	–2.973	2.702	–0.336	0.514
C	Percent minority	0.012	0.015	–0.008	0.008	–0.004	0.006
C	Percent bachelors+	0.051	0.034	0.033*	0.020	0.013	0.014
	Constant	8.473**	3.464	9.757***	3.015	8.170***	0.886
	Sample size	111		325		683	
		Adjusted R^2	.0906	Adjusted R^2	.1814	Prob > *F*	0

Note. Motivation Variable (M), Obstacle Variable (O), Control Variable (C). OLS = Ordinary Least Squares. Significance Levels: *** p<0.01, ** p<0.05, * p<0.1

policies than those reporting stronger political will. This suggests that listwise deletion lost a significant amount of variation by deleting observations with incomplete data. However, the concern associated with mean replacement is that the observed significant relationships between the variables may not be true due to underestimates of the standard deviation and standard error. OLS—regression to the mean—is not able to accurately measure variations from the mean (i.e., error) because observations are artificially held at the value of the mean. Therefore, even though these variables are significant, the resulting *p* values should be interpreted with caution.

The results from the analysis performed using informed *multiple imputation* are shown in the fourth column and yield a slightly different combination of statistically significant variables in the model, when compared with the other two approaches. Multiple imputation is typically accepted for use in the dependent as well as independent variables (Young and Johnson 2010), which enables the sample size to increase from 325 to 683. In this model, the motivation to build a sustainable community variable is statistically significant and positively associated with the policy index. ICLEI membership and lack of political will remain statistically significant, however, the magnitude of both decrease slightly compared with the other models. This model also yields statistically significant relationships for the motivation and the obstacle variables. Comparing these results with those from the listwise deletion model suggests that null findings in cases with large amounts of missingness may not be null findings after all. The standard errors in multiple imputation incorporate the uncertainty from the 20 imputation results, thus providing greater confidence in the resulting *p* values.

Discussion and Conclusion

Listwise deletion, mean replacement, and multiple imputation are common approaches for addressing missing data. Each is associated with advantages and disadvantages, and, depending on the nature of the missingness, using the wrong method may provide inaccurate, biased, or inappropriate null findings. This article elucidated these consequences and specifically described how inaccurate treatment can decrease the power of the sample size, increase the potential for biased results, and over or underestimate standard errors. This is not to say that multiple imputation is the correct or best solution for dealing with missing data. In fact, this article suggests that the categorization of missing data should drive the selection of an appropriate approach to dealing with missing data.

Although often a default, listwise deletion is not a blanket solution to missing data problems. Dropping observations from an analysis decreases its power, and its overuse may cause variables that help explain the outcome variable to be deemed insignificant. Also problematic is the potential of incorporated bias in the selection process. Listwise deletion might work for data that are MCAR, but data are very rarely MCAR. It is also possible that techniques such as mean replacement are suitable for use with MCAR data. However, it may result in the effect of these variables being vastly overestimated because the standard errors are made artificially smaller by holding the values to the mean. Multiple imputation, although more complicated, provides theoretically consistent results and works for data that are MAR. Incomplete observations are not dropped from the analysis and, by incorporating the uncertainty of missing responses into the standard errors, the magnitude and significance of the relationships between independent and dependent variables are appropriately measured.

Exploiting the ICSD allows us to examine the implications of various treatments of missing data. The second-generation ICSD contains data generated by informed multiple imputation, which enables analysis with larger sample size, less bias, and the ability to interpret the data as though they were not missing. In addition, this technique is applicable to data that are either MAR or MCAR. A large degree of the missingness in the ICSD can be attributed to survey recipient response, which makes multiple imputation an appropriate choice. However, some variables may not be MAR and, therefore, should be considered thoughtfully prior to applying this technique. In addition to being more complicated, a disadvantage to using multiply imputed data are that they are not conducive to the generation of standard descriptive statistics, including things such as variable means, and basic model fit indicators such as R^2.

In urban studies and across the social sciences, there are increasing expectations for rigor and transparency in the management of data including procedures for dealing with missing observations. This is manifested in the Transparency and Openness Promotion (TOP) guidelines that are being adopted by many journals (Nosek et al. 2015). It is our hope that urban scholars begin to treat missing data more explicitly and openly. Many different data sets may benefit from an exploration of why data are missing. Typically, surveys are a prime area for strategic nonresponse, which makes them a great example of how important missing data treatment is; however, secondary data with missing responses may face similar concerns. Included here is an online appendix, with multiple imputation code and description to aid in the utilization process. In 2018, the multiply imputed data included in the second-generation ICSD will be made publicly available. In the meantime, select variables from the first-generation ICSD are available at http://localgov.fsu.edu/ICSD/.

Authors' Note

Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgment

The University of Kansas Center for Research Methods and Data Analysis provided valuable assistance in the imputation process, as did our graduate research assistants Y.S. Angela Park and Christopher Holcomb.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the National Science Foundation under Grant Nos. 1461526/ 1461506/1461460.

Notes

1. For normally distributed parameters, the standard pooling process follows Rubin's Combination Rule, which incorporates the uncertainty generated by the process of imputation into the estimates of the standard errors. Rubin's Combination Rule incorporates the uncertainty or variation due to missing information and the results from just one data set. It does this by essentially averaging the variance over the imputed data sets and incorporating both within-imputation variance and between-imputation variance (White, Royston, and Wood 2011). Allison (2002) provides an overview of pooling methods for nonnormally distributed parameters. This pooling typically happens behind the scenes in software packages. Although the model outputs are the pooled coefficients from the individual analyses, the results can be interpreted in the same manner as one would in a normal setting.
2. The Integrated City Sustainability Database (ICSD) is a dynamic database that is expanding and anticipated to continue to grow over time as new data on city level sustainability are collected. The original ICSD establishes a 2010/2011 baseline on local sustainability initiatives. As more data are collected by the authors and others, they will be added to the ICSD to enable analyses of change over time.
3. The public release of the ICSD is scheduled for January 2018: http://localgov.fsu.edu/ICSD/.
4. The 0.2 correlation value selected for this specific data set indicated that a predictor was related to the variable being imputed. Anything below the 0.2 cutoff was deemed unrelated to the variable being imputed. The 0.2 correlation narrowed the related concept list enough to allow convergence and did not eliminate the theoretically related questions to zero in any case.
5. We only incorporate variables from three of the seven surveys in this model, which should keep the loss of observations from listwise deletion relatively low. This is done to demonstrate that a more advanced treatment of missing data may be valued even without extreme degrees of missing observations. In other words, we are giving the listwise deletion approach its "best chance" of success.
6. This is because using mean replacement for dependent variables is a debated procedure. If the dependent variable were mean replaced, the data would have the full 683 cities.

References

Abayomi, Kobi, Andrew Gelman, and Marc Levy. 2008. "Diagnostics for Multivariate Imputations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57 (3): 273–91.

Allen, Tammy D., Lillian T. Eby, and Elizabeth Lentz. 2006. "Mentorship Behaviors and Mentorship Quality Associated with Formal Mentoring Programs: Closing the

Gap Between Research and Practice." *Journal of Applied Psychology* 91 (3): 567–78.

Allison, Paul D. 2002. "Missing Data: Quantitative Applications in the Social Sciences." *British Journal of Mathematical and Statistical Psychology* 55 (1): 193–96.

Andridge, Rebecca R., and Roderick J. A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78 (1): 40–64.

Bae, Jungah, and Richard Feiock. 2013. "Forms of Government and Climate Change Policies in US Cities." *Urban Studies* 50 (4): 776–88.

Betsill, Michele M. 2001. "Mitigating Climate Change in US Cities: Opportunities and Obstacles." *Local Environment* 6 (4): 393–406.

Bodner, Todd E. 2008. "What Improves with Increased Missing Data Imputations?" *Structural Equation Modeling* 15 (4): 651–75.

Brehm, John O. 1993. "The phantom respondents: Opinion surveys and political representation." University of Michigan Press.

Brick, J. Michael, and Graham Kalton. 1996. "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5 (3): 215–38.

DeMaris, Alfred. 2014. "Combating Unmeasured Confounding in Cross-Sectional Studies: Evaluating Instrumental-Variable and Heckman Selection Models." *Psychological Methods* 19 (3): 380–97.

Donders, A. Rogier T., Geert J. M. G. van der Heijden, Theo Stijnen, and Karel G. M. Moons. 2006. "Review: A Gentle Introduction to Imputation of Missing Values." *Journal of Clinical Epidemiology* 59 (10): 1087–91.

Downey, Ronald G., and Craig V. King. 1998. "Missing Data in Likert Ratings: A Comparison of Replacement Methods." *The Journal of General Psychology* 125 (2): 175–91.

Feiock, Richard C., and Jungah Bae. 2011. "Politics, Institutions and Entrepreneurship: City Decisions Leading to Inventoried GHG Emissions." *Carbon Management* 2 (4): 443–53.

Feiock, Richard C., Rachel M. Krause, Christopher V. Hawkins, and Cali Curley. 2014. "The Integrated City Sustainability Database." *Urban Affairs Review* 50 (4): 577–89.

Feiock, Richard C., In Won Lee, Hyung Jun Park, and Keon-Hyung Lee. 2010. "Collaboration Networks Among Local Elected Officials: Information, Commitment, and Risk Aversion." *Urban Affairs Review* 46 (2): 241–62.

Fox, James Alan, and Marc L. Swatt. 2009. "Multiple Imputation of the Supplementary Homicide Reports, 1976–2005." *Journal of Quantitative Criminology* 25 (1): 51–77.

Gallimore, Jonathan M., Barbara B. Brown, and Carol M. Werner. 2011. "Walking Routes to School in New Urban and Suburban Neighborhoods: An Environmental Walkability Analysis of Blocks and Routes." *Journal of Environmental Psychology* 31 (2): 184–91.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8 (3): 206–13.

Hawkins, Christopher V., Rachel Krause, Richard Feiock, and Cali Curley. 2016. "Making Meaningful Commitments: Accounting for Variation in Cities' Investments of Staff and Fiscal Resources to Sustainability." *Urban Studies* 53 (9): 1902–24.

Johnson, David R., and Rebekah Young. 2011. "Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations." *Journal of Marriage and Family* 73 (5): 926–45.

Jones, Michael P. 1996. "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association* 91 (433): 222–30.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1): 49–69.

Kong, Augustine, Jun S. Liu, and Wing H. Wong. 1994. "Sequential Imputations and Bayesian Missing Data Problems." *Journal of the American Statistical Association* 89 (425): 278–88.

Krause, Rachel M. 2010. "Policy Innovation, Intergovernmental Relations, and the Adoption of Climate Protection Initiatives by U.S. Cities." *Journal of Urban Affairs* 33 (1): 45–60.

Krause, Rachel M. 2012. "Political Decision-Making and the Local Provision of Public Goods: The Case of Municipal Climate Protection in the US." *Urban Studies* 49 (11): 2399–417.

Krause, Rachel M. 2013. "The Motivations Behind Municipal Climate Engagement: An Empirical Assessment of How Local Objectives Shape the Production of a Public Good." *Cityscape* 15:125–41.

Krause, Rachel M., Richard Feiock, and Christopher Hawkins. 2016. "The Administrative Organization of Sustainability Within Local Government." *Journal of Public Administration Research and Theory* 26 (1): 113–27.

Little, Roderick J. A. 1988a. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association* 83 (404): 1198–1202.

Little, Roderick J. A. 1988b. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 6 (3): 287–96.

Little, Roderick J. A. 2016. "Selection Model (Missing Data)." In *Wiley StatsRef: Statistics Reference Online*. 1-5.

Little, Roderick J. A., and Donald B. Rubin. 2014. *Statistical Analysis with Missing Data*. Hoboken: John Wiley & Sons.

Lubell, Mark, Richard Feiock, and Susan Handy. 2009. "City Adoption of Environmentally Sustainable Policies in California's Central Valley." *Journal of the American Planning Association* 75 (3): 293–308.

Miyama, Eriko, and Shunsuke Managi. 2014. "Global Environmental Emissions Estimate: Application of Multiple Imputation." *Environmental Economics and Policy Studies* 16 (2): 115–35.

Myers, Teresa A. 2011. "Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data." *Communication Methods and Measures* 5 (4): 297–310.

Nosek, Brian A., George Alter, George C. Banks, Borsboom, D., S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422–25.

Park, Joohyung, and Sejin Ha. 2012. “Understanding Pro-environmental Behavior: A Comparison of Sustainable Consumers and Apathetic Consumers.” *International Journal of Retail & Distribution Management* 40 (5): 388–403.

Peng, Chao-Ying Joanne, Michael Harwell, Show-Mann Liou, and Lee H. Ehman. 2006. “Advances in Missing Data Methods and Implications for Educational Research.” In *Real Data Analysis,* 31–78.

Portney, Kent E. 2003. *Taking Sustainable Cities Seriously: Economic Development, the Environment, and Quality of Life in American Cities*. Vol. 67. Cambridge: MIT Press.

Royston, Patrick, and Ian R. White. 2011. “Multiple Imputation by Chained Equations (MICE): Implementation in Stata.” *Journal of Statistical Software* 45 (4): 1–20.

Rubin, Donald B. 1978. "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse." In Proceedings of the survey research methods section of the American Statistical Association (Vol. 1, pp. 20-34). American Statistical Association.

Rubin, Donald B. 1987. "Multiple Imputationfor Nonresponse in Surveys." New York: Wiley

Rubin, Donald B. 1996. “Multiple Imputation After 18+ Years.” *Journal of the American Statistical Association* 91 (434): 473–89.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. Hoboken: John Wiley & Sons.

Ryff, Carol D., and Corey Lee M. Keyes. 1995. “The Structure of Psychological Well-Being Revisited.” *Journal of Personality and Social Psychology* 69 (4): 719–27.

Salon, Deborah, Sinnott Murphy, and Gian-Claudia Sciara. 2014. “Local Climate Action: Motives, Enabling Factors and Barriers.” *Carbon Management* 5 (1): 67–79.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton: CRC Press.

Schafer, Joseph L., and John W. Graham. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7 (2): 147–77.

Schneider, Tapio. 2001. “Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values.” *Journal of Climate* 14 (5): 853–71.

Van der Heijden, Geert J. M. G., A. Rogier T. Donders, Theo Stijnen, and Karel G. M. Moons. 2006. “Imputation of Missing Values Is Superior to Complete Case Analysis and the Missing-Indicator Method in Multivariable Diagnostic Research: A Clinical Example.” *Journal of Clinical Epidemiology* 59 (10): 1102–1109.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.” *Statistics in Medicine* 30 (4): 377–99.

White, Ian R., Rhian Daniel, and Patrick Royston. 2010. “Avoiding Bias due to Perfect Prediction in Multiple Imputation of Incomplete Categorical Variables.” *Computational Statistics & Data Analysis* 54 (10): 2267–75.

Young, Rebekah, and David R. Johnson. 2010. “Imputing the Missing Y’s: Implications for Survey Producers and Survey Users.” In *Proceedings of the AAPOR Conference Abstracts,* 6242–48.

Zahran, Sammy, Himanshu Grover, Samuel D. Brody, and Arnold Vedlitz. 2008. “Risk, Stress, and Capacity: Explaining Metropolitan Commitment to Climate Protection.” *Urban Affairs Review* 43 (4): 447–74.

Zhang, Paul. 2003. “Multiple Imputation: Theory and Method.” *International Statistical Review* 71 (3): 581–92.

Author Biographies

Cali Curley is an assistant professor at Indiana University–Purdue University Indianapolis School of Public and Environmental Affairs. Her research is focused on environmental policy, energy policy, and local governance.

Rachel M. Krause is an associate professor at the University of Kansas School of Public Affairs and Administration. She researches urban sustainability, local governance, and climate protection initiatives.

Richard Feiock is a dynamic figure, often seen scaling walls and crushing ice. He can tread water for two days in a row. He is famous for wooing women with his sensuous and god-like guitar playing. He can pilot bicycles up severe inclines with unflagging speed, and cook thirty-minute brownies in twenty minutes. He is expert in stucco, a veteran in love, and an outlaw in Mexico. His deft floral arrangements have earned acclaim in international botany circles. Children trust him. He can hurl tennis rackets at small moving objects with deadly accuracy and he knows the exact location of every food item in the supermarket.

Christopher V. Hawkins is an associate professor in the School of Public Administration at the University of Central Florida. His research focuses on local economic development, metropolitan governance, and urban sustainability policy.