Integrated City Sustainability Database: Investigating the Advantages and Disadvantages of Bayesian Multiple-Imputation

Cali Curley, Indiana University-Purdue University, Indianapolis

Richard C. Feiock, Florida State University

Chris Hawkins, University of Central Florida

Rachel Krause, University of Kansas

Abstract:

The Integrated City Sustainability Database (ICSD) is the culmination of seven surveys on sustainability that have been combined and imputed to include a larger and more complete sample of US cities than any single survey yields or that exists elsewhere. A bayesian multiple-imputation method is employed to create an indication of the value for the missingness of each respondent across all survey questions. We used a theoretically informed, statistically supported set of predictor variables to develop a useable database. This paper describes the methodological approach to the imputation and offers a comparative between multiple methods of imputation. This paper is the methodological illustration of the ICSD.

Ninety percent of cities with population greater than 50,000 responded to at least one of the seven surveys used to create the ICSD. These surveys were conducted within an 18 month window, intended to be separate and distinct research projects. Thus, they each have varying samples and response rates. Some of these questions are identical in their phrasing and choices, other questions ask for information regarding similar concepts but change categories or the phrasing of the question, others still remain entirely unique to their survey. Therefore, the imputation of missing data points informed by responses provided to substantively related survey questions helps to improve the results drawn from quantitative analysis.

The Bayesian imputation process yielded 20 data sets pulling values from informed distributions to replace the missingness with the most likely response for each city given a set of predictor variables. The predictor variables vary for each survey question; the process of defining this predictor set is explained in detail in the paper. We compare the imputation approach utilized with other common methods for dealing with missing data and assess their relative strengths and weaknesses.

Integrated City Sustainability Database: Investigating the Advantages and Disadvantages of Bayesian Multiple-Imputation

Introduction:

Empirical research on urban politics and policy often relies on data collected through surveys of local officials. Surveys provide a relatively efficient way to collect a large amount of information that is 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 particular phenomenon by testing theory. However, missing data is a regular challenge in survey research. It often influences the selection of a statistical method of analysis, and, depending on its severity, can undermine the confidence of analysis that scholars can conduct. Nonetheless, the problems associated with missing data is one of the least acknowledged issues when conducting and reporting analysis.

A recent article in this journal by Feiock and colleagues (2014) describes the "Integrated City Sustainability Database (ICSD) as solution to the challenges associated with missing data. The ICSD combines the results of seven national surveys of city sustainability programs that were administered within an 18-month period in 2010-211 into one comprehensive national data set. On the surface the process of data harmonization yields a larger sample size, but the problems caused by missing data increase with the development of multivariate empirical models. To address this problem, the authors employ a two-stage informed multiple imputation technique. This "first generation" ICSD enabled more

confident conclusions to be drawn from the results of empirical analysis of local sustainability initiatives (Hawkins et al. 2016).

The previous version of the ICSD intended to provide one single data set available for use, we used a two-stage version of single imputation. The two stages were based on imputing within the survey recipients and then across those who did not receive the survey. Therefore taking in to consideration the different types of missingness. The first stage of imputation relied on single imputation and that imputed data was used to determine the across imputation. We then intended to provide only a single value for the missing data in the original ICSD. This way people could download the file and use the single imputation file. The end goal of the original paper was to provide one single data set for the ICSD. After much debate, the authors decided that while this is better than list-wise deletion, it is not the best treatment for missing data. Therefore, we are providing a multiple imputation version of the ICSD. This offers many advantages and advancements to the study of sustainability (specifically due to the survey research nature of the field). The user of the ICSD will now be running analysis across 20 data sets and reporting the average effects across those independent analyses. The advantages of multiple imputation is that the uncertainty of the imputation process is accounted for, and therefore we can consider the imputed value to be more reliable. In the case of single imputation, bias in the estimate is avoided but the uncertainty is misrepresented and therefore significance may be unreliable.

Our aim of this research note is to raise awareness about the treatment of missing data and demonstrate the usefulness of the ICSD in the study of urban sustainability. To accomplish this we first describe the problems associated with missing data and identify the common approaches that have been developed to address them. We then illustrates the relative performance of these approaches on data from the ICSD. We conclude with a discussion of the results and provide a set of recommendations for researchers using survey data.

Overview of Data Missingness and Techniques:

Missing data is a common challenge faced by quantitative analysts and is almost unavoidable in survey based research. There are three main reasons for the occurrence of missing data: 1.) non-coverage - the observation fell outside of the sample, 2.) total nonresponse - the observation failed to respond to the survey, and 3.) item non-response - the respondent skipped a question (Brick and Kalton, 1996). While each presents a subset of challenges for the researcher, list-wise deletion is applied as a common remedy. In this approach observations with missing values for a particular variable are simply removed from the dataset. This is the default operation in most statistical software packages. The application of list-wise deletion is widespread. Of the 1,087 studies examined by Peng et al (2006), 48 % contained missing data, 97% of which the researcher used listwise or pairwise deletion prior to conducting the empirical analysis.

Although list-wise deletion is convenient, there are two potential problems that arise from its application. The first is the potential for biased coefficient estimates. The idea of bias here is that by deleting those observations with non-response the sample mean is not representative of the population mean. This is particularly true if the missingness is non-random, and therefore there is a potential bias in the data from removing those non-respondent observations. The second problem associated with list-wise deletion is a reduction of statistical power. Statistical power stems from the size of the sample. This means that the larger the sample, the better the chances that a relationship between variables can be identified.

Deleting variables or observations for non-response is less consequential if the values are missing completely at random (MCAR). In other words, the variables or observations with missing values are not a function of other variables or observations. The second issue of reduced statistical power is a result of the decrease in sample size from listwise deletion. If the sample is still quite large or the percentage of missingness is small then this may be considered a non-issue.

A statistical approach, referred to as Little's test, has been developed to determine if the data is MCAR (Little 1988a). This test examines the explanatory power of other variables in the data to predict the missingness in the variable. If the test is significant, it suggests that some other indicator included in

the model explains some of the missingness and therefore the data is missing at random (MAR) rather than MCAR. This violates the assumptions of listwise deletion. Table 1 describes the different types of missing data.

Scholars utilize a variety of alternative techniques to list-wise deletion in order to accommodate data missingness and minimize its negative effects. Three of the most widely used approaches identified by Little (1988) are: 1.) examining the incomplete cases, 2.) imputation, and 3.) providing statistical weights to the complete cases (Little, 1988b). This research note focuses on the second strategy for resolving missingness in survey data.

Within the general category of data imputation, there are specific techniques that can be employed by researchers that vary in their complexity, strengths, and weaknesses. In this section we examine two commonly used techniques: single imputation via mean replacement and multiple imputation. These tools are designed to deal with data that is classified as either MAR or MCAR.[i]

Single imputation is a general term that describes a variety of missing data replacement techniques, including last value replacement, mean replacement and single regression replacement. A value replacement method, which can be used with panel or time series data, is referred to as the "last value" approach and 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 post-test value is missing. A second version of value replacement is to use information from similar observations, sometimes referred to as "hot-decking". It is built around a premise similar to that of propensity score matching. If observations can be matched with others that look similar across the known values for a set of variables, the missingness can be replaced by the value of its match. In categorical data, missingness is sometimes addressed by including a category for missing responses.

Mean replacement is a commonly used single imputation mechanism. This statistical approach replaces the missing observations with the mean value of the variable from all the observations in the

sample. This preserves the overall mean of each variable in a model, but not the relationships between the variables. Using mean replacement decreases the variance by holding unobserved variables to the mean. This automatically sets the sum of squared differences for these observations to zero, which likely results in an underestimation of variance. In order to identify the relationship between the dependent and mean replaced independent variables, the covariance between X and Y is divided by the variance of X. If the variance of X is under estimated it is likely that the coefficient estimator will be biased. Therefore, the relationship between X and Y will not reflect the true relationship. If the degree of missingness is small and the sample size is large this may not be problematic. However, the smaller the sample the larger effect missingness will have on these relationships.

An advanced version of single imputation is the single regression replacement method. In this approach missing values can be estimated using other 'informing' variables. The informing variables are a set of variables that help to determine the value of the missing response. These tend to be variables whose values are theoretically relevant or statistically correlated with the value of the variable that is missing. The informing variables are then used in a single regression to develop an estimate to replace the missing value. In single regression replacement, the missing value is only measured once. This creates the potential for biasing the standard errors in much the same way as mean replacement.

Multiple imputations is an extension of the single imputation regression replacement method. As its name suggests, missing values are estimated multiple times. Analyzing multiple imputed data follows three steps: the imputation of missing data, the analysis of the individual data sets, and the pooling of the results across the imputations.

The multiple imputation of missing data requires that a set of informing variables or an expected distribution of the data be used to determine the estimated value of the missing data. The ICSD uses an informed multiple imputation. These variables that inform the missing value can be determined through statistics and/or theory. If supported by theory they are described and explained thoroughly. If supported by statistics, correlation is used to determine the most relevant set of variables to use in the imputation.

These informing variables are used in regression analysis to determine an actual value representation of the missing data.

This is similar to the described process of single regression replacement above, the largest difference is that in multiple imputation this process is repeated. Rubin's (1978) formula suggests 3-10 imputations are required to produce acceptable results, however, some researchers argue the number of imputations should be similar to the percent of missing responses (Graham et al. 2007; Bodner 2008; Royston et al. 2011). This recommendation is made to ensure that the level of variation in the outcome of the imputation regression is captured in the standard errors. Therefore, the more uncertainty (missing values) the larger the number of imputations needed.

Once the data has been imputed, the researcher proceeds to analyzing the desired model across all of the imputed data sets. This means that the model is estimated across each of the imputed data sets. If the imputation occurs 20 times, then the model is estimated 20 times. Therefore, the results from the imputed data sets are pooled into one final result. The pooling process takes an average of the estimates from the imputed data sets. This incorporates the uncertainty generated by the process of imputation.

In the remainder of this research note we apply list-wise deletion, single mean replacement and multiple imputation techniques to data that makes up the Integrated City Sustainability Database. This comparative approach enables us to illustrate the relative advantages and disadvantages of each. Table 2 provides an overview and comparison of these different methods.

Description of Data used in ICSD Missingness Illustration

Table 3 presents the seven surveys on city-level sustainability initiatives that comprise the Integrated City Sustainability Database. [1] (See Feiock et al 2013 for additional detail). All seven of the surveys included in their sample all US cities with populations greater than 50,000. Survey response

[1] The ICSD is a dynamic database that is growing and, we anticipate, will continue to grow over time as new data on city level sustainability is collected. This paper utilizes the original ICSD data which establishes a 2010/2011 baseline on local sustainability initiatives. As more data is collected by the authors and others it will be added to the ICSD to enable analyses of change over time.

by cities over 50,000 was particularly strong, with approximately 90 percent of these cities responding to at least one of the seven surveys. This virtually eliminates self-selection bias and provides a unique opportunity to examine the sustainability policy, implementation, resources, obstacles, and motivations in medium and large US cities. However, although related, each survey utilized a somewhat different set of questions and received response back from a different set of cities. This creates problems in a multivariate context where models seek to draw information from across multiple surveys. However, it also creates an opportunity to examine different imputation techniques and assess the impact that they have on our ability to utilize the ICSD to examine city level sustainability policy.

One of the more complex approaches to dealing with missing data – and the approach ultimately used in the ICSD - is multiple imputation. In order to develop the multiple imputations version of the ICSD the process described in figure 1 is used. We employed a theoretic and statistically informed imputation technique. The theoretical linkages are determined by developing two concepts for each survey question a subject matter and an activity. A list of these concepts and how often they are attributed to a variable can be seen in table 7. This broad list of concepts is used to develop a very broad list of variables that may theoretically inform one another. This broad list can be considered a set of 'informing' variables. In other words, these are the variables that may help to inform the missingness of a particular variable. The list of potential informing variables is to0 long and therefore a statistical approach is used to narrow the set. The correlation between the variable being imputed and the potential informing variables must be 0.2 or greater to be included in the list of informing variables. Therefore, only variables that are theoretically and statistically relevant are retained in the imputation process. The imputation process is a regression analysis that determines potential values for a missing response based on informing variables that are statistically related in non-missing cases and deemed theoretically relevant. Twenty imputations are used to devise the results of the analysis using the multiple imputation technique. This process is repeated for all missingness across the seven surveys. This brings the sample size to the 683 number or

the cities with population 50,000 and above according to the 2010 census. Multiple imputation techniques are typically accepted for use in the dependent variable to eliminate missingness (Young and Johnson).

Dependent Variable

The primary purpose of this research note is to explain the construction of the ICSD and demonstrate the effect that the different approaches to handling missing data have on the outcomes of multivariate analyses. As such, for illustration purposes, we intentionally construct a standard empirical model that reflects many of those used in the extant literature to examine the factors that influence local action on sustainability. The dependent variable in this model is an additive index of the number of sustainability-related policies and actions that cities reported having implemented in their jurisdictions. Despite its flaws, the additive index is a particularly common dependent variable in quantitative studies of local sustainability (Portney 2003; Krause 2011; Bae and Feiock 2013). Moreover, to facilitate the illustration, we selected a dependent variable conducive to analysis using Ordinary Least Squares regression. Sixteen sustainability actions are included in this index and cluster in three primary areas: energy, transportation, and waste disposal. A full list of the actions that together comprise the index and their summary statistics are included in Table 5. These index components were all taken from a single survey in the ICSD, the Municipal Climate Protection Survey.

Independent Variables

The independent variables utilized in our model again reflect those commonly used in sustainability studies and relate to cities' motivations to engage in sustainability, the obstacles hindering their action, and a series of control variables (Krause 2013; Krause et al., 2016; Hawkins et al., 2015). Unlike the component parts of the dependent variable, which all come from a single ICSD survey, the independent variables are drawn from across multiple surveys. The EECBG Grantee Implementation Survey supplies the three motivation independent variables: energy cost savings, building a sustainable

community, and external public pressure motivations. 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.

Many of the surveys in the ICSD overlap in the inclusion of these questions. We only incorporate three of the seven surveys instead of maximizing across survey missingness, this model should minimize the degree to which missingness occurs. In other words, we are not imputing 80% of the observations for any of the variables included in the model. The point to choosing variables with less missingness is to demonstrate that in cases without high degrees of missing observations a more advanced treatment of missing data may be valued. In other words, we are giving the list-wise deletion approach its 'best chance' of success. .

Control variables include population density, per-capita income, form of government, ICLEI membership, percent minority, and citizens' educational attainment. Each of these control variables have been used in previous studies regarding sustainability policy (Krause 2010; Lubell 2009; Zahran et al. 2008; Feiock et al. 2010; Salon, Murphy & Sciara 2014). This data was collected from the US Census Bureau, the International City/County Management Association, and ICLEI Local Governments for Sustainability, and thus have near complete coverage.

Statistical Method

Ordinary least squares regression analysis is employed as the method of analysis. The goal of this analysis is not to make a theoretical claim about the relationships between these variables, but rather to examine the tradeoffs between list-wise deletion and two alternative types of imputation techniques. In order to examine these differences three variations of the same models are provided. The first model uses list-wise deletion to handle the missingness in the survey data, the second uses the single imputation

technique mean replacement, and the third uses multiple imputations, which is the approach utilized in the Integrated City Sustainability Database.

Results

Table 6, column 2 shows the results from the model utilizing *list-wise deletion*. Out of the full sample of 683 cities with population over 50,000, only 111 remain in the model after list-wise deletion removes incomplete observations (a loss of 572 observations). The results from the analysis using this approach suggests that only one variable – ICLEI membership – has a statistically significant effect on the additive policy index. None of the other independent variables are significant in explaining the variation in the dependent variable. The main concern with using this technique is the information loss resulting from the decrease in the sample size as well as potential bias that result from the process of deleting the missing variables.

The third column in Table 6 presents the results of the model using *mean replacement*. As discussed above, 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 sample size loss of 358 observations, over half of the available data. The results generated using mean replacement show several additional statistically significant relationships compared to list-wise deletion. The variable lack of political will, as well as the control variables population density and education, are now significantly related to variation in the additive policy index. The variable ICLEI membership remains significant and the magnitude of its effect is larger. Perhaps the most meaningful change in the results is that, in the second, a lack of political will has a negative statistically significant relationship to the policy index dependent variable. Cities characterized by a lack of political will towards sustainability implement approximately one half a policy less than those reporting more political will in their city governments. However, the concern associated with mean replacement is that the relationships between the variables will not be maintained due to underestimates of the standard deviation. Therefore,

even though these variables are significant in the mean replacement approach, the resulting p-values should be interpreted with greater caution.

The results from the analysis performed using informed *multiple imputation* result in a slightly different combination of statistically significant variables in the model, when compared to the other two approaches. The motivation to build a sustainable community is significant in this model and positively associated with the policy index. Cities with a higher motivation to build sustainable communities have an additive index of sustainable policies a third of a point higher than cities with lower motives to build sustainable communities, all else being equal. ICLEI membership and lack of political will remain statistically significant, however, the control variables that were significant under mean replacement have faded in significance. The magnitude of the ICLEI membership variable suggests that ICLEI members, all else being equal, have one more sustainability action in place than non-ICLEI members. The magnitude of the variable lack of political will decreased by a small amount compared to mean replacement. Multiple imputation is more typically used for dependent as well as independent variables than mean replacement, this increases the sample size from 325 to 683. The standard errors in multiple imputation incorporate the uncertainty from the 20 imputation results.

This illustration depicts that mean replacement may be an improvement upon list-wise deletion, but it may not tell the entire statistical story.

Discussion and Conclusion

All three of these techniques to dealing with missing data have been used throughout the literature. Each may have its advantages and disadvantages; however, using the wrong method may provide inaccurate, biased, or inappropriate null findings. The Integrated City Sustainability Database provides an opportunity to examine the implications of various treatments of missing data and its effect on the results of analysis. This examination also demonstrates that mistreating missingness in analysis can contribute to null findings.

We have selected to utilize informed multiple imputation for the ICSD because the capacity to perform statistical analysis is greatly improved by increasing the sample size (power). In addition, the data can be described as both MAR or MCAR. We can attribute a large degree of the missingness simply due to the random selection of survey recipients. This translates into a larger sample size, less bias by avoiding deletion, and the ability to interpret the data as if it were not missing. The disadvantages to this approach is that analysis takes longer, there is a lack of descriptive information for each individual data set, and the coding process is a little more involved.

It is our hope that scholars begin to treat missing data more explicitly and openly. We recognize that the coding process for utilizing imputed data may be more than some care to learn. Therefore, we have included, in an appendix, some basic multiple imputation code and description to aid in the utilization process.

References

Abayomi, Kobi, Gelman, Andrew, and Levy, Marc. 2008. "Diagnostics for Multivariate Imputations." Journal of the Royal Statistical Society 57.3:273-291.

Allen, Tammy D., Eby, Lillian T., Lentz, Elizabeth 2006. "Mentorship Behaviors and Mentorship Quality Associated with Formal Mentoring Programs: Closing the Gap Between Research and Practice." Journal of Applied Psychology 91.3:567-578

Andridge, Rebecca R. and Roderick J.A. Little (2010). "A Review of Hot Deck Imputation for Survey Non-Response" Int Stat Rev. 2010 April ; 78(1): 40–64

Andridge, Rebecca R., and Little, Roderick J.A. 2010. "A Review of Hot Deck Imputation for Survey Non-response." International Statistical Review 78.1:40-64.

Bae, Jungah, and Richard Feiock. "Forms of government and climate change policies in US cities." *Urban Studies* 50.4 (2013): 776-788.

Betsill MM. Mitigating climate change in US cities: opportunities and obstacles. Local Environ. 6(4), 393–406 (2001).

Bodner, Todd E. (2008) "What improves with increased missing data imputations?" *Structural Equation Modeling*: *A Multidisciplinary Journal* 15: 651-675.

Brick, JM and G. Kalton, 1996. "Handling Missing Data in Survey Research" Stat Methods Med Res, September 1996 vol. 5 no. 3 215-238

Deborah Salon, Sinnott Murphy & Gian-Claudia Sciara (2014) Local climate action: motives, enabling factors and barriers, Carbon Management, 5:1, 67-79, DOI: 10.4155/ cmt.13.81

Donders, A Rogier T. van der Heijden, Geert J.M.G. Stijnen, Theo, and Moons, Karel G.M. (2006). "Review: A Gentle Introduction to Imputation of Missing Values." Journal of Clinical Epidemiology 59:1087-1091.

Downey, Ronald G., and King, Craig V. 1998. "Missing Data in Likert Ratings: A Comparison of Replacement Methods." The Journal of General Psychology 125.2:175-191.

Feiock RC, Francis N, Kassekert T. Explaining the adoption of climate change policies in local government. Proceedings of the Pathways to Low Carbon Cities Workshop. Hong Kong, China, 13–14 December 2010

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

Gallimore, Jonathan M., Brown, Barbara B., and Werner, Carol M. 2011. "Walking Routes to School in New Urban and Suburban Neighborhoods: An Environmental Walkability Analysis of Blocks and Routes." Journal of Environmental Psychology 31:184-191.

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: 206–213.

Jones, Michael P. 1996. "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." Journal of the American Statistical Association 91.433:222-230.

King, Gary, Honaker, James, Joseph, Anne, and Scheve, Kenneth. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." American Political Science Review 95.1:49-69.

Krause R. Policy innovation, intergovernmental relations, and the adoption of climate protection initiatives by U.S. cities. J. Urban Aff. 33(1), 45–60 (2010).

Krause RM. Political decision-making and the local provision of public goods: the case of municipal climate protection in the US. Urban Studies 49(11), 2399–2417 (2011)

Krause RM. "The Motivations Behind Municipal Climate Engagement: An Empirical Assessment of How Local Objectives Shape the Production of a Public Good" Cityscape Vol. 15, No. 1, Climate Change and City Hall (2013), pp. 125-141

Little, Roderick J. A. 1988b. "Missing-Data Adjustments in Large Surveys" Journal of Business & Economic Statistics. July 1988, Vol.6, No.3

Little, Roderick J. A., and Rubin, Donald B. 1987 "Statistical Analysis With Missing Data." John Wiley & Sons, New York.

Little, Roderick J.A. 1988a. "A Test of Missing Completely at Random for Multivariate Data with Missing Values" Journal of the American Statistical Association. December 1988, Vol. 83, No. 404

Lubell M, Feiock R, Handy S. City adoption of environmentally sustainable policies in California's Central Valley. J. Am. Plann. Assoc. 75(3), 293–308 (2009)

Miyama, Eriko, and Managi, Shunsuke. 2014. "Global Environmental Emissions Estimate: Application of Multiple Imputation." Environmental Economics & Policy Stuies 16:115-135.

Park, Joohyung and Ha, Sejin. 2011. "Understanding Pro-Environmental Behavior." Journal of Retail & Distribution Management 40.5:388-403.

PortneyRubin, D.B. 1987. "Multiple Imputation for Nonresponse in Surveys." John Wiley & Sons, New York.

Rubin, Donald B. (1996). "Multiple Imputation After 18+ Years" Journal of the American Statistical Association, Vol. 91, No. 434, 473-489.

Ryff, Carol D. and Keyes, Corey Lee M. 1995. "The Structure of Psychological Well-Being Revisited." Journal of Pesonality and Social Psychology 69.4:719-727.

Schafer, Joseph L. 1997. "Analysis of Incomplete Multivariate Data." Chapman & Hall, Boca Raton, FL.

Schafer, Joseph L., and Graham, John W. 2002. "Missing Data: Our View of the State of the Art." Psychological Methods 7.2:147-177.

Schneider, T. 2001. "Analyisis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values." Journal of Climate 14.5: 853-871

van der Heijden, Geert J.M.G., Donders, A. Rogier T., Stijnen, Theo, Moons, Karel G.M. 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 59:1102-1109.

Zahran S, Grover H, Brody SD, Vedlitz A. Risk, stress, and capacity: explaining metropolitan commitment to climate protection. Urban Aff. Rev. 43(4), 447–474 (2008)

Zhang, Paul. 2003. "Multiple Imputation: Theory and Method." International Statistical Review 71.3:581-592.

Table 1: Types of Missingness Overview		
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, i.e. after observed values are accounted for, missingness is randomly distributed.	Missing observations are dependent upon unobserved values, and missingness cannot be accounted for by controlling for observed data.

Table 2: Techniques of Imputation*

TECHNIQUES	List-wise Deletion (Complete Case Analysis)	Single Imputation		Multiple Imputation
		Mean Replacement (Mean Substitution)	Single Regression Replacement	
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**	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 dataset; 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 is 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 was not missing.
References				
Method Exploration	Jones 1996, 223; Schafer and Graham 2002, 155.	Downey and King 1998; Shafer and Graham 2002, 159.	Donders et al. 2006, 1088-1089; Schneider 2001; van der Heijden et al. 2006;***	Donders et al. 2006, 1089; King et al. 2001; Rubin 1987; Schafer 1997; Zhang 2003;
Application	Park and Ha 2011, 394; Ryff and Keyes 1995, 722.	Allen et al. 2006, 572; Gallimore et al 2011, 186-187		Abayomi et al. 2008; Fox and Swatt 2009; Miyama and Managi 2014;

*Additional missingness reference can be found in Schafer and Graham 2002, 151.

**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.

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

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 over 10,000	442	26.6
EECBG Grantee Implementation Survey	970 municipal governments receiving EECBG awards, including all cities over 30,000	747	77
Implementation of Energy Efficiency and Sustainability Programs	1,180 cities: all with populations over 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 over 50000	263	44
Municipal Climate Protection Survey	664 cities with populations over 50000	329	49.5
Municipal Government Questionnaire	425 cities with populations over 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.

Table 4: Summary Statistics						
	Variable Name	Observations	Mean	Std Dev.	Min	Max
DV	Additive Policy Index	325	7.7169231	2.91863	1	16
C	Population Per Sq. Mile	690	3925.8988	3558.5722	171.2	51810
C	Per Capita Income	656	27017.646	8437.1326	10739	81198
C	ICLEI Membership 2010	713	.29312763	.45551601	0	1
C	Council	675	.6418148	.47992105	0	1
C	Mayor-Council	675	.34518519	.47578118	0	1
C	Percent Minority	690	44.43029	22.563246	8.4000015	99.199997
C	Percent Bachelors or more	656	30.4625	13.776391	4.6999998	79.300003
M	Reduced energy cost	469	2.7398721	.51112494	0	3
M	Sustainable Communities	468	2.2393162	.73632658	0	3
M	Public Pressure	457	1.0262582	.81607332	0	3
O	Staff Capacity	453	1.1743929	.72445121	0	2
O	Lack of Information	451	.75831486	.63710512	0	2
O	Lack of Political Will	333	.89189189	.67687292	0	2

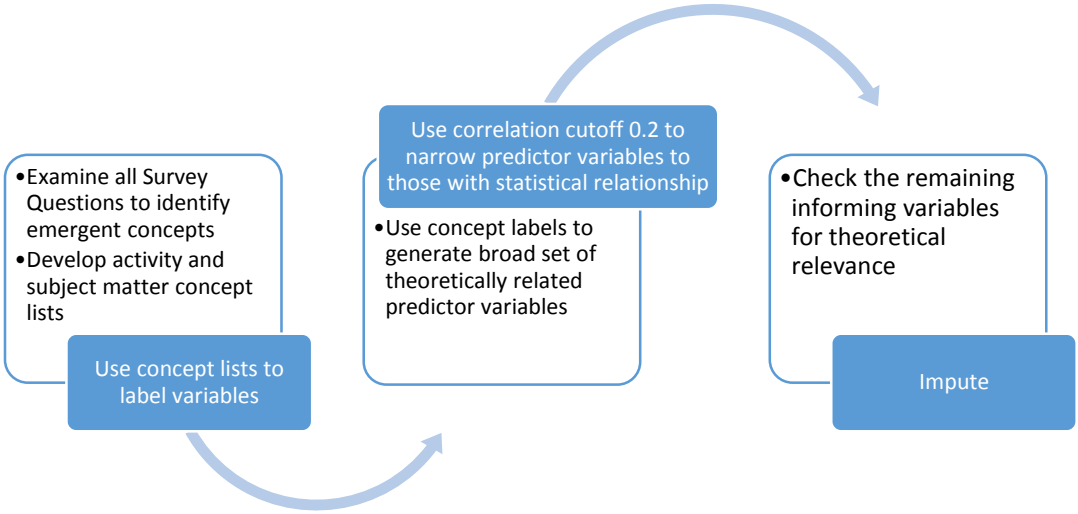
Table 5: Additive Policy Index					
Policy Name	Obs	Mean	Std	Min	Max
City Building High efficiency light bulbs	325	0.62	0.31338455	0	1
LED Streetlights	325	0.31692308	0.33267028	0	1
Green City Vehicles	325	0.44769231	0.26417101	0	1
Bike lane/ trails	325	0.66923077	0.38567044	0	1
Efficient city appliances	325	0.31384615	0.46477026	0	1
Renewable electric city buildings	325	0.37230769	0.48416521	0	1
Green city buildings	325	0.4	0.49065338	0	1
EE residential info	325	0.76923077	0.42197474	0	1
EE residential incentive	325	0.32307692	0.46837295	0	1
EE incentive development	325	0.23692308	0.42585036	0	1
EE regulate building	325	0.21846154	0.41383941	0	1
Commute incentive city staff	325	0.32307692	0.46837295	0	1
vehicle anti-idle policy	325	0.48615385	0.50057896	0	1
Public Transit services	325	0.60615385	0.48935487	0	1
Public transit incentive	325	0.25538462	0.43674963	0	1
Yard waste is composted or mulched	325	0.62769231	0.48416521	0	1
Recycling is picked up curbside	325	0.91076923	0.28551614	0	1
Methane Capture	325	0.13384615	0.26015008	0	1

Table 6: Comparison of techniques using the model results

		List-wise Deletion		Mean Replacement		Multiple Imputation	
		Coeff	Standard Error	Coeff	Standard Error	Coeff	Standard Error
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 Psq mile	0	0.000	0.0001*	0.000	0.000	0.000
C	Percapita income	0	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
	n	111		325		683	
		Adj R2	0.0906	Adj R2	0.1814	Prob >F	0

Table 7: General Concept Description			
General Concept	**Category**	**Description/Keywords**	**Count***
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, American Resource and Recovery Act (ARRA), stimulus	109
Energy	Subject Matter	Energy	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 (publicly-owned building, purchase (credits), incentives, utility retrofit)	128
Government Planning	Activity	goal, inventory from city government operations	9
Infrastructure	Activity	own..operate xxx facility	46
Inter-department	Activity	Coordinate within the city	46
Inter-governmental	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
*Represent number of variables characterized as general concept			

Figure 1: Process Flow of Informed Multiple Imputation



Appendix: STATA Multiple Imputation Code

***The following is Multiple Imputation code as related to using the ICSD imputed data for STATA. Please see (either appendix or website) for details on what is currently available for public use.

**Read in data as usual.

**Import the data as an imputed file or ice object

```
mi import ice, automatic
```

** Get a list of all commands for mi estimation, any of these commands can be used to analyze data as you normally would.

```
help mi estimation
```

** In order to use linear regression with continuous DV and an X variable. Options are typically added before the colon

```
mi estimate : regress Y_variablename X_variablename
```

**Logistic regression with dichotomous DV and an X variable and code to set a variables value to dichotomous

```
recode variablename 1 = 0 2 = 1
```

```
label define variablename 1 "Yes" 0 "no", replace
mi estimate : logistic Y_variablename X_variablename
```

**Ordinal-response regression

```
mi estimate : ologit Y_variablename X_variablename
```

**Multinomial logistic regression, items with more than 2 response options that are not ordered.

```
mi estimate : mlogit Y_variablename X_variablename
```

**In order to look at means across imputations or proportion of responses across imputations use the following code. These statistics are how to calculate the variance across imputations (level of uncertainty).

```
mi estimate : mean variablename

mean variablename if _mi_m == 0

mi estimate : proportion variablename

proportion variablename if _mi_m == 0
```

```
* Here's some code to run the individual regressions, save the
* R-squares, and summarize them for you.

* Define loop
qui sum _mi_m, detail
local imax = r(max)

* Create empty matrix for R-squared values
mata:  R = J(`imax',1,.)

* Run regressions, save R-squared
foreach j of numlist 1/`imax' {
        qui reg Y_variablename X_variablename if _mi_m==`j'   // the only thing to change is the
regression variables in this line //
        local r2 = e(r2)
        mata:  R[`j',1] = `r2'
}

mata:  mean = mean(R)
mata:  median = mm_quantile(R,1,.5)
mata:  st_numscalar("r2mean", mean[1,1])
mata:  st_numscalar("r2med", median[1,1])
```

```
di "The mean R-squared is: " r2mean
di "The median R-squared is: " r2med
```

[i] However, they are not well equipped to deal with data that is not missing at random (NMAR or MNAR). This category of missingness typically represents strategic non-response. It is problematic for most models of imputation because there is a specific underlying motivation for the missingness. This is especially troubling when there is no explanation captured in the existing data.