Dealing with Missing Data: A comparative exploration of approaches utilizing the Integrated City Sustainability Database

Cali Curley, Rachel Krause, Richard Feiock, and Chris Hawkins

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 added value 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 non-response 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.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Nos. 1461526/1461506/1461460. The University of Kansas Center for Research Methods and Data Analysis provided valuable assistance in the imputation process, as did our graduate research assistant XXX.

Introduction

This article seeks to raise awareness about the treatment of missing data in urban studies research generally. Much of the evidence base of empirical research on urban politics and policy relies on data collected through surveys of local officials. Surveys provide a relatively efficient way to collect large amounts of 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 particular phenomenon by testing theory. However, missing data is a common and significant challenge in survey-based research. It often influences the selection of a statistical method of analysis, and, depending on its 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 occurs for three reasons: 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 in the survey (Brick and Kalton, 1996). Although missing data resulting from each of these causes presents a subset of 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 simply removes observations with missing values for any variable included in analysis and despite its deficiencies, listwise deletion iswidely sued. Peng et al (2006) examined 1,087 published studies in education and psychology, of which 48% contained missing data. Within that subset, the researcher(s) used listwise or pairwise deletion 97% of the time prior to conducting their empirical analysis.

This paper offers a description of the basic classifications of missing data, the specific problems associated with each of them, and the common approaches that have been developed to address them. This is followed by a comparative 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 discusses their relative performance in analysis. In conclusion, a discussion of the missing data techniques based on the analysis results is offered to provide a set of recommendations for researchers using survey data.

Overview of Missing Data

Three classifications of missing data that 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 below provides a brief overview of the discussion that follows.

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 a strategic choice or a function of a captured or uncaptured variable. 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 reason that the respondent may have wanted to avoid answering that question.Utilizing Little's MCAR test is one piece of information that can help inform the decision as to whether data is truly MCAR or not. The application of this test is discussed in the discussion of listwise deletion below.

Data that are MAR are characterized by the fact that the value of the missingness can be predicted using observed variables. The observed variables may or may not be related to the cause of the missing value. Typically speaking, when a missing value can be explained by other observed data the missingness is determined to be randomly distributed by controlling for those explanatory variables. An example of this might be when an individual intentionally skips the question asking about his/her income in a survey; but the researcher has observed values for the respondent's employment status, education level, and experience at their current job. In this context, the value of the missing data is dependent on the value of observed responses.

Missing not at Random (MNAR) or non-ignorable missingness occurs when the explanation for why observations of a variable are missing is not available. Moreover, the researcher cannot approximate the missing values because the values of other relevant variables that could be used to do this are also not observed. Consider the previous example, if the observed data did not include education level or experience it would be very challenging to determine an expected value of the respondents' income.

[Insert table 1 about here]

The treatment of these missing observations has important ramifications for scholarship. The different approaches used to deal with missing data make specific mathematical assumptions about the type of missing data that they are handling. Their misuse may invalidate empirical results.

Approaches to handling missing data:

Scholars utilize a variety of alternative techniques in order to accommodate missing data and minimize its negative effects. Three of the most widely used approaches identified by Little

(1988) are: 1) examining the incomplete cases, 2) imputing values for missing data, and 3) providing statistical weights to complete cases (Little, 1988b). Within the general category of data imputation, there are specific techniques that vary in complexity as well as their relative strengths and weaknesses. In addition to listwise deletion, two commonly used techniques, single imputation via mean replacement and multiple imputation, are examined and compared.

Listwise deletion, the default approach to handle missing data is convenient, but it creates two problems for data analysis. The first is that if the missingness is MAR or MNAR, then the deletion of observations with missing values may lead to the sample mean being unrepresentative of the population mean. This is particularly problematic if the missingness is non-random because, if respondents are strategically opting out of responding to a question based on some unobserved trait, it may skew the sample mean. Non-response, especially strategic non-response, may result in a sample that is not representative of the population, violating assumptions of hypothesis testing, ordinary least squares, and other statistical analysis that suggests the population mean and the sample mean should be similar. The second issue with listwise deletion is that it reduces the sample size and thus the statistical power of the sample may be correspondingly reduced. Smaller samples are more likely to generate null results that might otherwise not be null with a larger sample.

Two conditions must be met for listwise deletion to be an appropriate treatment for dealing with missing data: the missingness must be MCAR and the sample is large even after the deletion occurs. Deleting observations for non-response is less consequential if the values are MCAR, because if missingess is completely random the data deleted would also be random and it would thus not cause the loss of important variation. A statistical approach, referred to as Little's test, has been developed to determine if the data can be classified as MCAR (Little

1988a). This test examines the explanatory power of other variables in the data to predict the missingness in the target 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. If the sample remains quite large after listwise deletion, or the percentage of missingness is small, then reduction of statistical power is not an issue.

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, sometimes referred to as "hot-decking," uses information from similar observations to replace missing data. 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, missing ones 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 replaces missing observations with the mean value of the variable from observed responses in the sample. This preserves the overall mean of each variable but not the variation of the sample. Therefore, using mean replacement decreases the potential variance by holding unobserved variables to the mean. This automatically sets the sum of squared differences for these observations to zero, which 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 estimated relationship between X and Y may not reflect the true relationship. There are cases where this technique may be appropriate, specifically when the degree of missingness is small and the sample size is large. 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. In this approach missing values can be estimated using observed variables that can predict the value of the missing response. These are referred to as informing variables since they inform the value that the missing variable likely would have taken had the respondent answered that question. The variables that are used to predict/inform missing responses tend to be variables whose values are theoretically relevant or statistically correlated with the variable that is missing. The informing variables are used in a single regression to develop an estimated value to replace the missing value. This allows the value of missing observations to vary based on responses to the informing variables. As an example, consider a scholar attempting to explain wages for a sample of respondents. However, her data contains several missing responses to a key variable associated with a survey question about perceived experience. The scholar knows that years at the current place of employment as well as education level are correlated with the observed values for perceived experience. Therefore, the scholar uses those two variables in a regression equation to develop a best guess for the value that the respondent would have given for perceived experience. This helps illustrate that the point of imputation is not necessarily 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 the inference of the desired model (Rubin 1987, 1996).

In single regression replacement, this missing value is only measured once, which creates the potential for biasing the standard errors similar to mean replacement. This is because the observed and missing variables are given the same weights in the regression estimating wage; in other words, there is no distinction between an estimated value and the true reported value. This is an important caveat of single regression replacement, because it does not mathematically consider the inherent uncertainty in the prediction of the missing value. Therefore, the desired analysis may be influenced by the predicted missing values more than the true observed data.

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: the imputation of missing data, the running of independent statistical analysis on the resulting individual data sets, and the pooling of the results across the imputations.

In order to impute missing data, several variables that predict or inform the missing values must be identified. These predicting variables should be theoretically related or statistically correlated to the missing values. After establishing the theoretic and/or statistical relationship, these variables can be used in multiple imputation to determine the values of the missing data. The model used to impute the missing values should match the type of data being generated. For example, if observations are being imputed for a continuous variable, an appropriate model such as ordinary least squares regression should be used; if the variable is binary, a model such as logit or probit is appropriate to impute their values.

The first step of multiple imputation is similar to the single regression replacement

method described above. However, in multiple imputation this process is repeated in order to incorporate the uncertainty in the prediction process, which is not captured in single imputation. Therefore, multiple imputation creates numerous data sets, each containing somewhat different estimates of the missing values. Rubin's (1978) formula suggests 3-10 imputations are necessary to produce results that incorporate enough variation in the prediction process; however, other 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 ensures that the level of variation in the prediction of missing values is large enough to be captured in the standard errors in the actual analysis of interest. This concludes the process of imputing. However, the key difference between single regression replacement and multiple imputation is in the analysis of the data.

Once the data is imputed, the researcher has a number of different data sets, so theory-based models can be estimated and tested simultaneously with each set of data. Many statistical programs enable data to be specified as imputed. For example, in STATA multiple imputed data must be set as *mi data*, which marks each imputation as a separate data set with its own estimation of the missing values. With this designation, , the theory-based model is estimated individually across each of the imputed data sets in the background (i.e., if the imputation occurs 20 times, then the model is estimated 20 times). The results are then pooled together and the pooled output reported. 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. Although the model outputs are the pooled coefficients from the 20 individual analyses, the results can be interpreted in the same manner as one would in a normal setting.

TABLE 2 HERE

Description and Illustration of ICSD Missing Data

To illustrate the relative advantages and disadvantages of each while simultaneously describing this novel database, listwise deletion, single mean replacement, and multiple imputation techniques are compared using data from the Integrated City Sustainability Database (ICSD).

A recent article in this journal by Feiock and colleagues (2014) describes the "Integrated City Sustainability Database (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 2010-211 into one comprehensive national data set. Table 3 presents basic information on the seven independently administered surveys.[1] The process of survey harmonization yields 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 utilizes 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 imputation technique, which produced a single unified data set through a two-stage version of single imputation. The first stage imputed missing data within each completed survey and the second used this data to impute across surveys, taking in to consideration the different types of missingness. This process generates a

[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. The original ICSD 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.

single unique value for each missing observation in the original ICSD and results in one single data set for the ICSD. This facilitates accessibility since users can download and use a single file of imputed data. This "first generation" ICSD is a significant advancement that enables more confident conclusions to be drawn from the results of empirical analysis of local sustainability initiatives (Hawkins et al. 2016). It provides imputed data for a large set of cities including smaller cities and has been used extensively in urban research.

While the two-stage single imputation approach of the first generation database is a significant improvement over listwise deletion, for the cities over 50,000 population that were included in the sample frames for all seven surveys, it can be improved through the process of multiple imputation. The primary advantage of multiple imputation is that the uncertainty of the imputation process is accounted for, and therefore the imputed value is considered 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. The second generation ICSD described here compliments the first generation database by providing a multiple imputation version for a subset of 683 ICSD cities with populations of 50,000 or more, which offers advantages and advancements to the study of sustainability.

Each of the seven surveys in the ICSD included all US cities with populations greater than 50,000 in their sample, with some also including smaller municipalities. Survey response by cities over 50,000 was particularly strong, with 90 percent of these cities responding to at least one survey. This virtually eliminates self-selection bias among this sub-sample 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 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.

TABLE 3 HERE

Figure 1 summarizes the process used to identify the theoretic and statistically relevant informing variables to use in the multiple imputation employed in the second generation ICSD. The theoretical linkages are determined via a process of developing two "general concepts" – one related to the "activity" and "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 "Energy".

[Figure 1 About Here]

A list of these concepts and how often they are attributed to variables in the surveys is presented in Table 4. This concept list is used to develop a broad list 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 and therefore a statistical approach is used to narrow the set. With the objective of identifying a small enough number of informing variables to enable statistical conversion, we selected 0.2 as the minimum correlation between the variable being imputed and the potential informing variables. As a result of this process, only variables that are theoretically and statistically relevant are retained as predictors, resulting in an average of 95 informing variable for each target variable.

A distribution of the non-missing cases is used to determine the expectation of the distribution for missing responses. For example, if the non-missing responses are normally distributed the imputed responses will maintain a normal distribution. The distribution assigned is variable specific. Twenty imputations are used to generate the results of the analysis using the multiple imputation technique. 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 is generated for each of the 1,010 variables in the ICSD.

TABLE 4 HERE

A Comparison of Approaches Using the ICSD

We utilize the ICSD in its raw and two imputed forms to demonstrate the relative performance of each of the three approaches to dealing with missing data: listwise deletion, value replacement, and multiple imputation. For illustration purposes, we construct a generic empirical model that reflects many of those used in the extant literature to examine the factors that influence local action on sustainability.

Dependent Variable

The dependent variable in this model is an additive index of the number of environmental sustainability-related policies and actions that cities reported having implemented in their jurisdictions. Despite its flaws, the additive index is a common dependent variable in quantitative studies of local sustainability (Portney 2003; Krause 2011; Bae and Feiock 2013). 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".

TABLE 5 HERE

Independent Variables

The independent variables reflect many commonly used in sustainability studies and relate to cities' motivations to engage in sustainability, obstacles hindering their action, and a series of control variables (Krause 2013; Krause et al., 2016; Hawkins et al., 2015). The independent variables are intentionally drawn from different ICSD component surveys. The "EECBG Grantee Implementation Survey" supplies the three motivation independent variables: 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.[2]

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). There data were collected from the US Census Bureau, the International City/County Management Association,

[2] 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 list-wise deletion approach its 'best chance' of success.

and ICLEI Local Governments for Sustainability, and thus have near complete coverage.

TABLE 6 HERE

Results

Ordinary least squares regression analysis is employed as the method of analysis. Our purpose here is not to make a theoretical claim about the relationships between these variables, but rather to examine the tradeoffs between using different approaches to deal with missing data. In order to examine these differences, three identical models are used to estimate the impact of the different missingness treatments. The first model uses listwise deletion to handle the missingness in the survey data, the second uses the single imputation mean replacement technique, and the third uses multiple imputations, which is the approach utilized in the second generation Integrated City Sustainability Database.

Table 7, column 2 reports the results from the model utilizing *listwise deletion*. Only 111 of the 683 cities with population over 50,000 remain in the model after listwise deletion removes incomplete observations (a loss of 572). The results using this approach indicate that only one variable – ICLEI membership – has a statistically significant effect on the additive policy index. The information loss resulting from the decrease in the sample size and potential bias from deleting the missing variables are major concerns with using this technique.

The third column in Table 7 presents the results of the model using *mean replacement*. 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, because mean replacement is not always accepted for use in the dependent variable. The results generated using

mean replacement show several additional statistically significant relationships compared to listwise 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, using mean replacement, 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, 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 to 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 is significant and positively associated with the policy index. ICLEI membership and lack of political remain statistically significant, however, the magnitude of both decrease slightly compared to the other models. The standard errors in multiple imputation incorporate the uncertainty from the 20 imputation results giving us confidence in the resulting p-values.

Discussion and Conclusion

The three techniques of listwise deletion, value replacement, and multiple imputation,

have been used throughout the literature to address missing data. Each is associated with particular advantages and disadvantages; however, depending on the nature of the missingness, 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. The second generation ICSD database 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 it was not missing. In addition, this technique is applicable to data that is either MAR or MCAR. A large degree of the missingness in the ICSD can be attributed to the random selection of survey recipients, 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. The disadvantages to analysis using multiple imputation is that the generation of the data is more complicated, the analysis takes longer, and involves more coding. Also multiply imputed data is conducive to the generation of standard descriptive statistics, including things like grand variable means, and basic model fit indicators like R^2.

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 my many journals (Nosek et al. 2015). It is our hope that urban scholars begin to treat missing data more explicitly and openly. We have included, in an appendix, some basic multiple imputation code and description to aid in the utilization process. In the near future we expect to make the multiply imputed data included in the first and second generation ICSD available to researchers and sustainability practitioners. In the meantime, select variables from the first

generation ICSD are available at http://localgov.fsu.edu/ICSD/.

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, R. & Bae, J. (2011). Politics, Institutions, and Entrepreneurship: City Decisions Leading to Inventoried Green House Gas Emissions. *Carbon Management 2(4*), 443-453.

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.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni 2015. Science 348 (6242): 1422-1425.

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

Peng CYJ, Harwell M, Liou SM, Ehman LH. Advances in missing data methods and implications for educational research. In: Sawilowsky SS, editor. Real data analysis. Charlotte, North Carolina: Information Age Pub; 2006. pp. 31–78

Portney, Kent E. 2013. *Taking Sustainable Cities Seriously: Economic Development, the Environment, and Quality of Life in American Cities*2nd edition. Cambridge, MA: MIT Press.

Rubin, 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.

Young, Rebekah and David R. Johnson. 2010. ‘‘Imputing the Missing Y’s: Implications for Survey Producers and Survey Users.’’ P. 186 in AAPOR 2010 Conference Abstracts. Retrieved June 2016. (http://www.aapor.org/AM/Template.cfm?Section=65th_Annual_Conference&Template =/CM/Con tent Display.cfm&ContentID=2385.)

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.

Appendix: STATA Multiple Imputation Code

***The following is Multiple Imputation code as related to using the ICSD imputed data for STATA.
***Please see http://XXXXXX/ 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
```

Table 1: Overview of Types of Missing Data		
Missing Completely at Random (MCAR)	**Missing at Random (MAR)**	**Missing Not at Random (MNAR, non-ignorable)**
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; missingness cannot be accounted for by controlling for observed data.

Table 2: Techniques of Imputation*

TECHNIQUES	Listwise 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: 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, 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 (publicly-owned building, purchase (credits), incentives, utility retrofit)	128
Government Planning	Activity	goal, inventory from city government operations	9
Infrastructure	Activity	own operate, 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			

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: 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 7: Comparison of techniques using the model results							
		Listwise 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

Figure 1: Process Flow of Informed Multiple Imputation