DUKE ENVIRONMENTAL AND ENERGY ECONOMICS WORKING PAPER SERIES
organized by the
NICHOLAS INSTITUTE FOR ENVIRONMENTAL POLICY SOLUTIONS
and the
DUKE UNIVERSITY ENERGY INITIATIVE

Creating Linked Datasets for SME Energy-Assessment Evidence Building: Results from the U.S. Industrial Assessment Center Program

Nicole M. Dalzell*
Gale A. Boyd§
Jerome P. Reiter‡

Working Paper EE 17-02

* Department of Mathematics and Statistics, Wake Forest University
§ Social Science Research Institute, Duke University
‡ Department of Statistical Science, Duke University

Acknowledgments
This work was supported by NSF Grant SES 1131897, and by the Duke University Energy Initiative Energy Research Seed Fund, with co-funding from the Information Initiative at Duke and was prepared while the authors were Special Sworn Status researchers at the Triangle Research Data Center, a member of the Federal Statistical Research Data Center Network. All results have been reviewed by the Census Bureau to ensure that no confidential information is disclosed. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau.

Creating linked datasets for SME energy-assessment evidence-building: results from the U.S. Industrial Assessment Center Program

Nicole M. Dalzell*
Department of Mathematics and Statistics, Wake Forest University
Gale A. Boyd
Social Science Research Institute, Duke University
Jerome P. Reiter
Department of Statistical Science, Duke University

June 15, 2017

Abstract

Lack of information is commonly cited as a market failure resulting in an energy-efficiency gap. Government information policies to fill this gap may enable improvements in energy efficiency and social welfare because of the externalities of energy use. The U.S. Department of Energy Industrial Assessment Center (IAC) program is one such policy intervention, providing no-cost assessments to small and medium enterprises (SME). The IAC program has assembled a wealth of data on these assessments, but the database does not include information about participants after the assessment or on non-participants. This study addresses that lack by creating a new linked dataset using the public IAC and non-public data at the Census Bureau. The IAC database excludes detail needed for an exact match, so the study developed a linking methodology to account for uncertainty in the matching process. Based on the linking approach, a difference in difference analysis for SME that received an assessment was done; plants that received an assessment improve their performance over time, relative to industry peers that did not. This new linked dataset is likely to shed even more light on the impact of the IAC and similar programs in advancing energy efficiency.

*Corresponding Author. *Email:* dalzelnm@wfu.edu

1 Introduction

The notion that energy demand suffers from the energy-efficiency gap is pervasive in the energy economics literature. Specific emphasis is placed on the apparent failure to fully implement cost effective technologies that would provide a needed energy service at lower levels of energy inputs (e.g., Allcott and Greenstone, 2012; Gerarden et al., 2015). Energy production and use also suffers from the potential for environmental externalities (air pollution, climate change, etc.) that may not be fully internalized in the price of energy. In short, if the energy-efficiency gap exists then it has both a private and social cost. Policymakers are interested in identifying market barriers that could be effectively removed to address the gap (Jaffe and Stavins, 1994).

One commonly cited potential market failure resulting in higher-than-optimal energy consumption is a lack of complete information regarding energy efficient technologies, or the high cost of acquiring this information. To address this potential barrier to energy efficiency, the U.S. Department of Energy implemented the Industrial Assessment Center (IAC) program in 1976. The IAC program provides no-cost energy assessments (or "energy audits") conducted by teams of university faculty and students who visit manufacturing plants to assess productivity, energy use and efficiency, and waste. Visits result in a report detailing all cost-saving opportunities identified during the assessment (U.S. Department of Energy, 2016). The IAC program targets small and medium sized enterprises (SME) under the assumption that such enterprises have fewer resources to dedicate to energy management than their larger counterparts. Additionally, work in Sweden (Thollander et al., 2007) and Portugal (Henriques and Catarino, 2016) suggest that there are difficulties in encouraging energy efficiency in SME.

The IAC program has conducted over 17,824 assessments and has made more than 135,918 individual recommendations through May 2017 (U.S. Department of Energy, 2016). Information about each assessment, including the provided recommendations, is maintained in a publicly available database. This means the IAC data are available to external researchers for conducting energy related research. For example, the IAC database has been used to statistically model decision making of plant reaction to IAC assessment reports, i.e., whether a plant chooses to implement a recommendation or not. Boyd (2001) and An-

derson and Newell (2004) both use discrete choice statistical models to examine the factors that influence the implementation choice. Perroni et al. (2016) compute enterprise level efficiency using DEA and SFA methods and determine that those SME that adopt energy efficiency are not statistically more productive in general; see Murillo-Zamorano (2004) for a review of the methods.

While the IAC data may be useful to understand factors in SME decision making regarding energy efficiency implementation, the data are not designed to assess long run energy savings that may, or may not, result from an assessment. Specifically, the data do not contain information about a plant post-assessment. For example, examining whether the implementation costs of recommendations are higher than predicted requires follow-up information on assessed plants. Examining whether the estimated savings from the assessment are fully realized, or if savings are persistent, also require information beyond that which is contained in the database.

There has been one follow-up survey of SME that received an IAC assessment, comprised of 42 respondents to a random selection of 100 firms from a frame of 2954 (Tonn and Martin, 2000). The study does not report on realized savings, but instead reports that firms say they had changed their approach to energy efficiency decision making after the IAC assessment, suggesting some possible positive spillovers. Papers examining other types of energy efficiency program interventions report mixed results. Abeelen et al. (2016) report that there are a large differences in projected and realized savings for firms participating in the Dutch voluntary agreements on energy efficiency. Fowlie et al. (2015) find a popular home weatherization program overestimated implementation costs and underestimated savings. Parfomak and Lave (1996) find that electric utility DSM program savings estimates in the industrial and commercial sector were largely accurate. The ability to conduct such research using the IAC could potentially extend the value of the data as a tool to assess the energy-efficiency gap. The benefits may also have positive spillovers to even more energy savings or to non-energy-benefits like labor productivity or improved market position. Allcott and Greenstone (2012) point to these unobserved cost and benefits as a confounding factor to fully determining the welfare effects of such policies.

A second factor limiting the research that can be conducted using IAC data is that

nothing is known about SME that do not receive an assessment. For instance, potential questions of interest include "how do the SME that receive an assessment compare to their peers in the industry?" or "are they initially less efficient than peers, i.e., do the audits reach firms needing the most 'help'?". Randomized control trials (RCT) are becoming increasingly popular way to study energy efficiency implementation, (e.g., Allcott, 2011), but the IAC is not a RCT and the IAC data by themselves are not readily amenable to quasi-experimental analyses. Anecdotally, participants may be offered an assessment "at random" by the local IAC university, in the form of cold calls and other searches, but ultimately the SME must agree to participate, i.e., self-select. Without more information about non-participants it is impossible to account for effect of the self-selection process on the outcomes.

Our study aims to address these factors and extend the research potential of the IAC data by linking to confidential plant level data from the U.S. Census of Manufactures (CMF). Effectively, this process focuses on identifying plants in the CMF that received IAC assessments. As the CMF contains longitudinal information on plants that received IAC assessments, as well those that did not, producing a linked IAC/CMF dataset creates the opportunity to explore research questions involving both the IAC recommendations as well the information contained in the CMF. Linking also facilitates the comparison of plants receiving assessments to a peer group of plants that did not. In short, the creation of such a linked data set creates the opportunity to conduct research into the energy-efficiency gap that cannot be conducted on either the public IAC database nor the CMF alone.

Due to privacy restrictions, identifying information such as plant name have been removed from the public IAC records. Linking IAC records to CMF records therefore relies on a set of categorical and continuous variables common to both files. However, linking in this scenario is complicated by a few characteristics of the data. First, the number of IAC assessments conducted in a given year is much smaller than the number of CMF records available for the same year. This means that linking on common categorical variables rarely results in unique matches in the CMF. Second, the continuous variables common to both files do not necessarily agree for true matches. For instance, values in the IAC relating to sales information appear to be rounded, and information for each completed

form is provided at potentially different times of the year. This means that even for true matches, we do not expect values of common continuous variables to agree exactly. To facilitate linking, we develop a linking methodology that incorporates these challenges into the modelling structure and accounts for some uncertainty in the matching process.

The remainder of this paper proceeds as follows. In Section 2, we detail challenges inherent in the linking process. These challenges motivated the development of a probabilistic method for linking the two data bases. Section 2.3 includes an overview of the file linking methodology. We do not provide mathematical details of the model in the main text in order to save space. These details are provided in Section 1 of the supplementary material. The supplementary material also includes results of simulation studies that illustrate the performance of the file linking methodology that we ultimately apply on the IAC and CMF data. In Section 3, we detail the steps of applying the model to create the desired linked data sets. We use data from the 2007 CMF, but the linking method can be applied to other years to facilitate future research. In Section 4, we present a difference in difference analysis to illustrate the type of research facilitated by the development of the new linked data sets. In Section 5, we conclude with further discussion of the potential policy impacts from future research created on these data sets and suggest directions for future work.

2 Methodology

Linking data from the IAC and the CMF is an example of file linking. The process of file linking involves selecting a record i in file F_1 and identifying which, if any, record $i^{'}$ in file F_2 contains information on the same plant. A record pair $(i, i^{'})$ corresponding to the same plant is called a match. The goal of linking is to produce a linked database in which each entry contains information on a matched pair.

Let n_1 and n_2 denote the number of records in F_1 and F_2, respectively. We assume each record corresponds to a single true plant; see Steorts et al. (2014) for linking data with replicates. Under this assumption, for a general linking application, the number of true matches is bounded by the lesser of n_1 and n_2. In the case of linking the IAC and CMF records, the number of IAC assessments conducted in a given year is much smaller than the number of records in the CMF. Treating the IAC as F_1 and the CMF as F_2, this means

that $n_1 << n_2$, and the number of true matches is bounded above by n_1. Furthermore, manufacturing plants that receive IAC energy assessments represent a subset of the plants in the United States, and hence a subset of the plants in the CMF. This implies every plant in the IAC should be in the census, i.e., every record in the IAC should have a match in the CMF. When linking the IAC and CMF, we therefore expect a linked database containing exactly n_1 linked pairs. However, as we discuss in Section 3.1, realities of the data result in some IAC records for which we cannot identify a CMF match.

In this section, we discuss specific considerations for linking the IAC and CMF. In Section 2.1, we discuss the use of common categorical variables to create groups, called blocks, of potential CMF matches for each IAC record. In Section 2.2, we explain how to use common continuous variables to select among these potential matches. In Section 2.3, we conceptually summarize our linking methodology. In Section 2.4, we discuss multiple imputation as a technique for accounting for uncertainty in the linking process.

2.1 Blocking

File linking is performed in a variety of ways, depending on data involved as well as the requirements of the linked data sets. Ideally, unique identifiers are utilized to readily identify records across files corresponding to the same plant. Such identifiers are available in the CMF data in the form of a longitudinal identifier; this identifier, as well as the rest of the CMF information, is confidential. In the CMF data, this identifier can be used to track a plant's records across years as well as across census surveys. However, there is no corresponding identifier in the IAC. Indeed, for privacy restrictions, identifying information such as plant name are removed from the public IAC database.

In the absence of unique identifiers, a standard file linking technique is to leverage categorical variables common to both files to help identify possible matches. This technique, known as blocking, restricts possible matches to pairs of records with identical values on a selected set of common categorical variables. We call these common categorical variables blocking variables, or BVs. The BVs available in the CMF and IAC are the state in which a plant is located and the 6-digit NAICS code (U.S. Census Bureau, 2016) corresponding to the specific products made by the plant. We require that all possible matches agree on

Table 1: Description of the MVs. First column: name of the variable in IAC database. Second column: the name of the associated variable in the CMF. Thrid column: abbreviations used for the MVs.

IAC Variable	**CM Variable**	**Abbr.**
Sales (in $)	Total Value of Shipments (in thousands of $)	TVS
Number of Employees	Total Employment	TE

both BVs. This is a reasonable first step because we expect that if (i, i') is a match, values of state and NAICS should agree across files. For each IAC record i, we call the set of all CMF records with identical BV values to i the block of possible matches. As an example, consider an IAC record for a plant in Oregon with NAICS code 342156. The block of possible matches, or block, for this record then contains all CMF records corresponding to plants in Oregon with the NAICS code 342156.

Blocking reduces the number of record pairs that must be considered when determining which IAC and CMF records are possible matches. However, the number of IAC assessments conducted in a given year is much smaller than the number of CMF records available for the same year. This means that while it does reduces the number of possible matches to consider, blocking on the BVs rarely results in unique matches in the CMF. Indeed, for some IAC records, blocking results in thousands of CMF possible matches. The size of these blocks necessitates the use of further information to determine which CMF record in a block is a likely match for its corresponding IAC record.

2.2 Continuous variables

To help identify plants within a block which are more (or less) likely to be matches, we use two continuous variables common to the IAC and CMF. We refer to these variables as matching variables, or MVs. The MVs are listed in Table 1, and we use the abbreviations from the CMF to denote the MVs [1].

[1]The scale of the TVS variable is different in the IAC than it is in the CMF. Specifically, the CMF variable TVS is recorded in thousands of dollars while the IAC variable is recorded in dollars. We divide

If, like the BVs, the MVs are believed to agree exactly across matched pairs, we can determine matches by requiring exact agreement on the values of these variables in each IAC record to the values in each in-block CMF record. However, this assumption is not reasonable for linking the IAC and the CMF. Values in the IAC relating to sales information appear to be rounded, and information for each completed form is provided at potentially different times of the year. This means that even for true matches, while we expect similar values of TE and TVS, we do not expect exact agreement.

2.3 Linking methodology

The fact that blocking on NAICS and state fails to yield exact matches, coupled with the reality that the common continuous variables may not exactly agree across matches, motivated the development of a probabilistic file linking technique suitable for linking the IAC and CMF data. The method utilizes the BVs to define blocks of records from the CMF containing possible matches for each record in the IAC. Once blocks of records are constructed, we utilize the MVs to estimate which in-block record in the CMF is a match for each record in the IAC. We assume that these MVs may not agree exactly across matched pairs. We refer to the entire Bayesian model as LFCMV, which stands for linking with faulty continuous matching variables.

In general, probabilistic file linking aims to estimate the probability that a given pair of records is a match. These probabilities are then used to assign match status to each pair of records. Record pairs estimated to be matches are called links, while pairs estimated as non-matches are called non-links. In our application, for a given IAC record, all CMF records within blocks defined by the BVs are potential links; all remaining CMF records are declared non-links. Within a block, the LFCMV model favors links with similar values of the MVs, where the probability that any two records within a block are linked reflects how similar the MVs tend to be across different types of matched pairs.

LFCMV linking proceeds in four main steps. First, we group records into blocks based on patterns of agreement in the BVs. Second, we utilize a linking model that models the distance between the MVs of records in the IAC and the MVs of their matching records in

the IAC value by 1000 to match the scale of the CMF variable.

the CMF. This linking model is conditional on a vector C, a linkage structure that indicates which record in the CMF is matched to each record in the IAC. Third, we specify a mixture model for the distance component of the linking model. This latent structure allows the distance between MVs in matched pairs to vary across types of matched pairs. For instance, if small aluminium plants have more similar values of the MVs across matched pairs than large dairy plants, the very flexible latent class model can allow these latent features in the data to be reflected in the linking process. Finally, within each block defined by the BVs, we specify a model for the corresponding elements of C. Full details of the model, including specifications for each of the sub-models, is included in the supplementary material.

2.4 Multiple linked data sets

For each IAC record i, the goal of file linking is to determine which of the records in its block is a likely match. We define $C_i = i^{'}$ where $i^{'}$ denotes a CMF record such that $(i, i^{'})$ is a match, i.e., CMF record $i^{'}$ is a match with IAC record i. The n_1 element vector $C = (C_1, \ldots, C_{n_1})$ then specifies which records from F_2 match the records in F_1, i.e., which records in the CMF received IAC assessments.

In LFMCV, we create estimates of C by drawing from the posterior distribution, i.e., a probability distribution specified by the structure of the LFCMV model. A draw for C specifies which CMF records to link to the IAC records, so conditional on a draw of C, we can link the IAC and CMF databases to create the desired linked data set. However, drawing a single estimate of C, and thereby creating a single linked data set, has some statistical and practical limitations. In file linking, C is a missing quantity we attempt to estimate; we have the information from the CMF and IAC, but the structure which connects the two, i.e., C, is unknown. This means file linking can be viewed as a missing data problem (e.g., Wu, 1995; McGlincy, 2004; Gutman et al., 2013).

There are many approaches for handling missing data, including imputing, or estimating, the missing quantities (e.g., Little and Rubin, 2002; Reiter and Raghunathan, 2007). We use multiple imputation, which is particularly suited to creating completed datasets for subsequent analyses. Multiple imputation is the process of filling in missing data by creating multiple estimates of the missing elements. Each estimate of the missing data is

used to create a completed data set, and the process is repeated $M > 1$ times to yield multiple, completed data sets. In the context of file linking as a missing data problem, multiple imputations are created for the "missing" linkage structure by using multiple draws of C. We draw C and link the IAC and CMF according to that draw. We then re-run the model for another iteration, draw a new sample of C, and link the IAC and CMF according the new draw, yielding the second version of the linked data. This process is repeated M times, producing M versions of the linked data set. The number M of data sets produced is at the discretion of the analyst, though multiple imputation is more effective with a large number of imputations (Reiter and Raghunathan, 2007).

From a statistical perspective, creating multiple imputations is more reflective of the uncertainty in the estimated links than as single imputation (Little and Rubin, 2002). C is a quantity that must be estimated, meaning that there will be some uncertainty in the estimated links. Create multiple linked data sets, each reflecting a potentially different estimate of C, is a way to reflect the uncertainty in that estimation process in any post-linking analysis performed on the data. Obtaining estimates of desired quantities, like means or regression parameters, from the M completed data sets can be readily accomplished using combining rules of Rubin (1976). Coding incorporating these rules is included in many standard statistical software platforms. In this paper, we refer to the data sets resulting from LFCMV linking as "the linked data sets". The M linked data sets will be made available in the Federal Statistical Research Data Center network for use by future researchers.

3 Preparing the data for linkage

In this section, we detail the practical steps used to create the linked data sets, including the necessary data pre-processing.

3.1 BVs and MVs

The IAC and CMF records are collected by two different organizations. As such, some variables common to both files are recorded using an organization specific scale, variable

name, and accuracy level. For instance, in the IAC, NAICS code is recorded as a 6-digit identifier, while in the CMF, up to 9-digits are recorded. Because of these discrepancies, pre-processing is required before linking. We begin with all records in the 2007 CMF, and all records in the IAC corresponding to assessments conducted in 2007 or 2008. There are 797 such assessments.

The first step is to examine the BVs used to block the records. In the CMF, state is denoted using the standard notation of two upper case letters. A small number of records are missing the variable for state, and we exclude these records from consideration for matching. In the IAC, however, the BV "state" is recorded with a variety of upper-lower case combinations. For instance, "AL","Al" and "al" all refer to the state Alabama. We convert the IAC labels to match the double capital format in the CMF, i.e., for plants referring to Alabama, state is recorded as "AL". Once the state variable is standardized, we create a list of all states associated with plants in the IAC. All CMF records from states not on this list are declared non-links. In addition to records pertaining to the 50 states, the IAC contains 11 records from plants in Puerto Rico. As there are no records corresponding to Puerto Rican plants in the CMF, we exclude these records from consideration for matching, meaning a total of 786 IAC records are considered for matching.

After blocking on state, we proceed to the BV of NAICS code. For each record in the IAC, we create a list of CMF records pertaining to the same 6-digit NAICS code and state as the IAC record. However, such direct blocking leads to a number of IAC records with no possible matches in the CMF. The IAC records are a subset of the CMF, and as such all records in the IAC should have a match in the CMF. The lack of matches for some records after direct blocking suggests that some of the NAICS codes in IAC or CMF may have some uncertainty. To account for this, we considering blocking on three different levels of NAICS agreement. Specifically, we create subsets with agreement on 4-digit, 5-digit and 6-digit NAICS codes. We refer to these as levels of NAICS blocking. NAICS codes are nested up to the 4-digit level, meaning that all plants that produce a certain type of product, say dairy products, must agree on at least the first 4 digits of their NAICS codes. However, the results of our difference in difference analysis suggest little distinction in results across the linked data products created at each level of NAICS blocking. We therefore present

results only from linking at the 6-digit NAICS level; further discussion is presented in the supplementary material.

3.2 Agreement filters

Blocking with respect to 6-digit NAICS leads to some large blocks. For file linking, smaller blocks tend to lead to a higher match rate, while for extremely large blocks, models often selects a match essentially at random. To reduce the size of the blocks, we apply what we call an agreement filter. We expect that values of TVS and TE should be similar across matched pairs. Therefore, for a given IAC record i, we compute the difference between the values of TVS for record i and each CMF record in its block. We then divide by the TVS value for record i, yielding a percent difference. If the difference for any pair $(i, i^{'})$ is more than 30%, we declare $i^{'}$ a non-link and exclude it from the block. We then repeat this process for values of TE. After applying the agreement filter, we have 330 IAC records with possible matches in the CMF.

The agreement filter chosen here reflects a belief that for true matches, the values of TVS and TE in the IAC and CMF should be similar. The threshold of 30% allows some room for noise, recording error and time variation. In general applications, the choice to apply an agreement filter, as well as the choice of appropriate threshold, is at the discretion of the analyst, and should be tuned to reflect the data involved.

4 Results and Discussion

Following a discussion of the linked data sets in Section 4.1, in Sections 4.2 we conduct an analysis illustrating the research potential of the linked data sets.

4.1 The linked data sets

In this section, we examine the links estimated by LFCMV. Specifically, we are interested in how often each IAC record i is matched to a given CMF record, i.e. the posterior probability that a given pair $(i, i^{'})$ is linked. This posterior probability is computed as proportion of draws for C such that each in-block record $i^{'}$ is linked to its corresponding

Table 2: Types of links obtained after LFCMV linking. Define a link as high probability if the posterior probability of linking $(i, i^{'})$ is greater than .5.

Link Type	High Prob	Low Prob	Total
Link Count	270	60	330

IAC record i. We run LFCMV to link F_1 and F_2. We draw 5000 estimates of C, producing $M = 5000$ linked data sets. For each IAC record i, we look at the posterior probabilities associated with possible in-block matches $i^{'}$. If at least one of the potential matches has a posterior probability $> .5$, we say i has a high probability match. Otherwise, we say i has low probability matches. Table 2 shows a summary of the counts of IAC records with high and low probability matches [2]. The results suggest that for most IAC records, there is a CMF record which favored to be linked over all other in-block options. For some records with low probability matches, the options for TE and TVS were similar for some in-block matches, leaving the model unable to favor one choice over the other. This is not undesirable, and indeed strengthens the value of creating of multiple completed data sets which are able to reflect these equally-likely match options. Other low probability matches were the results of no potential match having compelling values of the MVs, causing the model to select a match essentially at random from the in-block records. Again, creating multiple linked data sets allows this to be reflected in the final linked data sets and in resultant analyses.

4.2 Difference in difference

The motivation for linking the IAC and CMF is to create a data base that yields opportunities for more in depth policy analysis than can be conducted on either the IAC or CMF data independently. In this section, we conduct a difference in difference (DID) analysis designed to illustrate the potential of the linked data sets for energy-efficiency research. Specifically, we examine energy efficiency in plants that received IAC energy assessments versus those that did not. As the CMF is conducted every 5 years, we compare the differ-

[2]Counts have been rounded to satisfy disclosure protocols.

ence in efficiency trends between 2007 and 2012. We do not make a causal argument about the relationship between receiving an assessment and energy efficiency, as many factors, including plant size, may impact the relationships observed in the data. We do control for industry specific factors at the 6-digit NAICS level, as described in Section 4.2.2. For each plant, we refer to all plants with the same 6-digit NAICS code as the plant's peer group.

As a basis for our analysis, we begin with all CMF records from 2007 and 2012 that have a 4-digit NAICS code observed in the IAC. This limits our analysis to industries that were assessed in the time period of interest. We further limit our analysis sample to plants meeting IAC standards for assessment. IAC assessments are typically performed for plants with no more than 500 employees and with a gross annual sales of no more than $100 million (Muller, 2001, pg. 13). Let A_{2007} denote the set of CMF records from 2007 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. Similarly, let A_{2012} denote the set of CMF records from 2012 that satisfy the IAC restrictions for sales and employees and have 4-digit NAICS codes that appear in the IAC. We refer to A_{2007} and A_{2012} as our analysis samples.

4.2.1 Energy metric

Based on the linked data sets, we wish to investigate the energy efficiency of plants that received an assessment relative to those that did not, as well as the changes over time. This provides some insights into our linked data and lays the groundwork for further analysis, as discussed in the conclusions. Energy efficiency is a measure of relative performance capturing how much "better" or "worse" plants' energy use is for those that received assessments compared to a peer group of those that did not. We use a two-step process to create our energy efficiency metric.

We create a standardized metric of efficiency by first computing energy productivity, i.e., dollar sales per dollar of energy purchased. For each record $i^{'}$ in A_{2007} and A_{2012}, we compute an energy productivity metric as follows:

$$EF_{i'} = log\left(\frac{TVS_{i'}}{CF_{i'} + EE_{i'}}\right). \tag{1}$$

Here, (1) is the log of the inverse of the energy cost share. We use productivity rather than cost shares (or energy intensity) so that a "better" outcome is a positive number in

our final metric, i.e., more efficient. We also account for industry specific differences in two ways. The plant level energy productivity is normalized by the industry level mean, as defined by 6-digit NAICS, to account for difference in the inherent energy use in different products; this first step is similar to an approach used in Boyd and Curtis (2014). We further standardize the metric by the dispersion of industry level productivity. Boyd (2017) observes that industries that have low energy productivity (high cost shares) tend to have lower dispersion. Our efficiency metric accounts for both the industry level differences in the means and differences in dispersion. The result is an energy efficiency score that is zero for the "average" plant and positive for the more efficient plant. These scores are comparable across all NAICS codes.

4.2.2 Results

The records in A_{2007} and A_{2012} are divided into two categories. The first category, labelled "With Assessments", refers to plants in a given year that received an IAC assessment according to the linked data sets. The remaining plants are categorized as "Without Assessments", meaning that these plants did not receive an IAC assessment in 2007 or 2008. The process of assigning records in A_{2007} and A_{2012} to these categories is dependent upon the linking structures estimated in Section 4.1. For each posterior sample $s = 1, \ldots, M = 5000$ of C, we create a binary vector $I^{(s)}$ of length a_{2007}, where a_{2007} denotes the number of records in A_{2007}. Here, a 1 in position $i^{'}$ means that record $i^{'}$ in A_{2007} is linked to an IAC record at iteration s; such plants are categorized as "With Assessment" records for iteration s. All other plants are categorized as "Without Assessment" records.

In order to compare the energy scores in 2007 and 2012, we also consider records in A_{2012}. Linking is performed on the 2007 records, meaning that the indicators for IAC assessments are associated with a 2007 CMF record. However, the Census Bureau uses a longitudinal identifier that can be used to identify records for a given plant across time. For each iteration, we identify the set of such identifiers that, conditional on each $I^{(s)}$, received an energy assessment in 2007. The 2012 records corresponding to these unique identifiers compose the 2012 "With Assessment" samples.

For each iteration s, we compute the density of energy scores for "With Assessment"

plants, and separately for "Without Assessment" plants, for 2007 and 2012. We use fixed bins across the iterations [3]. Within each year, and each of the two categories, we then average the density in each bin at each iteration and plot the results. The resultant average density curves for 2007 are displayed in Figure 1, and the resultant averaged density curves for 2012 are displayed in Figure 2.

In Figure 1, the curve for the "With Assessment" group is shifted further left than the "Without Assessment" curve. This suggests that the energy scores for assessed plots are smaller than plants without assessments. In other words, relative to their peer group, plants receiving IAC audits appear less energy efficient in 2007. This difference appears to vanish in the plots corresponding to 2012. In Figure 2, there is no appreciable difference in the curves between the "Without" and "With Assessment" groups.

[3]The bin width is confidential to satisfy disclosure protocols.

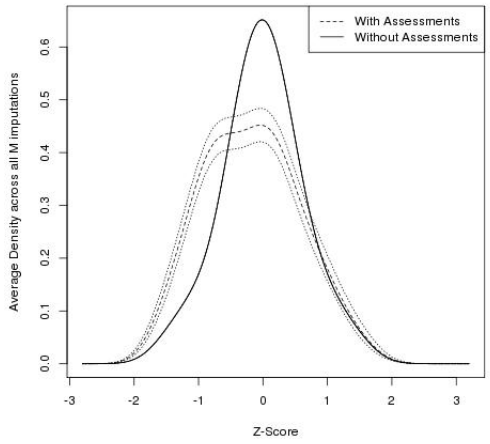


Figure 1: Kernel Density Plots for 2007. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

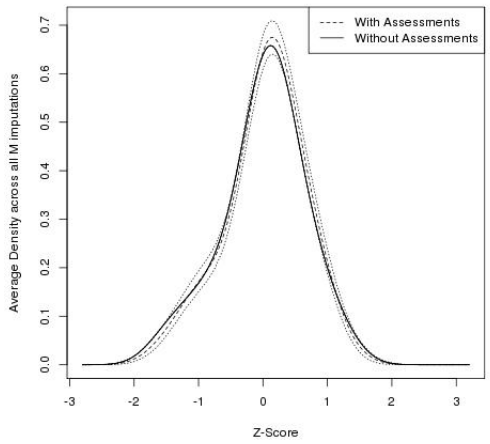


Figure 2: Kernel Density Plots for 2012. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

Table 3: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment" group.

		Mean	95% Interval
2007	With Assessments	-0.18	(-0.29,-0.08)
	Without Assessments	-0.007	(-0.009,0.005)
2012	With Assessments	0.04	(-0.05,.12)
	Without Assessments	0.009	(0.06,.011)

To supplement the graphical exploration of the data, we leverage the combining rules of Rubin (1987) to perform inference across the M data sets composing the linked data sets. Using the "With" and "Without Assessment" groups described above, we compute the average energy score for each of the two groups for each of the M data sets. Using the combining rules of Rubin (1987), we compute 95% confidence intervals for the mean of each group. These intervals account for some linkage uncertainty by combining results across the M data sets. The resultant intervals, displayed in Table 3, reveal the same pattern observed in the density curves. In 2007, the average energy score tends to be lower for the "With Assessment" groups, with non-overlapping intervals for the "With" and "Without" groups. This difference is much less pronounced in 2012, and the intervals overlap, suggesting no statistically significant difference in the means between the groups.

We performed a similar analysis for on the linked data sets using total employment instead of energy costs. This analysis examined labor efficiency differences in "With Assessment" versus "Without Assessment" plants. The resultant density curved showed minimal difference in labor efficiency between the assessed plant and the populations in 2007 and 2012. Since the IAC program is targeted at energy, not labor, we feel this more evidence that the linking approach works and that the IAC program has an impact on those plants that receive assessment.

5 Conclusions and Policy Implications

In this paper, we have described the creation of a database that allows novel research to be conducted on the policy interventions of detailed energy-efficiency assessments to SME. By combining information from the Census Bureau with publicly available IAC records, we increase the utility of the IAC database by enabling longitudinal research, as well as cross sectional comparisons to SME that do not receive assessments. The novel linking approach developed here allows any statistical inferences on the impact of this policy to account of the uncertainty in the linking itself.

A difference in difference analysis of the linked data sets presents suggestive evidence about the impact of the IAC assessment program. The data show that the distribution of plants that received an assessment appear to be less energy efficient than the population, but show no difference when observed five years later. These empirical observations do control for industry effects at the 6-digit NAICS level, the focus of the IAC on SME, and also for the underlying uncertainty in the matching process. We caution that these trends may not reflect a causal effect of the assessments; there may be sources of confounding that we have not accounted for. The observation that the linked plants are different from the population also speaks well of the linking process itself. If the linking were simply random guesses, conditional on the blocking variables, then we would expect no patterns to emerge. That there is no pattern for labor vis-à-vis energy provide even more support for the validity of the linkage, since the program is targeted at energy. It may well be that plants that receive assessment are, in fact, those that are more likely to benefit from them. The next step in answering such questions will be to explore the linked data further with more detailed statistical models.

Future research directions include various types of quasi-experimental approaches, beyond the scope of the difference in difference models in Section 4.2, i.e. using longitudinal analysis, and treatment effects models using propensity score matching or other sample selection corrections. Neither of these approaches would have been possible before the creation of this linked database. The detail in the IAC data on technologies and adoption paired with the additional plant and firm level information from the Census data could provide new insights about assessment / recommendation / technology adoption process

itself. The IAC is rich with detail about technology, but the Census data adds information about firm structure, profitability, plant age, investment spending, etc. The list of possible questions below that this new database enables is quite large and by no means exhaustive.

- What types of IAC recommendations have the largest impact?
- Is there heterogeneity amongst IAC locations?
- Are plants that receive large number of recommendations less efficient to start with?
- Are there within firm spillovers for an SME that receives an assessment at one plant, but not at others owned by the firm?
- Are there spillovers to productivity or are the impacts limited to energy? [4]
- Does firm structure have an impact on adoption as reported in the IAC, or impact of the assessments (as revealed by longitudinal Census analysis).

As interest grows by the utility industry in demand side management (DSM) program and by state/national government in ways to enable energy savings and increase profitability, particularly in SME manufacturing, the need to know what works, what does not, and how to better target information programs like the IAC only grows. This new linked database, which can be expanded to include other economic Census years and will be made available in the Federal Statistical Research Data Center network for use by other researchers, can contribute to this type of policy-based, evidence-building process.

Acknowledgements

This work was supported by NSF Grant SES 1131897, and by the Duke University Energy Initiative Energy Research Seed Fund, with co-funding from the Information Initiative at Duke. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

[4]Some results about labor productivity may suggest that this may not be the case, but the question still bears closer examination.

References

Abeelen, C., Harmsen, R., and Worrell, E. (2016), "Planning versus implementation of energy-saving projects by industrial companies: Insights from the Dutch long-term agreements," *Energy Efficiency*, 9, 153–169.

Allcott, H. (2011), "Social norms and energy conservation," *Journal of Public Economics*, 95, 1082 – 1095, Special Issue: The Role of Firms in Tax Systems.

Allcott, H. and Greenstone, M. (2012), "Is there an energy efficiency gap?" *Journal of Economic Perspectives*, 26, 3–28.

Anderson, S. and Newell, R. (2004), "Information programs for technology adoption: the case of energy-efficiency audits," *Resources and Energy Economics*, 26, 27–50.

Boyd, G. (2001), "A probit model of energy efficiency technology decision making," in *2001 ACEEE Summer Study on Energy Efficiency in Industry*, American Council for an Energy Efficient Economy.

Boyd, G. A. (2017), "Comparing the statistical distributions of energy efficiency in manufacturing: meta-analysis of 24 Case studies to develop industry-specific energy performance indicators (EPI)," *Energy Efficiency*, 10, 217–238.

Boyd, G. A. and Curtis, E. M. (2014), "Evidence of an 'Energy-Management Gap' in U.S. manufacturing: Spillovers from firm management practices to energy efficiency," *Journal of Environmental Economics and Management*, 68, 463 – 479.

Domingo-Ferrer, J. and Torra, V. (2002a), "Distance-based and probabilistic record linkage for re-identification of records with categorical variables," *Butlletí de lACIA, Associació Catalana dIntelligència Artificial*, pp. 243–250.

Domingo-Ferrer, J. and Torra, V. (2002b), "Validating distance-based record linkage with probabilistic record linkage," in *Topics in Artificial Intelligence*, pp. 207–215, Springer.

Domingo-Ferrer, J. and Torra, V. (2003), "Disclosure risk assessment in statistical microdata protection via advanced record linkage," *Statistics and Computing*, 13, 343–354.

Escobar, M. and West, M. (1995), "Estimating normal means with a dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Fowlie, M., Greenstone, M., and Wolfram, C. (2015), "Do energy efficiency investments deliver? Evidence from the Weatherization Assistance Program," Working Paper 21331, National Bureau of Economic Research.

Gerarden, T. D., Newell, R. G., and Stavins, R. N. (2015), "Assessing the energy-efficiency gap," Working Paper 20904, National Bureau of Economic Research.

Gutman, R., Afendulis, C., and Zaslavsky, A. (2013), "A Bayesian procedure for file linking to analyze end-of-life medical costs," *Journal of the American Statistical Association*, 18, 34–47.

Henriques, J. and Catarino, J. (2016), "Motivating towards energy efficiency in small and medium enterprises," *Journal of Cleaner Production*, 139, 42 – 50.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Jaffe, A. B. and Stavins, R. N. (1994), "The energy-efficiency gap: what does it mean?" *Energy Policy*, 22, 804 – 810.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014), "Multiple imputation of missing or faulty values under linear constraints," *Journal of Business & Economic Statistics*, 32, 375–386.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, John Wiley & Sons, Inc.

McGlincy, M. H. (2004), "A Bayesian record linkage methodology for multiple imputation of missing links," in *ASA Proceedings of the Joint Statistical Meetings*, pp. 4001–4008, American Statistical Association.

Muller, M. B. (2001), "IAC database manual," Tech. rep., U.S. Department of Energy, Center for Advanced Energy Studies, Rutgers University.

Murillo-Zamorano, L. R. (2004), "Economic efficiency and frontier techniques," *Journal of Economic Surveys*, 18, 33–77.

Pagliuca, D. and Seri, G. (1999), "Some results of individual ranking method on the system of enterprise accounts annual survey," *Esprit SDC Project, Deliverable MI-3/D2*.

Parfomak, P. W. and Lave, L. B. (1996), "How many kilowatts are in a negawatt? Verifying "Ex Post" estimates of utility conservation impacts at the regional level," *The Energy Journal*, 17, 59–87.

Perroni, M. G., da Costa, S. E. G., de Lima, E. P., and da Silva, W. V. (2016), "The relationship between enterprize efficiency in resource use and energy efficiency practices adoption," *International Journal of Production Economics*, pp. –.

Reiter, J. P. and Raghunathan, T. E. (2007), "The multiple adaptations of multiple imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, Wiley, New York, USA.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Sethuraman, J. (1994), "A constructive definition of dirichlet priors," *Statistica Sinica*, 4, 639–650.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2014), "SMERED: a Bayesian approach to graphical record linkage and de-duplication," *Journal of Machine Learning Research*, 33, 922–930.

Thollander, P., Danestig, M., and Rohdin, P. (2007), "Energy policies for increased industrial energy efficiency: Evaluation of a local energy programme for manufacturing {SMEs}," *Energy Policy*, 35, 5774 – 5783.

Tonn, B. and Martin, M. (2000), "Industrial energy efficiency decision making," *Energy Policy*, 28, 831 – 843.

Torra, V., Abowd, J. M., and Domingo-Ferrer, J. (2006), "Using mahalanobis distance-based record linkage for disclosure risk assessment," in *International Conference on Privacy in Statistical Databases*, pp. 233–242, Springer.

U.S. Census Bureau (2016), "North American Industry Classification System," `http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007`.

U.S. Department of Energy (2016), "Industrial Assessment Centers (IAC)," `https://energy.gov/eere/amo/industrial-assessment-centers-iacs`.

Wu, Y. (1995), "Random shuffling: a new approach to matching problems," in *ASA Proceedings of the Statistical Computing Section*, pp. 69–74, American Statistical Association.

Supplementary Material

The following sections comprise supplementary material to complement the main article. In Section A, we present the LFCMV model, including notation and sub-model specifications. In Section B, we describe the Gibbs sampling steps used to obtain posterior estimates of the linkage structure. In Section C, we use simulation studies to illustrate the performance of LFCMV. In Section D, we present results of additional studies that investigate the performance of LFCMV under a variety of fault scenarios. In Section E, we present extended results relating to Section 4 in the main text.

A Model

In this section, we describe the LFCMV methodology. Notation and key concepts are presented Section A. We then present the model for C in Section A.2, and the linking model in Section A.3.

A.1 Notation

Let the two files to be linked be denoted F_1 and F_2, containing n_1 and n_2 records, respectively. Without loss of generality, we assume that $n_1 < n_2$. We further assume that each record $i \in F_1$ has a match $i^{'} \in F_2$, that is, the set of individuals in F_1 is a subset of the set of individuals in F_2. This assumption is motivated by the application in which the IAC assessment records (F_1) represent a subset of the records in the CMF (F_2). We assume each record corresponds to a single true individual.

Our goal is to link each record $i \in F_1$ to a record $i^{'} \in F_2$. To reduce the number of possible matches for each i, we place each of the n_1 records in F_1 in its own block. Let (i) denote the block defined by $i \in F_1$. For each record i, we limit the possible matches to a set of records $i^{'} \in F_2$ which are assigned to block (i). Possible matches $i^{'} \in F_2$ are assigned to (i) as follows.

Let J denote the number of BVs, i.e., the number of categorical variables common to F_1 and F_2 that are used for blocking. It is not necessary to use all common categorical variables for blocking; analysts can select variables appropriate for blocking for each application. Let

B_{fik} denote the BV value for record i in file f on field k, where $k = 1, \ldots, J$, $i = 1, \ldots, n_f$, and $f = 1, 2$. For all (f, i), let $B_{fi} = (B_{fi1}, \ldots, B_{fiJ})$. For each $i \in F_1$, define $F_{2(i)}$ as the set of all $i^{'} \in F_2$ such that $B_{1ik} = B_{2i^{'}k}$ for all $k = 1, \ldots, J$. Assign all records in $F_{2(i)}$ to block (i). In other words, for each i, we restrict the set of possible matches to records $i^{'}$ in F_2 with the same BV values as record i.

Under this blocking structure, it is possible that a record $i^{'}$ in F_2 can be considered a possible match for more than one record i in F_1. Each record i in F_1 is placed in its own block, and the set of possible matches $F_{2(i)}$ are defined only by the BV combination B_{fi}. If two records a and b in F_1 have the same BV combination, their sets of possible matches, $F_{2(a)}$ and $F_{2(b)}$, contain the same records from F_2. This yields the possibility that both a and b can be linked to the same record in F_2. For our application, this is not a concern. Our goal is to identify records in the CMF that have received an IAC assessment, not necessarily to generate a set of matched pairs. If unique matches are required for a given application, one could imagine adapting this blocking structure to suit this need. For instance, blocks may be defined such that more than one F_1 record may be assigned to a block. Matching would then assign each record in $F_{2(i)}$ to at most one record i in F_1.

With n_1 records in F_1, we have n_1 blocks of records from F_2. Each block $F_{2(i)}$ represents potential matches for the accompanying record $i \in F_1$. Let $n_{(i)}$ be the number of records in $F_{2(i)}$. For each $i \in F_1$, the goal of file linking is to determine which of the $n_{(i)}$ records is a match for i. We define $C_i = i^{'}$ such that $i^{'} \in F_{2(i)}$ and $(i, i^{'})$ is a match. The n_1 element vector $C = (C_1, \ldots, C_{n_1})$ then specifies which records from F_2 match the records in F_1.

Finally, let p be the number of common continuous variables (MVs) used for matching. For all $j = 1, \ldots, p$, let Y_{ij} denote the standardized value for MV j for record i in F_1. We provide details on this standardization in Section C.1. For all $i = 1, \ldots, n_1$, let $Y_i = (Y_{i1}, \ldots, Y_{ip})$. Similarly, let $\mathbf{Y}_j = (Y_{ij}, \ldots, Y_{n_1 j})$ for all j. We define $\mathbf{Y}$ as the $n_1 \times p$ matrix of MV values for F_1. Similarly, for all j, let $X_{i^{'}j}$ denote the standardized value for MV j for record $i^{'} \in F_2$ where $i^{'} = 1, \ldots, n_2$. Let $X_i = (X_{i1}, \ldots, X_{ip})$ and $\mathbf{X}_j = (X_{ij}, \ldots, X_{n_1 j})$. We define $\mathbf{X}$ as the $n_2 \times p$ matrix of MV values for F_2.

A.2 Model for C

Within a block $F_{2(i)}$, there are $n_{(i)}$ possible values of C_i. We denote these values as ℓ, where $\ell \in \{1, \ldots, n_{(i)}\}$. We assume that *a priori*, each $i^{'}$ in $F_{2(i)}$ is equally likely to be linked with i, so that for all ℓ,

$$p(C_i = \ell \mid B_{1i}) = \frac{1}{n_{(i)}}. \tag{2}$$

Conditional on B_{1i}, which defines the block for record i, each C_i is modeled independently.

A.3 Linking model

Conditional on C, we assume that for each matching pair $(i, i^{'} = C_i)$, the values of Y_i are related to X_{C_i} through a linking model. This model reflects the belief that for true matches $(i, i^{'} = C_i)$, we may have $Y_{ij} \neq X_{i^{'}j}$ for some $j = 1, \ldots, p$. Such a discrepancy could occur due to incorrect recording, time differences in data collection, different accuracy thresholds, etc. The notion of linking based on examining distances between variables common to both files is an example of distance-based record linkage. Distance-based record linkage is commonly used for re-identification purposes in disclosure and privacy research (e.g., Pagliuca and Seri, 1999; Domingo-Ferrer and Torra, 2002a, 2003; Torra et al., 2006). In these applications, distances between each record i and $i^{'}$ are computed using some distance metric, and the record pair that has the smallest distance according to this metric is considered a link (Domingo-Ferrer and Torra, 2002b).

We model Y_i with a multivariate normal distribution centering Y_i at its corresponding X_{C_i}. This choice of linking model favors matched pairs with similar values of Y and X. The variance component of the linking model represents the distance between Y_i and X_{C_i} that is considered plausible for matched pairs. In our application, each IAC record corresponds to a manufacturing plant. It is reasonable to assume that the distance in the MVs across matched pairs may vary with state, plant type, or other characteristics of the records. In more general applications, the distance may vary across certain blocks, or with other features in the data. To allow the linking model to capture these distributional features, we utilize a mixture of multivariate normals. Mixtures of multivariate normals are highly flexible and, with enough components, can represent any distribution. For this application,

the mixture framework allows the distance across types of matched pairs to vary with latent class.

Following standard conventions for a mixture model, assume each pair $(i, i^{'} = C_i)$ belongs to one of H latent classes. For convenience, we associate the latent class with record i. Let $z_i \in \{1, \ldots, H\}$ represent the latent class assignment for $i \in F_1$, with $z = (z_1, \ldots, z_{n_1})$. For $h = 1, \ldots, H$, we assume that $Pr(z_i = h) = \pi_h$ for all i. Let $\pi = (\pi_1, \ldots, \pi_H)$.

Conditional on z and C, we model Y_i with a multivariate normal distribution centered at X_{C_i} and class-specific variance Σ_h. This mixture model can be written as

$$Y_i|\mathbf{X}, C, \mathbf{\Sigma}, z_i \sim N_p\left(X_{C_i}, \Sigma_{z_i}\right) \tag{3}$$

$$z_i|\boldsymbol{\pi} \sim Multinomial\left(1; \pi_1, ..., \pi_H\right). \tag{4}$$

This is an adaption of the standard mixture of multivariate normals in which each Y_i is modeled as $Y_i \mid \boldsymbol{\mu}, \mathbf{\Sigma}, z \sim N_p\left(\mu_{z_i}, \Sigma_{z_i}\right)$. In the standard formulation, each Y_i is assumed to have a component specific mean μ_{z_i}. In our framework, we center each Y_i at its corresponding X_{C_i}. This reflects a belief that for matched pairs, Y_i and X_{C_i} should be close together. In our application, $p = 2$, so (3) is bivariate normal.

Following Kim et al. (2014), we select a conjugate prior structure for each Σ_h as follows:

$$\Sigma_h \sim InvWishart\left(f_0, G_0\right), \tag{5}$$

where $G_0 = Diag(\phi_1, ..., \phi_p)$ and

$$\phi_j \sim Gamma(a_\phi, b_\phi), \tag{6}$$

with $E(\phi_j) = a_\phi / b_\phi$. This conjugate prior structure facilitates posterior updates, as discussed in Section B. In order to ensure a proper posterior distribution, we set $f_0 = p + 1$. We set the hyper-parameters $a_\phi = b_\phi = 0.25$, a choice which allows substantial prior mass at modest sized variances (Kim et al., 2014).

For the latent class weights π, we use a stick-breaking representation of the truncated Dirichlet process (Sethuraman, 1994; Ishwaran and James, 2001). In this framework, the

mixture probabilities are

$$\pi_h = V_h \prod_{g<h} (1 - V_g), h = 1, ..., H \tag{7}$$

$$V_h \sim Beta(1, \alpha), V_H = 1 \tag{8}$$

$$\alpha \sim Gamma(a_\alpha, b_\alpha). \tag{9}$$

We set $a_\alpha = b_\alpha = 0.25$, reflecting a low prior sample size and hence a vague prior for α. As noted in Escobar and West (1995), such a choice allows the data to dominate in the posterior.

The stick-breaking representation allows the data to inform the number of latent classes that are occupied at any iteration of the posterior sampler. As H represents the maximum number of occupied classes, we recommend starting with a large value, such as $H = 30$. When posterior runs indicate that all H classes are consistently occupied, we increase H and restart the sampling.

There are a few considerations to take into account when using LFCMV for file linking. First, as is true for file linking methodologies in general, block size is important to the performance of the model. Smaller blocks tend to lead to a higher match rate, while for extremely large blocks, the model often selects a match essentially at random. It is therefore important to select a blocking scheme that results in small blocks. Second, the linking model described in Section A.3 is based on distance between Y_i and its corresponding X_{C_i}. As we illustrate in Section C, in the posterior, the model tends to assign high probability to matches with small distances across matched pairs. The determination of what is a reasonable distance across a matched pair is influenced by the class-specific variance component of the linking model.

B Posterior sampling

In this section, we describe the Gibbs sampler used to estimate the posterior distribution of the LFCMV model.

B.1 Initialization

We initialize $C^{(0)}$ such that

$$C_i^{(0)} = \left(i^{'} \mid d(i, i^{'}) = min\left(d(i, \ell), \ell \in F_{2(i)} \right) \right), \tag{10}$$

where

$$d(i, i^{'}) = \sqrt{\sum_{j=1}^{p} \left(\frac{Y_{ij} - X_{i^{'}j}}{Y_{ij}} \right)^2}. \tag{11}$$

In other words, within a block, we initialize C_i such that $d(i, i^{'})$ is minimized. The two closest records according to this metric will be initialized as links. Because Y_i and X_{C_i} are standardized variables, $d(i, i^{'})$ assigns equal weight to the distances in each of the p MVs. The importance of such standardization in distance-based linkage is discussed in Pagliuca and Seri (1999).

For the linking model, we set $\alpha^{(0)} = 1$, and initialize V, π, z, and $\mathbf{\Sigma}$ by drawing from the appropriate distributions from Section A.3.

B.2 Posterior sampling algorithm

With the truncated representation of the Dirichlet process, posterior sampling for the entire model is facilitated with a Gibbs Sampler as follows.

1. For $h = 1, ..., H$, update Σ_h from the full conditional

$$\Sigma_h^{(s+1)} | \mathbf{Y}, \mathbf{X}, C^{(s)}, z^{(s)} \sim InvWishart\left(f_0 + n_h^{(s)}, G_0 + S_h^{(s)} \right), \tag{12}$$

where $n_h^{(s)} = \sum_i I(z_i^{(s)} = h)$ denotes the number of individuals in latent class h at iteration (s) and $S_h^{(s)} = \sum_{i:[z_i^{(s)}=h]} (Y_i - X_{C_i^{(s)}})(Y_i - X_{C_i^{(s)}})^T$.

2. For $h \in \{1, ..., H-1\}$, update V_h from the full conditional,

$$V_h^{(s+1)} | z^{(s)}, \alpha^{(s)} \sim Beta\left(1 + n_h^{(s)}, \alpha^{(s)} + \sum_{g=h+1}^{H} n_g^{(s)}\right). \tag{13}$$

   Set $V_H^{(s+1)} = 1$.

3. Set $\pi^{(s+1)} = V_h^{(s+1)} \prod_{g<h} (1 - V_g^{(s+1)})$ for all $h \in \{1, ..., H\}$ per (7).

4. For $j = 1, \ldots, p$, update ϕ_j from the full conditional

$$\phi_j^{(s+1)} | \mathbf{\Sigma}^{(s+1)} \sim Gamma\left(a_\phi + \frac{1}{2} H f_0, b_\phi + \frac{1}{2} \sum_{h=1}^{H} \left(\Sigma_h^{-1}[j,j]^{(s+1)}\right)\right), \tag{14}$$

   where $\left(\Sigma_h^{-1}[j,j]\right)^{(s+1)}$ represents the j^{th} diagonal entry of $(\Sigma_h^{-1})^{(s+1)}$.

5. Update α from the full conditional,

$$\alpha^{(s+1)} | \pi^{(s+1)} \sim Gamma\left(a_\alpha + H - 1, b_\alpha - log(\pi_H^{(s+1)})\right). \tag{15}$$

6. For $i \in 1, ..., n$, sample the latent class indicator $z_i \in \{1, ..., H\}$ from a multinomial full conditional,

$$z_i^{(s+1)} | \phi^{(s+1)}, C^{(s)} \sim Multinomial({\pi_1}^*, {\pi_2}^*, ..., {\pi_H}^*), \tag{16}$$

   where

$${\pi_h}^* = \frac{\pi_h^{(s+1)} f(Y_i | X_{C_i^{(s)}}, \Sigma_h^{(s+1)})}{\sum_{g=1}^{H} \pi_g^{(s+1)} f(Y_i | X_{C_i^{(s)}}, \Sigma_g^{(s+1)})}. \tag{17}$$

7. For $i = 1, ..., n_1$, we update each C_i. For each $i' \in F_{2(i)}$,

$$Pr\left(C_i = i' | Y_i, \mathbf{X}, z_i^{(s+1)}, \mathbf{\Sigma}^{(s+1)}\right) = \frac{f(Y_i | \mathbf{X}, C_i = i', \Sigma_h^{(s+1)})}{\sum_{\ell \in F_{2(i)}} f(Y_i | \mathbf{X}, C_i = \ell, \Sigma_h^{(s+1)})}. \tag{18}$$

The update in (18) represents draws from the multinomial Bernoulli posterior distribution of C. Because each block contains only one Y_i, computing (18) requires enumerating $n_1 \times \left(\sum_i n_{(i)}\right)$ multivariate normal likelihoods at each MCMC iteration. Unlike many file linking applications in which there are multiple records from each file in each block, for LFCMV this direct updating is fairly efficient in terms of computation time.

C Illustrative simulation

In this section we illustrate the performance of the LFCMV model. For this simulation, we use IAC data to create both F_1 and F_2. F_1 and the corresponding matches are constructed using records from 2007 and 2008, while the remainder of F_2 is constructed using records from the remaining IAC years. In Section C.1, we describe the process of creating F_1 and F_2 and introduce the BVs and MVs used in linking. Simulation results are presented in Section C.2.

C.1 Data

The complete downloaded IAC database (up through the year 2016) contains 17583 records with 56 variables. In this simulation, we use two categorical variables, NAICS code and state, as the BVs. The state variable refers to the state in which a manufacturing plant is located. NAICS code refers to a 6-digit numerical code corresponding to the specific products made by a given plant (U.S. Census Bureau, 2016). NAICS codes are nested up to the 4-digit level, meaning that all plants that produce a certain type of product, say dairy products, must agree on at least the first 4 digits of their NAICS codes. In this simulation we define blocks using state and the first 4-digits of the 6-digit NAICS code.

Two continuous variables, sales and number of employees, serve as our MVs. Sales is a continuous variable for the sales of each plant in U.S. dollars, and employees refers to the total number of employees for the plant. We let $j = 1$ refer to sales and $j = 2$ to number of employees.

C.1.1 Create F_1

To create F_1, we use the IAC records from 2007 and 2008, comprising 797 records. We reduce this set to the final set of 512 F_1 records using the steps outlined in the second column of Table 4. For this simulation, we use use 16786 IAC records from years other than 2007 or 2008 to create possible matches for F_1. If a record in F_1 has a state and 4-digit NAICS combination which is not found in these 16786 records, we remove these records from consideration for F_1. For simulation purposes, we also exclude from F_1 any records

Table 4: Summary of the process of creating F_1 and F_{2U}. The "Action" column describes the data cleaning step. The center column describes the number of records in F_1 after each cleaning step. The far right column describes the number of records in F_{2U} after each cleaning step. The final row gives the number of records in the F_1 and F_{2U} used for linking.

Action	F_1 Records	F_{2U} Records
Initialize	All 2007/2008: 797	All other years: 16786
Remove records from Puerto Rico	786	16779
Remove States Unique to F_1 or F_2	778	16440
Remove Missing States	778	13622
Remove Missing Sales	764	13476
Remove Missing Employees	764	13464
Remove Missing Electricity Usage	763	13429
Remove Missing Cost of Energy	763	13422
Require Matching State/4-Digit NAICS	512	1170
Final Total:	512	1170

containing missing data. The process of incorporating missing data imputation into the file linking process adds an additional level of uncertainty. The selection process for F_1 results in the final count of $n_1 = 512$ records.

C.1.2 Create F_2

We create F_2 in two stages. First, for each record i in F_1, we create a matching record by adding noise to the MVs in record i. The collection of these n_1 matches is denoted F_{2M}. Second, we use the 16786 IAC records from years other than 2007 or 2008 to create sets of possible matches for each record i in F_1. These records represent the false matches, or possible options for i based on state and NAICS agreement which are in fact not a match for i. The collection of these records is denoted F_{2U}. We let $F_2 = (F_{2U}, F_{2M})$.

We create F_{2M} as follows. For each record i in F_1, we generate a matching record using

the following steps. Assign each record i in F_1 to one of $H = 30$ latent classes z_i with

$$z_i \sim Multinomial(1; 1/H, \ldots, 1/H). \tag{19}$$

To generate the variance matrix, we sample from the following model:

$$\Sigma_h \sim InvWishart\left(50, G_{0h}\right), \tag{20}$$

where $G_{0h} = Diag(\phi_{1h}, \phi_{2h})$ and

$$\phi_{1h} \sim Gamma(1, 1), \phi_{2h} \sim Gamma(25, 3), \tag{21}$$

with $E(\phi_j) = a_\phi / b_\phi$. Let $\mathbf{Y}^{(0)}$ be the $n_1 \times p$ matrix containing all non-standardized MV values in F_1. For each i with latent class $z_i = h$, we define

$$F_{2M,i} = Y_i^{(0)} + \epsilon_i, \epsilon_i \sim N_2((0,0), \Sigma_{z_i}). \tag{22}$$

Here $F_{2M,i}$, $Y_i^{(0)}$ and ϵ_i are vectors of length $p = 2$. We repeat this match generating process 100 times for each i, creating 100 replicates $F_{2M,i}^{(m)}$, $m = 1, \ldots, 100$. Let $F_{2M}^{(m)} = \{F_{2M,i}^{(m)} \mid i = 1, \ldots n_1\}$.

We create F_{2U} using the 16786 IAC records from years other than 2007 or 2008. After removing missing data and other records as outlined in Table 4, we select 1170 records which have a state/4-digit NAICS combination that is observed in the F_1 data. Denote these 1170 records as F_{2U}. We let $F_2^{(m)} = (F_{2U}, F_{2M}^{(m)})$, where $m = 1, \ldots, 100$. For each replicate m, we link F_1 and $F_2^{(m)}$.

Each of the n_1 records from F_1 is assigned to a block (i). Based on the BVs, we assign each of the 1170 records in F_{2U} to these blocks. For each record $i \in F_1$, let $F_{2U(i)}$ denote the subset of F_{2U} assigned to block (i). As seen in Table 5, this results in blocks which range in size from 2 to 26. Blocks sizes, F_{2U}, and F_1 are consistent across replicates.

C.1.3 Standardization of Y and X

Before linking F_1 and $F_2^{(m)}$, we log transform the MVs. Let $\tilde{\mathbf{Y}}_j$ refer to the log-transformed MV values for field j in F_1, and $\tilde{\mathbf{X}}_j^{(m)}$ refer to the log-transformed MV values for field j in $F_2^{(m)}$. For ease of posterior computation in the linking model, we standardize $\tilde{\mathbf{Y}}_j$ and $\tilde{\mathbf{X}}_j^{(m)}$

Table 5: Block sizes for the illustrative simulation. Block Size: the number of records from F_{2U} assigned to each block (i). Count: the number of blocks of each block size.

Block Size	2	3	4	5	6	7	8	9	10	11	12	13	14
Count	129	106	76	39	20	17	30	14	11	23	4	4	10
Block Size	16	17	18	19	24	25	26						
Count	2	3	4	4	8	3	5						

as follows. Denote $S^{(m)} = (\tilde{\mathbf{Y}}_1, \tilde{\mathbf{X}}_1^{(m)})$ and $E^{(m)} = (\tilde{\mathbf{Y}}_2, \tilde{\mathbf{X}}_2^{(m)})$. Let $\bar{E}^{(m)}$ be the mean of all $n_1 + n_2$ elements of $E^{(m)}$ and let $\bar{S}^{(m)}$ be the mean of all $n_1 + n_2$ elements of $S^{(m)}$, with $sd(E^{(m)})$ and $sd(S^{(m)})$ representing the standard deviations of $E^{(m)}$ and $S^{(m)}$, respectively. We then standardize each transformed value as

$$\mathbf{Y}_1^{(m)} = (\tilde{\mathbf{Y}}_1 - \bar{S}^{(m)})/sd(S^{(m)}),\ \mathbf{X}_1^{(m)} = (\tilde{\mathbf{X}}_1^{(m)} - \bar{S}^{(m)})/sd(S^{(m)}), \tag{23}$$

$$\mathbf{Y}_2^{(m)} = (\tilde{\mathbf{Y}}_2 - \bar{E^{(m)}})/sd(E^{(m)}),\ \mathbf{X}_2^{(m)} = (\tilde{\mathbf{X}}_2^{(m)} - \bar{E}^{(m)})/sd(E^{(m)}). \tag{24}$$

C.2 Results

For each of 100 replicates, we apply the LFCMV model to link F_1 and the replicate specific $F_2^{(m)}$. We run the Gibbs sampler described in Section B.2 for 5000 iterations with a burn in length of 200 iterations. For purposes of comparison, we also link each pair of data sets using a "naive" matching method. Specifically, for the naive method, we select a match $i^{'} = C_i$ such that

$$C_i = \left(i^{'} \mid d(i, i^{'}) = min\left(d(i, \ell), \ell \in F_{2(i)}\right)\right), \tag{25}$$

where

$$d(i, i^{'}) = \sqrt{\sum_{j=1}^{p} \left(\frac{Y_{ij} - X_{i^{'}j}}{Y_{ij}}\right)^2}. \tag{26}$$

The primary metric used in this simulation is the match rate (MR), or the percentage of the n_1 records from F_1 that are correctly matched. For each replicate, the naive matching method yields one estimate of C. Conditional on this C, we compute a single estimate of

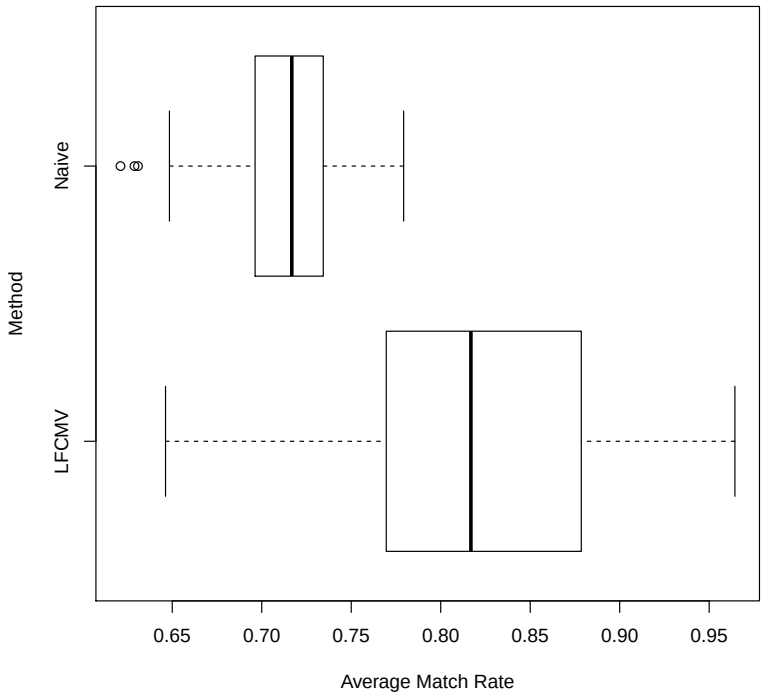


Figure 3: Match rates. Upper Boxplot: Match rate for each replicate obtained by linking records by the naive method. Lower Boxplot: Composed of average MR for each of 100 replicates of LFCMV.

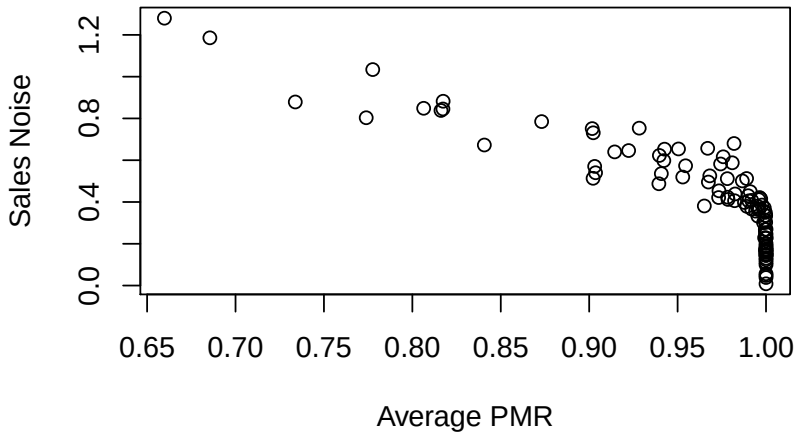


Figure 4: Average posterior match rates by average distance across matches pairs for sales from the 100 replicates. Lower PMR values are associated with greater distance.

the match rate. As a Bayesian procedure, LFCMV yields a set of posterior draws for C. For each replicate, we compute the MR for each posterior draw of C. We then average these MR values to obtain one estimate of the posterior MR for each replicate.

Results are displayed in Figure 3. For all but 2 of the 100 replicates, the MR obtained using LFCMV is higher than the MR obtained using the naive approach. A 95% posterior interval for the improvement in MR using LFCMV over the naive method is (.2%, 22%) with an average increase in match rate of 9%. This translates to an average of 47 additional records that are correctly matched using LFCMV.

As seen in Figure 3, the range of MR values is fairly wide. This range is due to the process of generating the X_i values. The average distance across matched pairs, i.e., the average value of $Y_{ij} - X_{C_i j}$, is different across each replicate. Consider Figure 4. Replicates with smaller average PMR have, on average, more distance between the MV values for sales across matched pairs. By design, LFCMV is searching for a match with small distance between Y_i and X_{C_i}. If the data generation process creates a matching X_{C_i} which is far from Y_i, the model sometimes selects records $X_{i' \neq C_i}$ with a smaller distance between Y and X. This behavior is consistent with the intuition discussed in Section A.3.

D Sensitivity to distance across matched pairs

In this section, we illustrate the performance of LFCMV linkage when the mechanism that introduces distance across matched pairs of MVs is not normally distributed. The creation of F_1 and F_{2U} is the same as the simulation study in Section C. However, we vary the process of creating F_{2M} by using different distributions to create the distance across matched pairs. In Section D, we apply uniform distance at three levels: 10%, 20% and 30%. For this simulation, we are interested in examining the performance of LFCMV with increasing distances between matches pairs. In Section D.2, we introduce distance across matched pairs dependent upon the number of employees in a plant. This simulation leverages the ability of LFCMV to allow the distance between matched pairs to vary across types of pairs.

D.1 Uniform distance

In Section C.2, the MR obtained by LFCMV tends to vary with the distance across matched pairs. To further explore this concept, we conduct a simulation in which distance between matched Y and X is generated uniformly at three different levels. Specifically, for each replicate, we generate $F_{2M}^{(m)}$ using

$$F_{2M,i1}^{(m)} = Y_{i1}^{(0)}(1+\kappa_{ij}), \kappa_{ij} \sim Uniform(0,u) \tag{27}$$

$$F_{2M,i2}^{(m)} = Y_{i2}^{(0)}(1+\kappa_{ij}), \kappa_{ij} \sim Uniform(0,u), \tag{28}$$

where $u = .1, .2$, and $.3$. We run 100 replicates at each setting of u.

Figure 5 displays the results across all three settings of u. As is true in Section C, LFCMV generally yields higher match rates than the naive approach, and this result holds across all three settings of u. As u increases and true matched pairs have a greater distance between the MVs of matched pairs, both the naive and LFCMV linking methods have reduced performance. As previously mentioned, block size also has an impact on MR. Table 6 shows the breakdown of posterior MR by block size. The average PMRs for smaller blocks tend to be higher than average PMR for larger blocks, with some variation attributable to the number of blocks of a given size.

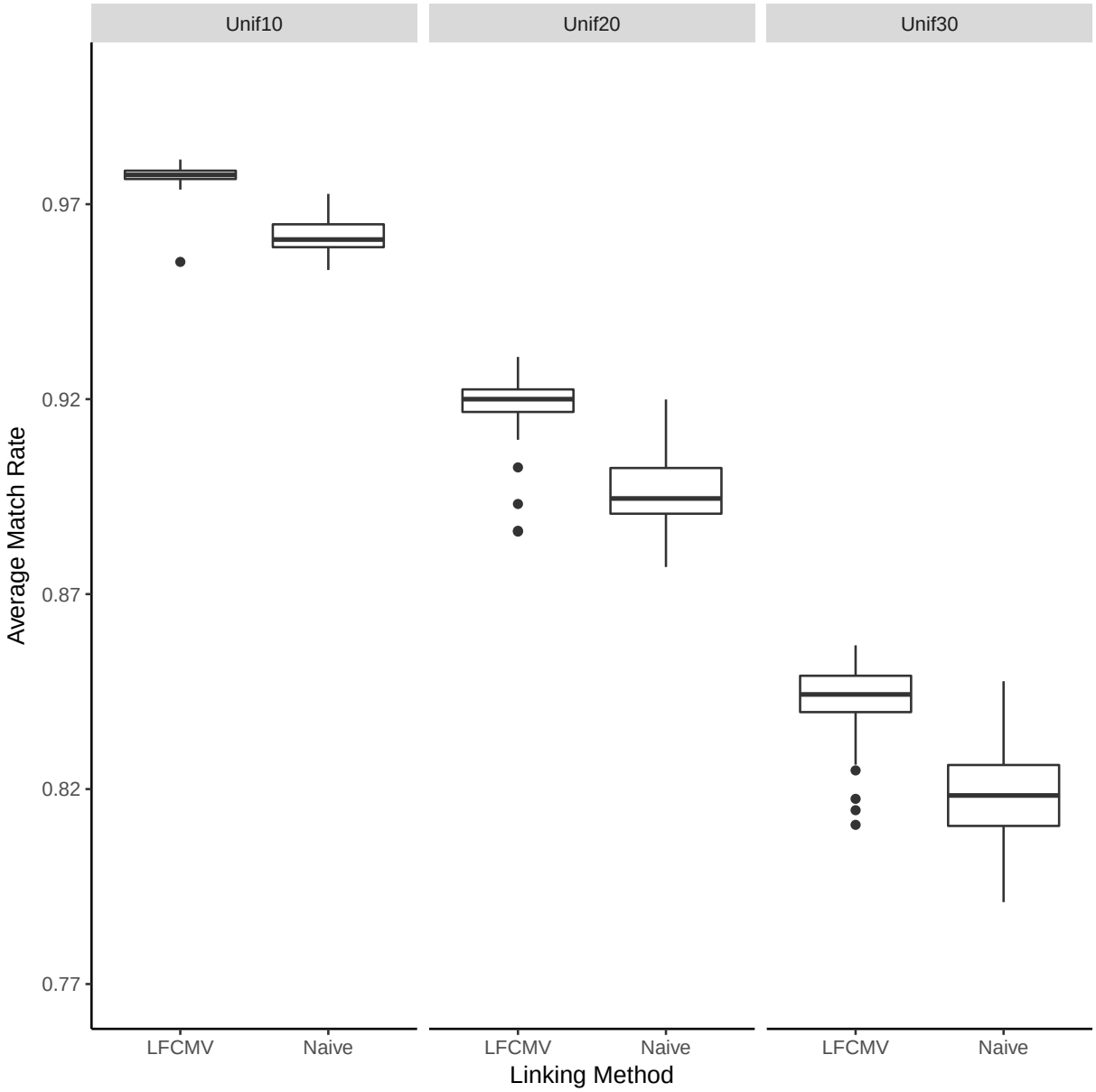


Figure 5: Match rates under uniform distance. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.

Table 6: Average posterior match rate for each block size. The first column indicates the size of the block, with the number of blocks of that size in parentheses. Block sizes are consistent for each simulation. Remaining columns denotes the average posterior MR, as computed by averaging the block-specific MR across each of the 100 replicates.

Block Size (Count)	10%	20%	30%
2 (*129*)	.99	.98	.94
3 (*106*)	.99	.97	.93
4 (*76*)	.99	.94	.88
5 (*39*)	.98	.92	.80
6 (*20*)	.97	.90	.82
7 (*17*)	.98	.90	.78
8 (*30*)	.95	.89	.80
9 (*14*)	.91	.78	.66
10 (*11*)	.82	.56	.41
11 (*23*)	.93	.79	.64
12 (*4*)	1	1	1
13 (*4*)	.97	.85	.67
14 (*10*)	.99	.89	.72
16 (*2*)	1	.80	.57
17 (*3*)	1	.98	.82
18 (*4*)	1	.89	.61
19 (*4*)	.91	.50	.30
24 (*8*)	.99	.69	.40
25 (*3*)	1	.83	.57
26 (*5*)	1	.95	.80
Overall PMR	.98 (.003)	.92 (.007)	.84 (.008)

D.2 Size dependent distance

One of the features of LFCMV is the ability to model distances between the MVs that vary with types of matched pairs. To illustrate this capability, we conduct a simulation in which we introduce correlation between the number of employees in a plant and the distance across matched pairs of MVs. We classify the 261 plants in F_1 with more than 130 employees as "large" and classify the remaining 251 plants in F_1 with 130 or fewer employees as "small". We run a simulation, denoted LD, in which MVs for matched pairs corresponding to large plants tend to be farther apart than the MVs for matched pairs corresponding to small plants. We then run a second simulation, denoted SD, in which the opposite is true. For simulation LD, we generate $F_{2M}^{(m)}$ from

$$F_{2M,i,j}^{(m)} = Y_{ij}(1+\kappa_{ij}), \kappa_{ij} \overset{iid}{\sim} \begin{cases} Uniform(0,.4) & : \text{ if } Y_{i2}^{(0)} > 130 \\ Uniform(0,.15) & : \text{ otherwise.} \end{cases} \tag{29}$$

Similarly, for simulation SD, we generate $F_{2M}^{(m)}$ from,

$$F_{2M,i,j}^{(m)} = Y_{ij}(1+\kappa_{ij}), \kappa_{ij} \overset{iid}{\sim} \begin{cases} Uniform(0,.4) & : \text{ if } Y_{i2}^{(0)} \leq 130 \\ Uniform(0,.15) & : \text{ otherwise.} \end{cases} \tag{30}$$

We run 100 replicates of the LD and SD simulation settings.

With these simulations, we are interested in assessing two aspects of LFCMV. First, we examine the match rate for different types of matched pairs for both the SD and LD simulations. The matching results are summarized in Table 7. As with the results from Section C, larger distance across matched pairs tends to be associated with lower MRs than smaller distance across matched pairs. However, an examination of the match rate by plant size reveals that the naive approach and LFCMV perform differently in the different size groups. Consider Figure 6, displaying the match rates for the LD replicates. For large plants, which have greater distance across matched pairs in this simulation, LFCMV matches more accurately than the naive approach. However, the opposite trend is evident for smaller plants. This suggests that LFCMV overestimates the distance reasonable for matched pairs in the smaller plants, resulting in a lower match rate. Figure 7 shows similar tendencies for the SD simulation; LFCMV's match rate is better on average for the smaller plants, with the naive approach matching more accurately for the larger plants

Table 7: Average posterior match rate for the SD and LD simulations. Bold values indicate the PMR associated with the plant size with smaller distance between matched pairs in each simulation. Large plants: the PMR for large plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Small plants: the PMR for small plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Overall PMR: the PMR for all plants, averaged over the 100 replicates of the SD and LD simulations, respectively. Values in parentheses represent the standard error.

	SD	LD
Large plants	**.90**(.008)	.78(.01)
Small plants	.76(.009)	**.88**(.003)
Overall PMR	.85(.006)	.84(.006)

with less distance across matched pairs. Despite this behavior, LFCMV results in an overall improvement in match rate over the naive approach in both the SD and LD replicates.

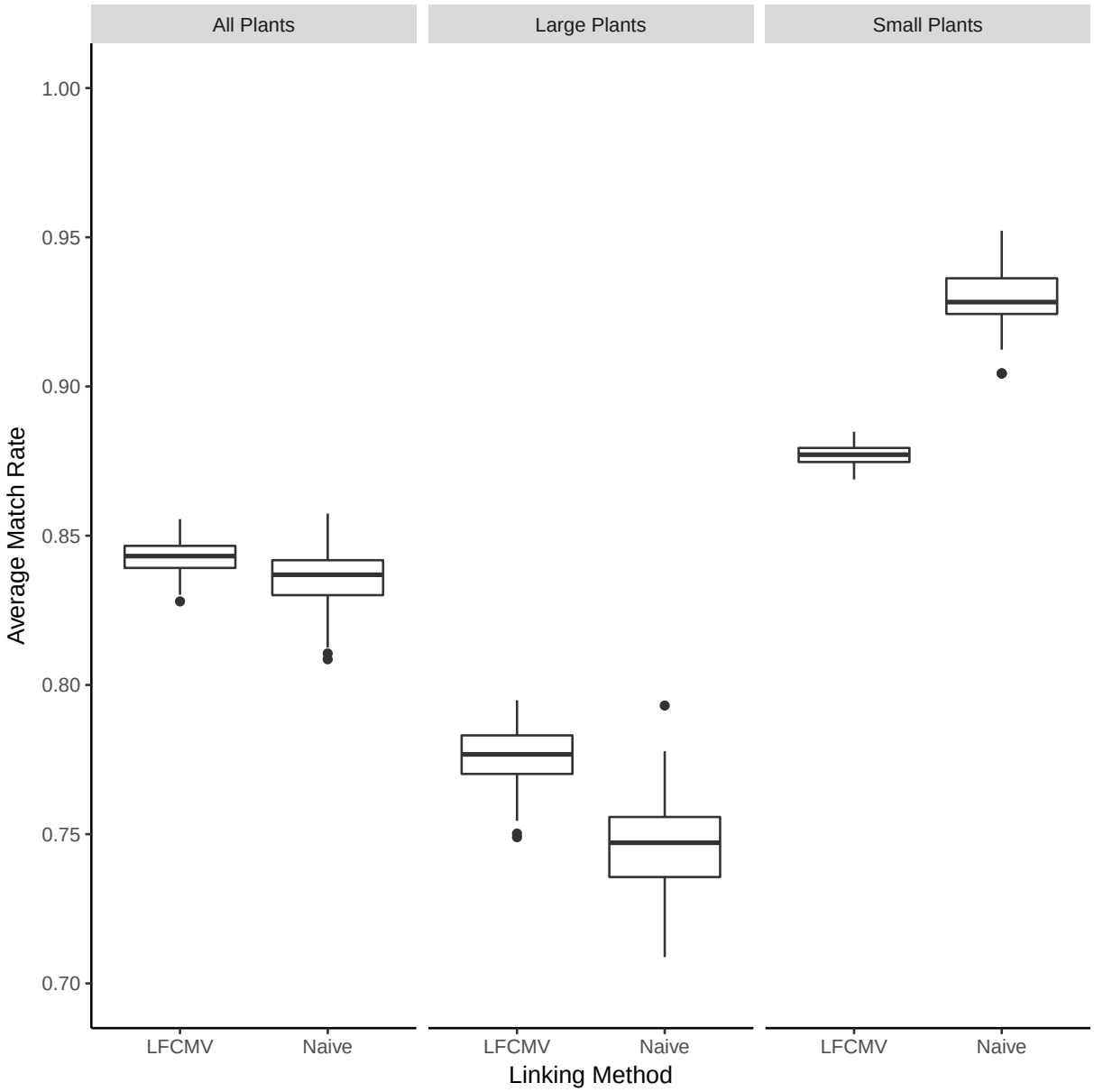


Figure 6: LD: Match rates under when large plants have greater distance across matched pairs than small plants. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.

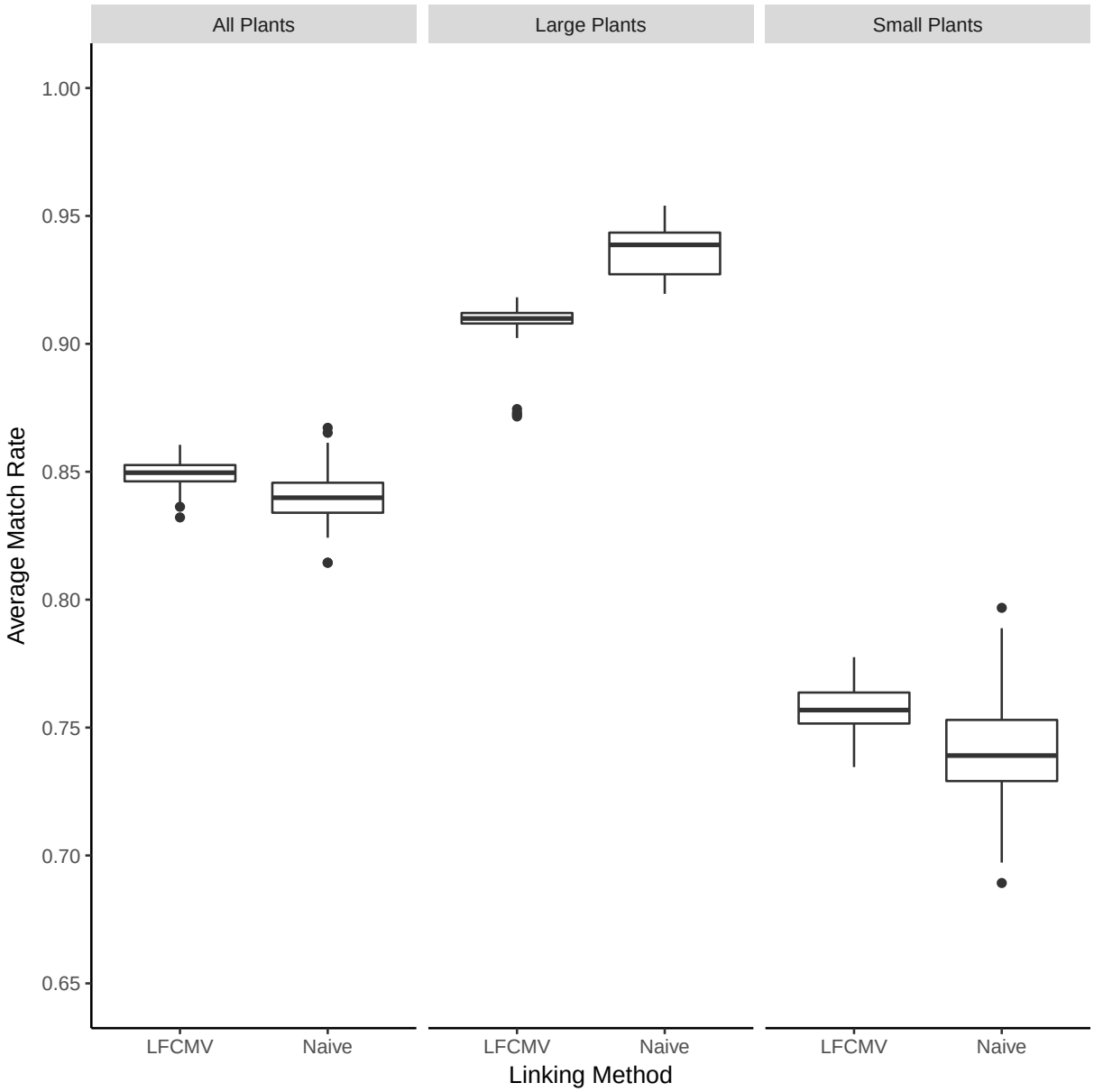
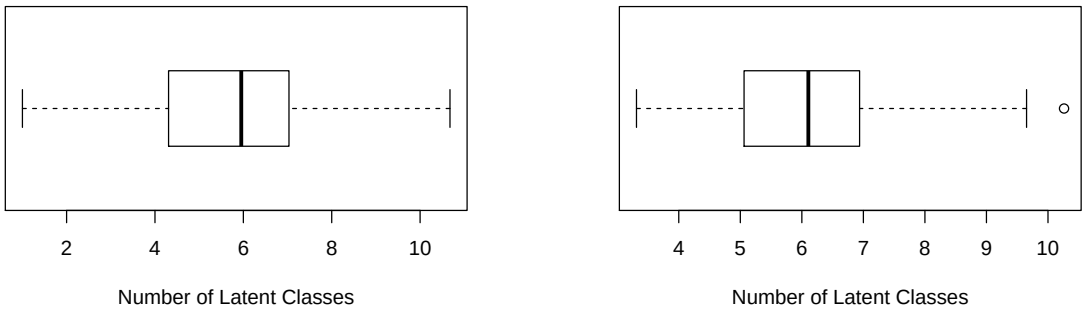


Figure 7: SD: Match rates under when small plants have larger distance across matched pairs than larger plants. Left Boxplot in each panel: Composed of average PMR for each of 100 replicates. Right Boxplot in each panel: Match rate for each replicate obtained by linking records by the "naive" approach, i.e., minimizing $d(i, i^{'})$.



(a) SD (b) LD

Figure 8: Average number of latent classes for the SD and LD simulations. For each of the 100 replicates of simulation SD (Figure 8a) and LD (Figure 8b), we compute the number of latent classes at each posterior iteration with more than 5 records. We then average these counts to obtain a single estimate of the average number of occupied latent classes for each replicate. The boxplots are composed of the 100 replicate-specific counts.

The second aspect of LFCMV we illustrate with the SD and LD simulations is the ability of LFCMV to capture the relationships inherent in the different types of matched pairs using the mixture model formulation of the linking model. In the simulation studies described in Section D, the distance across matched pairs is uniform, so latent classes are not needed to describe the distribution of distances across matched pairs. Accordingly, the number of occupied mixture components collapses to a single component during the posterior sampling. However, for the SD and LD simulations, there are two clearly defined types of matched pairs, as for each simulation, small and large plants are associated with a certain degree of distance across matches. Accordingly, Figure 8 indicates that multiple mixture components are occupied in the SD and LD simulations. In Figure 9, we examine these components for a single iteration of a replicate of the SD simulation. The figure indicates that latent class 1 is associated with smaller plants than latent class 2, while latent class 3 contains a blend of plant sizes. As distance across matched pairs is directly related to plant size, these results suggest that the latent class structure of the linking model may be useful for capturing features in the data associated with differing amounts

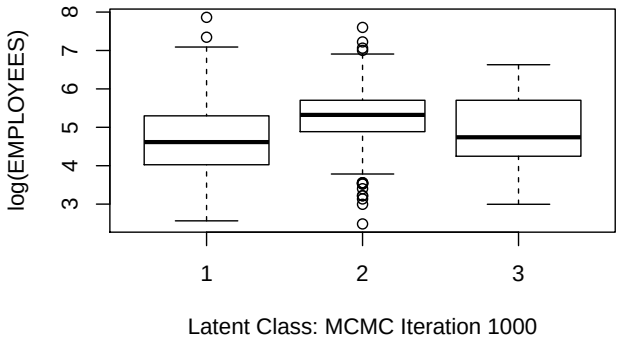


Figure 9: SD: The box plots display the log(employees) value for the three latent classes at MCMC iteration 1000. Each boxplot represents a latent class, and the y-axis represents the value of $\tilde{\mathbf{X}}_2$, denoted log(employees), for each record in that latent class. A paired t-test comparing the log(employees) value of latent classes 1 and 2 leads statistically significant results ($p < 0.05$). Specifically, latent class 1 tends to contain smaller plants than latent class 2.

of distance across matched pairs.

D.3 Summary of findings

The simulations conducted in Sections D and D.2 highlight a few features of LFCMV linking. First, as is evident in Section C.2, large distances among the MVs for matched pairs can lead to a reduction in PMR. This underscores the importance of the considerations discussed in Section A.3. Second, the simulations in Section D.2 provide an example the utility of the latent class structure of the linking model. The latent class structure provides the potential for distribution in distance to vary across matched pairs. If such a structure is unnecessary in an application, the sampler tends to collapse to using single class to estimate $\boldsymbol{\Sigma}$.

E Extended Results and Discussion

In this section, we present supplemental plots to Section 4 in the main text. In Section 2.1 of the main text, we note that blocking on state and 6-digit NAICS codes leads to a number of IAC records with no possible matches in the CMF. The IAC records are a subset of the CMF, and as such all records in the IAC should have a match in the CMF. The lack of matches for some records after direct blocking suggests that some of the NAICS codes in IAC or CMF may have some uncertainty. To account for this, we considering blocking on three different levels of NAICS agreement. Specifically, we create subsets with agreement on 4-digit, 5-digit and 6-digit NAICS codes. We refer to these as levels of NAICS blocking. NAICS codes are nested up to the 4-digit level, meaning that all plants that produce a certain type of product, say dairy products, must agree on at least the first 4 digits of their NAICS codes. However, the results of our preliminary analysis suggest no difference in results across the linked data products created at each level of NAICS blocking. We therefore present results only from linking at the 6-digit NAICS level; we present these results here.

The results in Table 8 and Figures 10 through 13 were produced by the same methods described in Section 4 of the main text. However, blocking was performed on either 4-digit or 5-digit NAICS code agreement, rather than the 6-digits codes as in the main text. As is evident in the table and figures, the results of the analysis are comparable across the three levels of NAICS blocking.

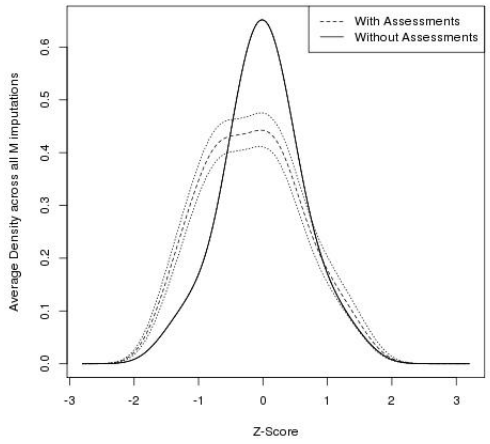


Figure 10: NAICS 4 Blocking, 2007: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

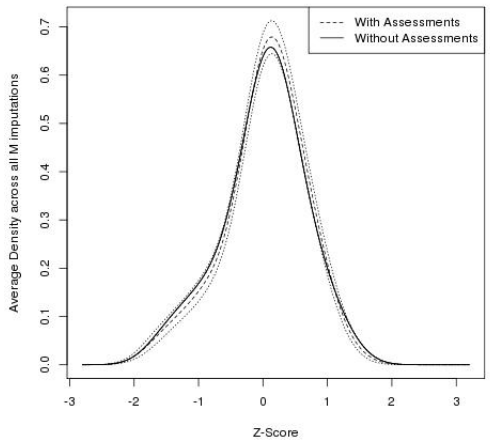


Figure 11: NAICS 4 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

Table 8: Displays 95% MI intervals for the mean of the "With Assessments" versus "Without Assessment" group at each level of NAICS blocking. Similar trends are evident across all three levels of NAICS blocking.

NAICS4		Mean	95% Interval
2007	With Assessments	-0.18	(-.27,-.09)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.03	(-0.04,0.11)
	Without Assessments	0.008	(0.006,0.011)
NAICS5		Mean	95% Interval
2007	With Assessments	-0.17	(-0.26,-0.08)
	Without Assessments	-0.007	(-0.01,-0.005)
2012	With Assessments	0.02	(-0.06,0.10)
	Without Assessments	0.009	(0.006,0.011)
NAICS6		Mean	95% Interval
2007	With Assessments	-0.18	(-0.29,-0.08)
	Without Assessments	-0.007	(-0.009,0.005)
2012	With Assessments	0.04	(-0.05,.12)
	Without Assessments	0.009	(0.06,.011)

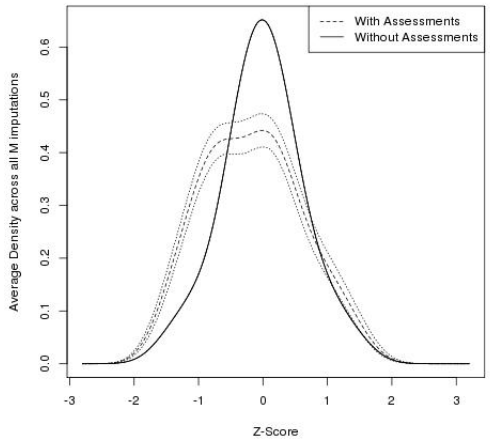


Figure 12: NAICS 5 Blocking, 2007 Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.

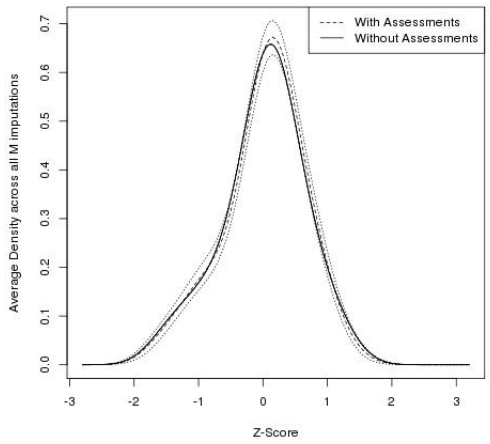


Figure 13: NAICS 5 Blocking, 2012: Kernel Density Plots. Solid line: Average Density of energy scores for CMF plants that did not receive assessments. 95% intervals are displayed but are very tight to the curve. Dashed lines: Average Density of energy scores for CMF plants that did receive assessments. 95% intervals are displayed. The fuzziness in the plot results from disclosure review requirements of the Census Bureau.