The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1997–2020
DOI: 10.1214/16-AOAS962

MAXIMIZING THE INFORMATION CONTENT OF A BALANCED MATCHED SAMPLE IN A STUDY OF THE ECONOMIC PERFORMANCE OF GREEN BUILDINGS[1]

BY CINAR KILCIOGLU AND JOSÉ R. ZUBIZARRETA

Columbia University

Buildings have a major impact on the environment through excessive use of resources, such as energy and water, and large carbon dioxide emissions. In this paper we revisit a previously published study about the economics of environmentally sustainable buildings and estimate the effect of green building practices on market rents. For this, we use new matching methods that take advantage of the clustered structure of the buildings data. We propose a general framework for matching in observational studies and specific matching methods within this framework that simultaneously achieve three goals: (i) maximize the information content of a matched sample (and, in some cases, also minimize the variance of a difference-in-means effect estimator); (ii) form the matches using a flexible matching structure (such as a one-to-many/many-to-one structure); and (iii) directly attain covariate balance as specified—before matching—by the investigator. To our knowledge, existing matching methods are only able to achieve, at most, two of these goals simultaneously. Also, unlike most matching methods, the proposed methods do not require estimation of the propensity score or other dimensionality reduction techniques, although with the proposed methods these can be used as additional balancing covariates in the context of (iii). Using these matching methods, we find that green buildings have 3.3% higher rental rates per square foot than otherwise similar buildings without green ratings—a moderately larger effect than the one found by the prior study.

1. Introduction: Green buildings, buildings data, overview of matching, outline.

1.1. *Market performance of environmentally sustainable buildings.* Buildings have a major impact on the environment through greenhouse gas emissions and excessive use of natural resources. For example, the United States Environmental Protection Agency (EPA) reported that in 2013 nearly 39% of total U.S. carbon dioxide emissions were due to residential and commercial buildings.[2] For the same year, the U.S. Energy Information Administration reported that about 40% of total

Received June 2015; revised June 2016.

[1]Supported in part by a grant from the Alfred P. Sloan Foundation.

Key words and phrases. Causal inference, matched sampling, observational studies, propensity score.

[2]http://www.epa.gov/climatechange/Downloads/ghgemissions/US-GHG-Inventory-2015-Main-Text.pdf, Table ES-7.

U.S. energy consumption was from these types of buildings.[3] At the same time, there is growing scientific consensus that current levels of carbon dioxide and related greenhouse gas emissions greatly increase the risks of climate change, and that excessive use of resources can lead to resource depletion and habitat degradation. For these reasons, the construction and operation of buildings can have a substantial impact on the earth's environment.

In an interesting and relevant study, Eichholtz, Kok and Quigley (2010) analyzed the effect of environmentally sustainable building practices on their rents and selling prices. This is an important study subject for the reasons already stated and also because there is not much empirical evidence for the development of environmentally sustainable or green buildings. Among the available evidence, there are the results of a study by the U.S. General Service Administration Public Buildings Service that analyzed the performance of 22 green buildings and found that, compared to national averages, green buildings have 36% fewer carbon dioxide emissions and 25% less energy use, in addition to 19% lower aggregate operational costs and 27% higher occupant satisfaction.[4] Given the environmental and social benefits of green buildings, one important question is how much these benefits affect the rent of green commercial buildings. This is important to investors, developers and property owners in order to invest in green buildings.

In their study, Eichholtz, Kok and Quigley (2010) analyzed a large sample of commercial green- and nongreen-rated buildings in the United States. Using linear regression and propensity score methods, they found that buildings with green ratings have approximately 2.8% higher rental rates per square foot compared to similar buildings without green ratings. In this paper, we revisit this important question using new matching methods that adjust more precisely for covariates and better exploit the structure of the buildings data.

1.2. *Buildings data*. In the United States, green buildings are certified as energy efficient or sustainable by different agencies. The EPA gives the "Energy Star" certification to commercial buildings if their amount of energy used meets certain criteria.[5] The Green Building Council (USGBC) labels a building as LEED (Leadership in Energy and Environmental Design) based on its performance in different categories such as indoor environmental quality, site sustainability and water conservation. Following Eichholtz, Kok and Quigley (2010), we consider a building to be green if it is certified as Energy Star or LEED and focus our analysis on commercial buildings.

[3] http://www.eia.gov/totalenergy/data/monthly/pdf/mer.pdf, Table 2.1.

[4] http://www.gsa.gov/graphics/pbs/Green_Building_Performance.pdf.

[5] Specifically, the EPA can give the "Energy Star" certification to buildings in the top quarter of energy efficiency compared to similar buildings nationwide. The energy efficiency calculation is done by the EPA using a scoring algorithm that takes into account the characteristics of the building, such as size, location and number of occupants.

To estimate the effect of energy efficiency and sustainability on the economic returns of buildings, we compare green-rated buildings to similar nongreen-rated buildings in the same market. For this, we use multivariate matching methods and find matches of green and nongreen buildings that are nearby and similar along a number of covariates, including age, amenities, number of stories, quality and whether the building was recently renovated. However, standard matching methods do not have the flexibility to exploit the particular structure of the buildings data and will typically result in imbalanced or inefficient analyses. In particular, the data consists of 694 green buildings and 7411 nongreen buildings, organized in 694 geographic clusters. In each of these clusters, there is one green building and one or more nongreen buildings not further apart than one quarter mile from the green building. While some clusters have only one nongreen building, others have as many as 83 nongreen buildings. As a result of this structure, pair matching (or matching with a $1:1$ ratio) would result in many nongreen buildings not being used in the analysis, and matching with a fixed $1:\kappa$ ratio (where κ is an integer greater than 1) would result in some clusters not being used at all. Naturally, we would like to use a flexible matching ratio in order to match as many buildings as possible, while precisely balancing covariates. However, to our knowledge, existing matching methods are not able to achieve all of these goals simultaneously. In the following section, we give an overview of standard matching methods, and then, in the next section, explain more carefully the contribution of the proposed methods.

1.3. *Overview of matching in observational studies.* In observational studies, matching methods are often used in an attempt to compare like with like, that is, units that are the same ideally in every respect except in their assignment to a treatment [Cochran and Rubin (1973)]. In our study, these units are buildings similar in terms of age, amenities, number of stories, etc., except in their green building practices. Of course, this comparison can be assessed in terms of observed covariates only, and with matching methods (the same as with other regression or weighting methods of adjustment for observed covariates), the question about the influence of unobserved covariates in effect estimates remains open [see, for instance, Chapter 4 of Rosenbaum (2002) for a formal discussion]. With standard matching methods, other devices such as differential effects, evidence factors, multiple control groups and sensitivity analyses can be used to limit and assess the influence of such unobserved covariates [see Rosenbaum (2015) for a review of these devices].

The appeal of matching as a method of adjustment lies in part in its conceptual simplicity [comparing like with like while keeping the unit of analysis intact; Rosenbaum and Silber (2001)], that its adjustments are an interpolation instead of an extrapolation based on a parametric model [Rosenbaum (1987), Imbens (2015)], and in the fact that it is conducted without using outcomes, thus preventing exploratory expeditions in the data to choose the adjustments that better suit the hypotheses of the investigation [Rubin (2008)]. It is for this last reason that

matching is considered to be part of the design as opposed to the analysis of an observational study [Rosenbaum (2010)]. However, some matching methods are cumbersome in practice.

The main goal of matching is to find matched groups with similar or balanced observed covariate distributions [Stuart (2010)]. Ideally, these groups would be formed by units identical in every way (by "clones" of treated and control units), but usually this is not feasible in practice. There is a course of dimensionality in exact matching: as the number of observed covariates increases, there is a combinatorial explosion in the resulting types of units. Thus, for an observational study of the typical size (like our building study with a few thousand observations), there will not be enough units to match each treated unit to one control exactly. It is for this reason, and also because randomization does not produce exact matches but balance in expectation, that weaker, aggregate forms of covariate balance than exact matching tend to be pursued in practice, leaving exact matching for a few covariates of overriding prognostic importance [see Sections 3.3 and 9.3 of Rosenbaum (2010)]. The propensity score [Rosenbaum and Rubin (1983)] is an important tool used to achieve aggregate covariate balance.

The propensity score is the probability of treatment assignment given the observed covariates. It constitutes a dimensionality reduction technique in which a P-dimensional observed covariate is summarized into a single scalar with important theoretical properties. Informally, Theorems 1 and 3 in Rosenbaum and Rubin (1983) state that matching on the propensity score tends to balance the P observed covariates used to estimate the score, and that for balancing the P covariates, it suffices to balance the one-dimensional propensity score. However, these are stochastic properties that hold over repeated realizations of the data-generation mechanism, and, for a given realization (that is, for a given data set), even if the true treatment assignment is known, it is not certain that the propensity score will balance the observed covariates [especially if the covariates have many categories or are sparse; see, e.g., Yang et al. (2012) and Zubizarreta et al. (2011)]. Also, in practice, the true assignment mechanism is unknown, and this makes the task of balancing the observed covariates even more difficult due to misspecification of the propensity score model. Furthermore, while matching on the propensity score is typically used for balancing means, in some settings it is desirable to balance other features of the distribution of the P observed covariates, such as its marginal distributions, and this can be very difficult by matching on the propensity score. It is for these reasons that matching on the propensity score involves a considerable amount of guesswork in practice.

A recent method that speaks to these limitations is the covariate balancing propensity score [Imai and Ratkovic (2015)]. Under a generalized method-of-moments or empirical likelihood framework, this method estimates the propensity score penalizing fits for which the covariate distributions are not balanced. In a similar way to the propensity score, the covariate balancing propensity score can be used for matching and weighting.

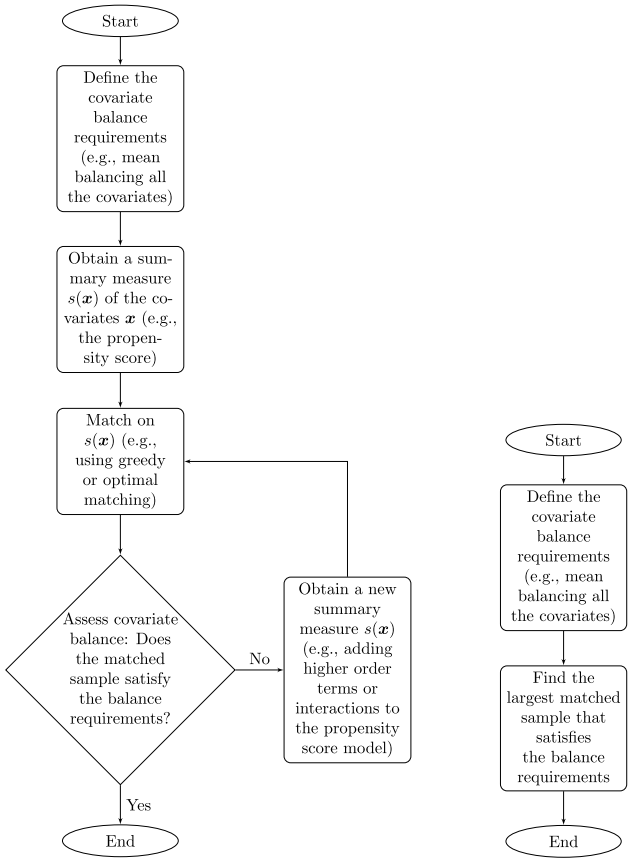
Related weighting methods include inverse probability tilting [Graham, De Xavier Pinto and Egel (2012)], entropy balancing [Hainmueller (2012)], the stable balancing weights [Zubizarreta (2015)], calibration weighting [Chan, Yam and Zhang (2016)] and the overlap weights [Li, Morgan and Zaslavsky (2016)]. However, two differences between weighting and matching (as typically used in statistics) are that in matching the unit of analysis typically remains intact, facilitating simpler outcome analyses and complementary, qualitative descriptions of the data [Rosenbaum and Silber (2001)], and that the structure of the data after the matched adjustments resembles more closely that of a randomized experiment, facilitating more transparent sensitivity analyses to hidden biases [Rosenbaum (2002)].

Other recent matching methods include the following: coarsened exact matching [Iacus, King and Porro (2012)], which matches exactly treated and control units after coarsening the observed covariates; genetic matching [Diamond and Sekhon (2013)], which uses a genetic search algorithm to maximize a measure of covariate balance consisting of several statistics; balance optimization subset selection [Nikolaev et al. (2013)], which minimizes a global imbalance measure subject to equally sized matched groups; and optimal matching with refined covariate balance [Pimentel et al. (2015)], which uses network flow algorithms to balance nominal covariates and their interactions according to a prioritized list.

Another recent matching method that addresses the aforementioned limitations of matching based on the propensity score is optimal cardinality matching, or cardinality matching for short [Zubizarreta, Paredes and Rosenbaum (2014)]. Unlike the previous matching methods, cardinality matching solves an integer programming problem to maximize the cardinality or size of a matched sample subject to more flexible constraints on covariate balance. In their weakest form, these constraints can require the means to be balanced [see Zubizarreta (2012) for details], but they can also require other forms of distributional balance such as fine balance [Rosenbaum, Ross and Silber (2007)] and strength-k balancing [Hsu et al. (2015)]. In this way cardinality matching balances the covariates directly.

The flowcharts in Figure 1 compare the basic steps involved in cardinality matching and in more standard matching methods based on the propensity score or other summary measures of the observed covariates (such as the Mahalanobis distance). While these standard matching methods can entail many iterations to meet the covariate balance requirements by fine-tuning the summary measure, cardinality matching directly finds the largest matched sample that meets these requirements. In a sense, with cardinality matching, subject matter knowledge of the scientific question at hand comes naturally into the matching problem through the balancing constraints, finding the largest matched data set that satisfies the investigator's specifications for covariate balance or comparability between treated and control units.[6] In contrast, with cardinality matching the possibility of covariate

[6]For simplicity, in Figure 1(a) we omit the decisions involved in propensity score matching about overlap, but typically additional steps would be present [see, e.g., Chapter 16 of Imbens and Rubin (2015)].



(a) Common matching methods (b) Cardinality matching

FIG. 1. *Flowcharts of common matching methods and cardinality matching.*

distributions exhibiting limited overlap is addressed in terms of the original covariates, finding the largest match that meets the investigator's specifications for covariate balance.

1.4. *Outline.* To analyze the effect of energy efficiency and sustainability on the economic returns of buildings, in this paper we build on the method of cardinality matching and propose a general matching framework to maximize the information content of a balance matched sample. Within this framework, we present new matching methods that simultaneously achieve three goals: (i) to maximize the information content of a matched sample and, in some cases, minimize the variance of a widely used effect estimator; (ii) to form the matched groups of the matched sample using a flexible matching structure [such as a one-to-many/many-to-one or, in a sense, a full matching structure; Rosenbaum (1989), Hansen (2004)]; and (iii) to directly attain covariate balance as specified—before matching—by the in-

vestigator. On the one hand, standard matching methods such as the ones illustrated in Figure 1(a) are not designed to achieve goals (i) and (iii), but, on the other hand, cardinality matching does not allow flexible matching structures beyond a one-to-many fixed matching ratio. Achieving these three goals simultaneously poses a number of difficulties. First, maximizing the size of a matched sample with a flexible matching ratio requires a different notion of sample size than the one used in cardinality matching, since, for instance, two one-to-one treated and control matches should not count the same as one one-to-two match. This requires defining the information content of the matched sample. Second, the differential weighting of the different matched groups needs to be taken into account when assessing covariate balance and in the analyses, but this poses a number of challenges in building a mathematical program and in computing its optimal solutions. Third, a sound implementation of this method needs to take advantage of modern advancements in parallel computing.

This paper is organized as follows. In Section 2 we review cardinality matching, discuss different matching structures, and finally present a definition of the information content of a matched sample for a simple difference-in-means effect estimator. In Section 3 we first introduce a general framework for matching to maximize the information content of a balanced matched sample, then show that cardinality matching is a particular case of this framework, and present a formulation for matching with a variable one-to-many ratio (in two other appendices, we present formulations for matching to minimize the variance of the difference-in-means effect estimator and matching with a flexible one-to-many/many-to-one or full matching structure). In Section 4 we evaluate the building matches in terms of covariate balance and effective sample sizes, and also describe the details of the computational implementation. In Section 5 we investigate the economic effects of green buildings. In Section 6 we discuss the proposed matching methods. In Section 7 we close with a summary and remarks.

2. Review: Cardinality matching, matching structures, information content.

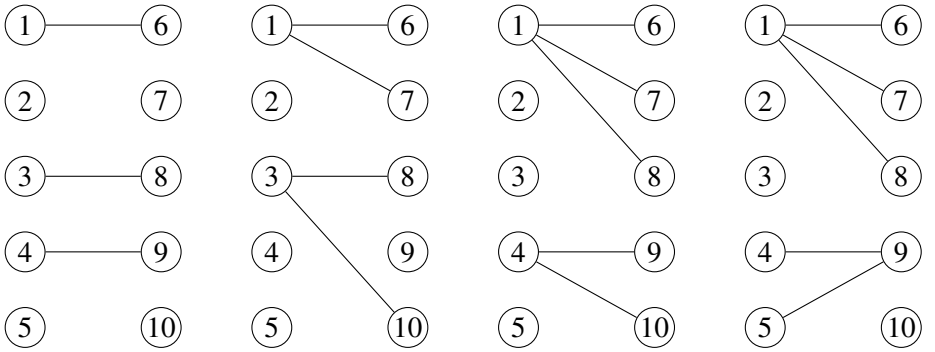
2.1. *Cardinality matching*. As described above, cardinality matching uses the original covariates instead of a summary of them to match units and directly balance their distributions [Zubizarreta, Paredes and Rosenbaum (2014)]. Specifically, cardinality matching finds the *largest* matched sample that satisfies the investigator's specifications for covariate balance. For example, cardinality matching will find the largest matched sample in which all the marginal distributions of the covariates are balanced. In this manner, cardinality matching first focuses on covariate balance in aggregate, allowing the investigator to then rematch the treated and control units in the balanced matched sample to emphasize covariates that are strongly correlated with the outcome. As illustrated in Zubizarreta, Paredes and

Rosenbaum (2014), this has the effect of reducing the heterogeneity of matched-group differences in outcomes and, in turn, also reducing sensitivity to biases due to unmeasured confounders [see Rosenbaum (2005) for a detailed exposition of this argument and Baiocchi (2011) for an original alternative approach].

From a computational standpoint, cardinality matching requires solving a linear integer programming problem, and while it has not been found that a polynomial time algorithm solves the cardinality matching problem, there is considerable structure in this problem and many instances of it can be solved in time that from a user perspective is comparable to that of common matching methods [see Appendix A in the supplemental article Kilcioglu and Zubizarreta (2016)]. At present, the cardinality matching problem can be solved with the optimization solvers CPLEX, GLPK, Gurobi and Symphony via the statistical package `designmatch` for `R` [Zubizarreta (2012), Zubizarreta and Kilcioglu (2016)].

2.2. *Matching structures*. In its simplest form, a matched sample is assembled by pairs of treated and control units selected from larger reservoirs. As in our buildings study, the reservoir of controls is often much larger than the one of the treated units, and it is feasible to match more than one control to each treated unit. One possible way of doing this is by matching with a fixed $1:\kappa$ ratio, and either matching each treated unit to κ controls or not matching it at all. A more flexible structure is a variable $1:\kappa$ ratio, in which each treated unit is matched at most to κ controls (if matched at all). The most flexible structure is matching with a one-to-many/many-to-one structure or, loosely speaking, full matching [Hansen (2004), Rosenbaum (1989)]. (In rigor, the term full match refers not only to a one-to-many/many-to-one structure, but also to an optimal design for an observational study in which all the treated units are matched to controls forming groups as similar as possible in terms of a summary of the covariates, $s(\boldsymbol{x})$; see Section 10.3.6 of Rosenbaum (2002) In this sense, a one-to-many/many-to-one matching structure always dominates a many-to-many structure [Rosenbaum (1991)].) We denote the one-to-many/many-to-one structure as $1:\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}:1$, where $\kappa_{\mathcal{C}}$ is the maximum number of control units matched to each treated unit, and $\kappa_{\mathcal{T}}$ is the maximum number of treated units matched to each control. These different matching structures are illustrated in Figure 2 below.

It is desirable to extend cardinality matching to matching with a variable one-to-many or a one-to-many/many-to-one structure, but a question that arises is how to define the size of the matched sample with these flexible matching structures. Naturally, five $1:1$ matches of green and nongreen buildings [exemplified in Figure 3(a)] should count more than two $1:2$ matches plus one $1:1$ match [Figure 3(b)], and this, in turn, should count more than one $1:5$ match [Figure 3(c)]. Although the first and second matchings have the same number of different controls, in the second matching there are only two different treated units, and so, subject to the same constraints on covariate balance, the first matching should be

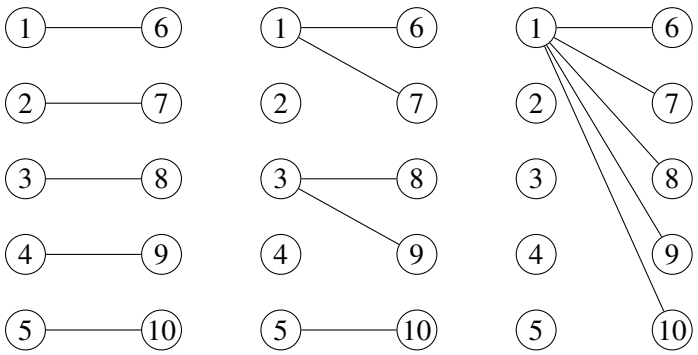


(a) 1 : 1 or pair matching (b) $1:\kappa$ fixed (c) $1:\kappa$ variable (d) $1:\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}:1$ or full matching

FIG. 2. *Different matching structures.*

preferable. Intuitively, there is more information in the first match. In the following section we formalize this notion using the concept of information content of a matched sample for a difference-in-means effect estimator.

2.3. *Information content of a matched sample.* Let $i \in \mathcal{I} = \{1, 2, \ldots, I\}$ index the set of matched groups and $j \in \mathcal{J}_i = \{1, 2, \ldots, J_i\}$ index the set of units (in our study, buildings) within each of these matched groups. Using this notation, for example, in Figure 2(a), $J_i = 2$ for each $i \in \mathcal{I}$ and the matched groups constitute pairs and, in Figure 2(c), $J_1 = 4$ and $J_2 = 3$, and so the groups form quadruples and triples, respectively. To accommodate the more general one-to-many/many-to-one or full matching structure, we adopt the convention that the first unit in each group is either a treated unit and all the other units are controls, or that the first unit is a control and all the other units are treated.



(a) 5 matches with 10 units (b) 5 matches with 8 units (c) 5 matches with 6 units

FIG. 3. *Different matching structures with the same number of matches.*

Following Haviland, Nagin and Rosenbaum (2007), we pose a simple treatment effect model

$$Y_{ij} = \alpha_i + \beta Z_{ij} + \varepsilon_{ij}, \tag{2.1}$$

where Y_{ij} is the observed outcome of unit j in matched group i, α_i is a group effect for all the units in group i (this indicates there is dependence between units in each group, but that it may be eliminated by taking differences within groups), Z_{ij} is the treatment assignment indicator, and ε_{ij} is a residual term with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Here, we use model (2.1) for variance calculations to motivate our matching approach, but we do not explicitly use it for inference, as we want our inference to be valid under more general conditions than those of the model [Tukey (1986)]. Consider the matched group difference in outcomes

$$D_i = Z_{i1}\left(Y_{i1} - \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right) + (1 - Z_{i1})\left(-Y_{i1} + \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right), \tag{2.2}$$

where κ_i is the number of control units in matched group i. We can calculate the variance of this difference and find that

$$\operatorname{Var}(D_i) = \sigma^2\left(1 + \frac{1}{\kappa_i}\right) \propto \left(\frac{2}{\frac{1}{1} + \frac{1}{\kappa_i}}\right)^{-1}. \tag{2.3}$$

In other words, the variance of the difference is inversely proportional to the harmonic mean of the number of treated and control units in each matched group [Kalton (1968); see also Hansen and Bowers (2008)]. We denote $h^{(\kappa)}$ as the harmonic mean of the number of units in a matched group with a $1:\kappa$ (or $\kappa:1$) matching ratio

$$h^{(\kappa)} = \frac{2}{\frac{1}{1} + \frac{1}{\kappa}} = \frac{2\kappa}{1+\kappa}. \tag{2.4}$$

In this manner, in a $1:1$ match or pair match, $h^{(1)} = 1$; in a $1:2$ match, $h^{(2)} = 4/3$; in a $1:3$ match, $h^{(3)} = 3/2$; and so on.

We call the information content of a matched sample the sum of the harmonic means of the number of treated and control units in each matched group, $\sum_{i\in\mathcal{I}} h^{(\kappa_i)}$, that is, the sum of the Fisher information of the matched groups. In this way, for example, the information content of two $1:1$ matches will be 50% larger than the information of one $1:2$ match ($1+1=2$ instead of $4/3$), and the information of three $1:1$ matches will be the same as the information of two $1:3$ matches ($1+1+1 = 3/2+3/2$).

Another way of defining the information content in a matched sample about the parameter β is the reciprocal of the variance of an effect estimator, for example, the average of the group differences

$$\hat{\delta} = \frac{1}{I}\sum_{i\in\mathcal{I}}\left(Z_{i1}\left(Y_{i1} - \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right) + (1 - Z_{i1})\left(-Y_{i1} + \frac{\sum_{j\neq 1} Y_{ij}}{\kappa_i}\right)\right). \tag{2.5}$$

However, other estimators may be preferable in practice, such as regressing the group differences in outcomes on group differences in covariates as in Rubin (1979), or using the weighted M-statistics in Rosenbaum (2014). Also, this definition is less intuitive and more difficult to implement in practice (see Appendix B), and has a weaker connection with cardinality matching. Clearly, if the matching ratio given by κ_i is constant, then maximizing the information content is equivalent to cardinality matching with a fixed $1:\kappa$ ratio as in Zubizarreta, Paredes and Rosenbaum (2014), and so this provides a more general framework and a richer interpretation for cardinality matching.

For these reasons we consider maximizing the sum of the harmonic means of the number of treated and control units in each matched group, in other words, maximizing the sum of the Fisher information of the matched groups. Building upon this notion of information content, in the next section we present a matching framework and specific matching formulations within this framework that maximize the information content of a matched sample subject to covariate balance and matching structure constraints.

3. Maximizing the information of a balanced matched sample.

3.1. *A general matching framework.* Let $t \in \mathcal{T} = \{1, \ldots, T\}$ index the set of treated units (in our study, green buildings) and $c \in \mathcal{C} = \{1, \ldots, C\}$ index the set of controls (nongreen buildings) with $T \leq C$. Define $p \in \mathcal{P} = \{1, \ldots, P\}$ as the label of the P observed covariates. Each treated unit $t \in \mathcal{T}$ has a vector of observed covariates $\boldsymbol{x}_t = \{x_{t,1}, \ldots, x_{t,P}\}$, and each control $c \in \mathcal{C}$ has a similar vector $\boldsymbol{x}_c = \{x_{c,1}, \ldots, x_{c,P}\}$. We introduce the decision variable m_{tc}, which is 1 if treated unit t is matched with control c, and 0 otherwise.

In the abstract, we want to solve

$$\max_{\boldsymbol{m}} \{\mathbb{I}(\boldsymbol{m}) : \boldsymbol{m} \in \mathcal{M} \cap \mathcal{B}\}, \tag{3.1}$$

where $\mathbb{I}(\boldsymbol{m})$ is the information content of the matched sample, and $\mathcal{M}$ and $\mathcal{B}$ are matching and balancing constraints, respectively. This general formulation pursues the goal of finding the largest matched sample—or, in general, the matched sample with the largest information content—that satisfies certain requirements for matching structure $\mathcal{M}$ and covariate balance $\mathcal{B}$. Generally, the requirements for covariate balance are guided by scientific knowledge of the research question at hand. Often one would match with a flexible matching structure, but, as we discuss below, this imposes computational restraints. We now discuss the specific forms of $\mathbb{I}$, $\mathcal{M}$ and $\mathcal{B}$ when matching with a $1:\kappa$ fixed ratio, a $1:\kappa_{\mathcal{C}}$ variable ratio, and, due to space considerations, we relegate the case of matching with a flexible $1:\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}:1$ matching ratio to Appendix C.

3.2. *Matching with a fixed* $1:\kappa$ *ratio.* Matching with a fixed $1:\kappa$ ratio is equivalent to cardinality matching. In (3.1), $\mathbb{I}$, $\mathcal{M}$ and $\mathcal{B}$ take the forms

$$\mathbb{I}(\boldsymbol{m}) = \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}, \tag{3.2}$$

$$\begin{aligned}\mathcal{M} = \Bigg\{ & \sum_{c\in\mathcal{C}} m_{tc} = \kappa, t\in\mathcal{T} \text{ if } \kappa > 1 \text{ and } \sum_{c\in\mathcal{C}} m_{tc} \le \kappa, t\in\mathcal{T} \text{ if } \kappa = 1; \\ & \sum_{t\in\mathcal{T}} m_{tc} \le 1, c\in\mathcal{C}; m_{tc}\in\{0,1\}, t\in\mathcal{T}, c\in\mathcal{C}\Bigg\},\end{aligned} \tag{3.3}$$

$$\begin{aligned}\mathcal{B} = \Bigg\{ & -\varepsilon_p \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc} \le \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}\big(f(x_{t,p}) - f(x_{c,p})\big) \\ & \le \varepsilon_p \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}, \\ & m_{tc}\in\{0,1\}, t\in\mathcal{T}, c\in\mathcal{C}; p\in\mathcal{P}\Bigg\},\end{aligned} \tag{3.4}$$

where $\varepsilon_p \ge 0$ is a given constant, and $f(\cdot)$ is a suitable transformation of the covariates. For example, if $f(x_{\cdot,p}) = x_{\cdot,p}$, then (3.4) constrains the matched samples to have means that differ at most by ε_p for covariate p. Also, if $f(\cdot)$ is a binary indicator for the categories of a nominal covariate p and $\varepsilon_p = 0$, then (3.4) requires the matched samples to have the same number of treated and control units within each category, but without constraining which units are matched together [Rosenbaum, Ross and Silber (2007)]. Similar ideas can be used to balance the interactions of several nominal covariates. See Zubizarreta (2012) and Zubizarreta, Paredes and Rosenbaum (2014) for more balancing examples.

3.3. *Matching with a variable* $1:\kappa_{\mathcal{C}}$ *ratio.* To generalize cardinality matching for maximizing the information content of the matched sample with a variable $1:\kappa_{\mathcal{C}}$ matching ratio, we introduce a new decision variable n_t, the number of control units that treated unit t is matched to, which is bounded above by $\kappa_{\mathcal{C}}$. Then problem (3.1) becomes

$$\mathbb{I}(\boldsymbol{m},\boldsymbol{n}) = \sum_{t\in\mathcal{T}} h^{(n_t)}, \tag{3.5}$$

$$\begin{aligned}\mathcal{M} = \Bigg\{ & \sum_{c\in\mathcal{C}} m_{tc} = n_t, t\in\mathcal{T}; n_t \le \kappa_{\mathcal{C}}, t\in\mathcal{T}; \sum_{t\in\mathcal{T}} m_{tc} \le 1, c\in\mathcal{C}; \\ & m_{tc}\in\{0,1\}, t\in\mathcal{T}, c\in\mathcal{C}; n_t \ge 0, t\in\mathcal{T}\Bigg\},\end{aligned} \tag{3.6}$$

$$\mathcal{B} = \Bigg\{ -\varepsilon_p \sum_{t\in\mathcal{T}} h^{(n_t)} \le \sum_{t\in\mathcal{T}} h^{(n_t)} x_{t,p}$$

$$
\begin{aligned}
(3.7)\qquad & -\sum_{c\in\mathcal{C}}\Bigg(\sum_{t\in\mathcal{T}} m_{tc}\frac{h^{(n_t)}}{n_t}\Bigg)x_{c,p} \le \varepsilon_p \sum_{t\in\mathcal{T}} h^{(n_t)}, \\
& \qquad p\in\mathcal{P}, m_{tc}\in\{0,1\}, t\in\mathcal{T}, c\in\mathcal{C}; n_t\ge 0, t\in\mathcal{T}\Bigg\}.
\end{aligned}
$$

Here, we let $f(x)=x$ for mean balance. Note that, by using transformations of the covariates, it is possible to balance other statistics besides means [e.g., by mean balancing indicators for the quantiles of x in the treated units, it is possible to approximately balance its marginal distribution; see Zubizarreta (2012) for details]. Also, note that $h^{(\kappa)}$ is an increasing, convex transformation of κ; that is, $h^{(\kappa)}$ increases as κ increases at a decreasing rate. However, this optimization problem has the expressions $h^{(n_t)}$ and $m_{tc}\frac{h^{(n_t)}}{n_t}$ which are not linear in m_{tc} and n_t. To linearize $h^{(n_t)}$, we define a new decision variable $m_t^{(r)}$, which is 1 if treated unit t is matched with at least r controls, and 0 otherwise ($t\in\mathcal{T}, r\in\{1,\dots,\kappa_{\mathcal{C}-1}\}$). This new decision variable can be written using linear constraints as

$$
(3.8)\qquad m_t^{(r)} \le n_t - \sum_{s=1}^{r-1} m_t^{(s)}, \qquad t\in\mathcal{T}, r\in\{1,\dots,\kappa_{\mathcal{C}-1}\},
$$

$$
(3.9)\qquad \kappa_{\mathcal{C}} m_t^{(r)} \ge n_t - \sum_{s=1}^{r-1} m_t^{(s)}, \qquad t\in\mathcal{T}, r\in\{1,\dots,\kappa_{\mathcal{C}-1}\}.
$$

Here we do not need to define the decision variable $m_t^{(\kappa_{\mathcal{C}})}$ since $m_t^{(\kappa_{\mathcal{C}})} = n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}$. Using the variables $m_t^{(r)}$, we can rewrite $h^{(n_t)}$ as

$$
\begin{aligned}
(3.10)\qquad w_t^{(1)} &:= h^{(n_t)} \\
&= \sum_{s=1}^{\kappa_{\mathcal{C}}-1} \big(h^{(s)} - h^{(s-1)}\big) m_t^{(s)} + \big(h^{(\kappa_{\mathcal{C}})} - h^{(\kappa_{\mathcal{C}}-1)}\big)\Bigg(n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}\Bigg).
\end{aligned}
$$

Hence, we can write the objective function in the linear form $\sum_{t\in\mathcal{T}} w_t^{(1)}$.

The next step is to write $m_{tc}\frac{h^{(n_t)}}{n_t}$ in linear form. For this, define

$$
\begin{aligned}
(3.11)\qquad w_t^{(2)} &:= \frac{h^{(n_t)}}{n_t} \\
&= \sum_{s=1}^{\kappa_{\mathcal{C}}-1}\Bigg(\frac{h^{(s)}}{s} - \frac{h^{(s-1)}}{s-1}\Bigg) m_t^{(s)} + \Bigg(\frac{h^{(\kappa_{\mathcal{C}})}}{\kappa_{\mathcal{C}}} - \frac{h^{(\kappa_{\mathcal{C}}-1)}}{\kappa_{\mathcal{C}}-1}\Bigg)\Bigg(n_t - \sum_{s=1}^{\kappa_{\mathcal{C}}-1} m_t^{(s)}\Bigg),
\end{aligned}
$$

where $\frac{h^{(0)}}{0}$ is set to 0. The expression of interest becomes $m_{tc}w_t^{(2)}$, which is still not linear. To linearize it, we define the decision variable $q_{tc} = m_{tc}w_t^{(2)}$, which is

equal to $w_t^{(2)}$ if $m_{tc} = 1$, 0 otherwise, and write

$$q_{tc} \leq m_{tc}, \qquad t \in \mathcal{T}, c \in \mathcal{C}, \tag{3.12}$$

$$q_{tc} \leq w_t^{(2)}, \qquad t \in \mathcal{T}, c \in \mathcal{C}, \tag{3.13}$$

$$q_{tc} \geq w_t^{(2)} - (1 - m_{tc}), \qquad t \in \mathcal{T}, c \in \mathcal{C}. \tag{3.14}$$

Last, we define $w_c = \sum_{t \in \mathcal{T}} q_{tc}, c \in \mathcal{C}$, and rewrite the mean balancing constraints as

$$-\varepsilon_p \sum_{t \in \mathcal{T}} w_t^{(1)} \leq \sum_{t \in \mathcal{T}} w_t^{(1)} x_{t,p} - \sum_{c \in \mathcal{C}} w_c x_{c,p} \leq \varepsilon_p \sum_{t \in \mathcal{T}} w_t^{(1)}, \qquad p \in \mathcal{P}. \tag{3.15}$$

This program is no longer a pure integer programming (IP) problem, as in cardinality matching; it is a mixed integer programming (MIP) problem with considerably less structure than the MIP problem solved by Zubizarreta (2012). In fact, the constraints (3.8)–(3.15) make the program quite complicated to solve in general.

3.4. *Matching with a flexible* $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ *ratio*. One step further is to formulate (3.1) to match with a flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ matching ratio or full matching. Due to space constraints, this is discussed in Appendix C.

4. Description of the matches. In our study, we find the matched sample of green and nongreen buildings with largest information (3.5) that satisfies the matching structure (3.6) and that balances the original covariates according to (3.7). In particular, we match with a variable $1 : \kappa_{\mathcal{C}}$ matching ratio because each geographic cluster has only one green building and a variable number of nongreen buildings. We choose $\kappa_{\mathcal{C}} = 4$ because the gains from matching with a higher $1 : 5$ or a $1 : 6$ ratio are not very marked assuming the same number of treated units are matched [see Table 2 of Haviland, Nagin and Rosenbaum (2007)] and because increasing the maximum matching ratio by one adds $2T$ constraints and T binary variables to the mathematical program, making it more difficult to solve (see Section 4.4 below). In accordance with Eichholtz, Kok and Quigley (2010), we match green and nongreen buildings within the same geographic clusters that define the markets. For this, we use constraint (E.1) in Appendix E, although this constraint is not necessary for solving (3.5)–(3.7) in general.

4.1. *Covariate balance*. Table 1 shows the absolute standardized differences in means of the observed covariates before and after matching with a variable $1 : 4$ ratio. In the table, before matching there are a number of substantial differences, most notably in the building classes, age ($>$40 years) and amenities, whereas after matching all these differences are smaller than 0.1. Within the framework of (3.1), we designed the matched sample to be balanced in this way.

TABLE 1
Standardized differences in means before and after matching

	Standardized difference in means	
Covariate	**Before matching**	**After matching**
Building size	0.362	0.076
Building class A	1.005	0.096
Building class B	−0.650	0.053
Building class C	−0.557	−0.068
Net contract	0.127	0.020
Employment growth	0.043	0.000
Employment growth missing	−0.010	0.000
Age $\leq$10 years	0.323	0.049
Age 11–20 years	0.400	0.034
Age 21–30 years	0.392	0.018
Age 31–40 years	−0.066	−0.044
Age $>$40 years	−0.974	−0.050
Age missing	−0.150	−0.007
Renovated	−0.389	0.033
Stories low	−0.145	−0.066
Stories intermediate	0.032	0.046
Stories high	0.141	0.031
Stories missing	−0.061	−0.014
Amenities	0.474	0.079

4.2. *Information of the matched samples*. Table 2 below shows the information content or, loosely speaking, the effective samples sizes of the samples matched with fixed 1 : 1, 1 : 2, 1 : 3 and 1 : 4 ratios, and with a variable 1 : 4 ratio. With a 1 : 1 ratio or pair matching, the resulting information content is 666, meaning that 666 buildings were paired. With fixed 1 : 2, 1 : 3 and 1 : 4 ratios, the information content is equivalent to 757, 708 and 642 pairs, respectively, whereas with a variable 1 : 4 ratio, it is 941. In other words, matching with a variable 1 : 4

TABLE 2
Effective sample sizes as measured by $\mathbb{I}$ *in* (3.5)

Matching structure	**Information or effective sample size**
1 : 1 fixed	666
1 : 2 fixed	757.3
1 : 3 fixed	708
1 : 4 fixed	641.6
1 : 4 variable	940.6

ratio produces an effective sample size 47% larger than matching with a fixed $1:4$ ratio. This shows the gains from matching with a variable ratio.

4.3. *Comparison to optimal matching*. Following the suggestion of a reviewer, we compare our method to optimal matching as implemented in `optmatch` [Hansen (2007)]. In optimal matching, we calculate the Mahalanobis distance with propensity score calipers as suggested in Rosenbaum and Rubin (1985). For a strict comparison, in both methods we use a variable $1:4$ matching ratio. As a result, with optimal matching the effective sample size is smaller than with our method (730 versus 940.6) and there are imbalances in several covariates (more than half of the covariates exhibit differences in means larger than 0.1 standard deviations). Arguably, covariate balance could be improved by recalculating the covariate distances, but this would involve iteration in order to achieve covariate balance [as described in Figure 1(a) above]. With the proposed method, the differences in means are constrained to be at most 0.1 standard deviations by design. However, `optmatch` is optimal in another important sense—it minimizes the total sum of covariate distances between matched units—and it runs in polynomial time, so relatively large data sets can be handled quickly [Hansen and Klopfer (2006)]. As we discuss in the following section, computation is an important aspect to consider in the implementation of our method.

4.4. *Computation and details of the implementation*. Matching with a variable $1:\kappa_C$ ratio, (3.5)–(3.15), as in our study, and also matching with a flexible $1:\kappa_C/\kappa_T:1$ ratio, (C.3)–(C.29), as in Appendix C, have more complicated structure than cardinality matching, mainly due to the harmonic means used in the objective function and mean balancing constraints. Specifically, while cardinality matching with a $1:1$ ratio and mean balancing has $T \times C$ binary decision variables and $T + C + 2 \times P$ constraints, matching with a variable $1:\kappa_C$ ratio with harmonic means has additional $T \times (\kappa_C + C)$ continuous decision variables and $T \times (2 \times \kappa_C + 3 \times C - 1)$ constraints after some simplifications.

Although these two matching problems are considerably larger than cardinality matching, by using optimization solvers such as CPLEX and Gurobi, it is still possible to reach solutions with a small optimality gap in a reasonable amount of time depending on the problem size (see Appendix D for a simulation study using the buildings data). Nemhauser (2013) reports that algorithmic speed in solvers such as CPLEX and Gurobi has increased 256,000 times between 1991 and 2013. This, combined with a modest computer speedup of 1000 times, translates into the ability to solve problems that took nearly seven years in the early 1990s to one second today [Nemhauser (2013)]. These major improvements have been made possible by a combination of advancements in preprocessing and heuristics for finding good feasible solutions quickly, branch-and-bound methods to reduce the feasible set, linear programming implementations as the basic tool for solving IP and MIP problems, and parallel computing [Bixby and Rothberg (2007), Linderoth

and Lodi (2010), Nemhauser (2013); see also Bertsimas (2014) for a related discussion and applications of MIP to statistical and machine learning].

In addition to these optimization techniques, we used exact matching constraints on the location covariate (see Appendix E), and divided the problem into 10 subproblems to solve each of them in parallel. Using the `R` packages `doParallel` and `foreach` [Weston and Calaway (2014)], we solved the 10 subproblems independently and simultaneously using 10 processors with a 15-minute time limit. Among these subproblems, one gives the optimal solution within the time limit, and the others give solutions with about a 2% optimality gap at the end of the specified time. This computational implementation method enables us to solve this problem under 20 minutes. It would take more than 2 hours to reach the same solution if no parallel computing methods were used. At the present time, the code that we used for the analyses is available upon request, but soon it will be available within the package `designmatch` for `R`.

5. Economic performance of green buildings. From our balanced matched sample, we find that green buildings have 3.3% higher rental rates per square foot than otherwise similar nongreen buildings, with a 95% confidence interval of [1.3%, 5.5%].

To obtain these estimates, we used the weighted test statistic in Hansen, Rosenbaum and Small [(2014); Sections 2.3 and 2.4], and, under an additive constant treatment effect model, solved the Hodges–Lehmann estimating equation to derive the point estimate, and then inverted the test to obtain the confidence interval [see Chapter 2 of Rosenbaum (2002) and Chapter 3 of Lehmann (2006) for details]. Following Eichholtz, Kok and Quigley (2010), we used log rents as the outcome variable and interpreted differences within matched groups as a percentage change or rate. For comparison, the 3.3% estimate is moderately larger than the one of Eichholtz, Kok and Quigley (2010), who reported that green buildings have rental rates 2.8% higher per square foot than similar nongreen buildings (with a 95% confidence interval of [1%, 4.6%]). However, our estimand is not strictly comparable to the one of Eichholtz, Kok and Quigley (2010) since they use linear regression, in principle weighting all the observations [Aronow and Samii (2016)], whereas we use matching, restricting the analysis to the sample with the largest information that is balanced (in our study, these are 675 out of the 694 green buildings available before matching).

In order to get a better understanding of the representativeness of our matched sample, in Table 4 of Appendix F we provide a description of the samples of green buildings before matching, after matching, and of those green buildings that were unmatched and left out from the analyses. Overall, this sample closely resembles that of all the available green buildings before matching, and so, in principle, these results can be generalized to a population of buildings of similar characteristics.

Next, when conducting a sensitivity analysis to hidden biases, we find that, for an unobserved covariate to explain away the estimated effect of 3.3%, it would

need to simultaneously increase the odds of a building having green ratings and of a positive difference in rent both by a factor of 1.9, and so the results are only moderately insensitive to hidden biases [see Rosenbaum and Silber (2009) and Hansen, Rosenbaum and Small (2014) for details of this analysis].

To interpret these results, let us remember that approximately 30% of building operating costs are driven by energy consumption and that green buildings typically have 25% less energy use and 19% lower operating costs. Therefore, in broad terms, savings from operating costs overcome the extra amount paid for a green building rent if the rent to operating costs ratio is 5.75 ($= 0.19/0.033$) or more. Thus, it is an economically sound decision for some companies to prefer green buildings and pay more rent. Moreover, as Eichholtz, Kok and Quigley (2010) discuss, a small improvement on the energy use of existing buildings can also have a big impact on the environment. In this way, companies are also willing to pay more to "go green" for a sustainable environment.

6. Discussion. The main objective of matching in observational studies is to balance observed covariates and thereby remove biases due to systematic differences in their distributions [Cochran (1965), Section 2.2]. As discussed in Section 8.7 of Rosenbaum (2010), efficiency is a secondary concern in observational studies. The explanation is that if there is a bias that does not decrease as the sample size increases, then it tends to dominate the mean squared error in large samples, resulting in a very precise estimate of the wrong quantity [Haviland, Nagin and Rosenbaum (2007)]. For these reasons, in view of the bias-variance—or, stated differently, the balance-precision—trade-off involved in matching, we give priority to balance over precision, and, subject to removing systematic biases by balancing covariates, we maximize precision or, more specifically, the information content of the matched sample.

The framework we proposed in Section 3.1 encompasses these objectives in a general way. Within this framework, cardinality matching is a special case when matching with a fixed $1:\kappa$ ratio. Also, the formulations presented in Section 3.3, and in Appendices B and C, are different methods for maximizing the information content of a balanced matched sample. Ideally, if the outcome model follows (2.1) and if the outcome analyses use the effect estimator (2.5), then one would solve the matching problem in Appendix B, but, as discussed, this is a very complicated optimization problem because the number of matched pairs I is also a decision variable. If the solution to the cardinality matching problem uses all the available treated units, then this solution also minimizes the variance of the effect estimator (2.5). With other estimators or nonconstant variances across units, the formulations in Section 3.3 and Appendix C may be more appropriate.[7] As discussed in Section 2.3, these formulations are not only easier to implement but also

[7]In model (2.1), we assumed that the variance is constant across units. One way to relax this assumption is to assume instead that the variance in the treated group is f times larger than the

more intuitive, as they maximize the sum of the Fisher information of the matched groups.

Building on cardinality matching, the proposed methods do not require estimation of the propensity score as they directly balance the original covariates. Nonetheless, the propensity score may be used as an additional covariate in the balancing constraints $\mathcal{B}$. In this paper we mainly discussed mean balancing constraints, but other constraints can be implemented for distributional balance [Zubizarreta (2012)].

Assessing common support or overlap in covariate distributions is a common practice in observational studies to avoid extrapolating or fabricating results from regression models that assume a particular functional form [Rosenbaum (2010), Section 18.2; Imbens and Rubin (2015), Chapter 14]. This is typically done in two steps: first, by trimming the sample on the propensity score and, second, by checking balance. For instance, Imbens (2015) suggests dropping units with extreme values of the estimated propensity score and then checking balance in normalized differences in average covariates. As in cardinality matching, the methods proposed in this paper directly "trim" the sample to satisfy the requirements for covariate balance of the original covariates. To the extent that these requirements balance the covariates adequately, these methods will avoid extrapolation by restricting the analysis to the matched treated and control samples that overlap the most (again, in the sense of information and the balance requirements).

Of course, restricting the analysis to the samples of treated and control units that overlap will typically change the estimand. In the case that treated units are matched to a subset of the controls, the estimand will cease to be the effect of treatment on the treated and it will become a more local estimand that depends on the sample data [Crump et al. (2009)]. In view of this limitation imposed by the data, one way to proceed without further modeling assumptions is by describing both the matched and unmatched samples as in Appendix F. This provides a basic understanding of the population to which, in principle, the results of the matched analysis can be generalized [Hill (2008); see also Traskin and Small (2011), Fogarty et al. (2016) and Silber et al. (2015)]. Another way to proceed is by weighting the matched samples to a target population of greater policy interest, for example, by using methods in Hartman et al. (2015).

In cardinality matching, finding the largest balanced matched sample is followed by rematching the pairs or groups that constitute the matched sample to

---

one in the control group. Then $h^{(\kappa)}$ becomes the harmonic mean of the sum of 1 and κ_i/f units for each matched group. As another example, suppose that the variance in one category of a binary covariate is f times larger than the one in the other category. Then the weighting becomes $h^{(\kappa_i)}/f$ for the matched group with greater variance, therefore requiring to match f times as many groups from the strata with smaller variance. Extending this example, there may be important strata, and one could estimate the variance in those strata and plug in the estimates, but this would require using the outcomes for matching. In general, if the variances vary arbitrarily, then the weights become intractable.

minimize their total sum of covariate distances. If these covariates are predictive of the outcome, this rematching will reduce heterogeneity within matched groups and therefore sensitivity to biases due to unobserved covariates [Rosenbaum (2005)]. A possible direction for future research would be to extend the proposed methods along these lines. Also, the proposed methods can be used for adjustment in observational studies with a time-dependent treatment and time-dependent covariates via risk set matching [Li, Propert and Rosenbaum (2001), Lu (2005)]. Under weaker identification assumptions than "no unmeasured confounders," the proposed methods can also be used for treatment effect estimation with an instrumental variable [Baiocchi et al. (2010), Zubizarreta et al. (2013)] or a discontinuity design [Keele, Titiunik and Zubizarreta (2015)].

7. Summary. In this paper we revisited the study of Eichholtz, Kok and Quigley (2010) about the market performance of green buildings. To analyze the effect of energy efficiency and sustainability on the economic returns of buildings, we used new matching methods that take more advantage of the clustered structure of the buildings data than standard matching methods. We proposed a general framework for matching in observational studies and specific matching methods within this framework that simultaneously achieve three goals: (i) maximize the information content of a matched sample (and, in some cases, also minimize the variance of a widely used effect estimator); (ii) form the matches using a flexible matching structure (such as a one-to-many/many-to-one structure); and (iii) directly attain covariate balance as specified—before matching—by the investigator. To our knowledge, existing matching methods are only able to achieve, at most, two of these goals simultaneously. Using these methods, we obtained a larger effective sample size and found that green buildings have 3.3% higher rental rates per square foot than otherwise similar buildings without green ratings [a moderately larger effect than the one previously found by Eichholtz, Kok and Quigley (2010)]. Thus, besides being environmentally responsible, it is also an economically sound decision to pursuel environmentally sustainable building practices.

Acknowledgments. We thank Jake Bowers, Luke Miratrix and Paul Rosenbaum for comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings" (DOI: 10.1214/16-AOAS962SUPP; .pdf). In this on-line supplement, we include the appendices to "Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings" by Kilcioglu and Zubizarreta (2016).

REFERENCES

ARONOW, P. M. and SAMII, C. (2016). Does regression produce representative estimates of causal effects? *Amer. J. Polit. Sci.* **60** 250–267.

BAIOCCHI, M. (2011). Designing robust studies using propensity score and prognostic score matching. Chapter 3 in Methodologies for Observational Studies of Health Care Policy, Dissertation, Department of Statistics, The Wharton School, Univ. Pennsylvania, Philadelphia, PA.

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550

BERTSIMAS, D. (2014). Statistics and Machine Learning via a Modern Optimization Lens. The 2014–2015 Philip McCord Morse Lecture.

BIXBY, R. and ROTHBERG, E. (2007). Progress in computational mixed integer programming—A look back from the other side of the tipping point. *Ann. Oper. Res.* **149** 37–41. MR2313358

CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700.

COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **128** 234–266.

COCHRAN, W. and RUBIN, D. (1973). Controlling bias in observational studies: A review. *Sankhya* **35** 417–446.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DIAMOND, A. and SEKHON, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* **95** 932–945.

EICHHOLTZ, P., KOK, N. and QUIGLEY, J. M. (2010). Doing well by doing good? Green office buildings. *Am. Econ. Rev.* **100** 2492–2509.

FOGARTY, C., MIKKELSEN, M., GAIESKI, D. and SMALL, D. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Amer. Statist. Assoc.* **111** 447–458.

GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition model with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387

HANSEN, B. B. (2007). Flexible, optimal matching for observational studies. *R News* **7** 18–24.

HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.* **23** 219–236. MR2516821

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151

HANSEN, B. B., ROSENBAUM, P. R. and SMALL, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *J. Amer. Statist. Assoc.* **109** 133–144. MR3180552

HARTMAN, E., GRIEVE, R., RAMSAHAI, R. and SEKHON, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *J. Roy. Statist. Soc. Ser. A* **178** 757–778. MR3348358

HAVILAND, A., NAGIN, D. and ROSENBAUM, P. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychol. Methods* **12** 247.

HILL, J. (2008). Discussion of research using propensity-score matching: Comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, *Statistics in Medicine* [MR2439882]. *Stat. Med.* **27** 2055–2061. MR2439884

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552

IACUS, S. M., KING, G. K. and PORRO, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* **20** 1–24.

IMAI, K. and RATKOVIC, M. (2015). Robust estimation of inverse probability weights of marginal structural models. *J. Amer. Statist. Assoc.* **110** 1013–1023. MR3420680

IMBENS, G. W. (2015). Matching methods in practice: Three examples. *J. Hum. Resour.* **50** 373–419.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

KALTON, G. (1968). Standardization: A technique to control for extraneous variables. *Appl. Statist.* **17** 118–136. MR0234599

KEELE, L., TITIUNIK, R. and ZUBIZARRETA, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. Roy. Statist. Soc. Ser. A* **178** 223–239. MR3291769

KILCIOGLU, C. and ZUBIZARRETA, J. R. (2016). Supplement to "Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings." DOI:10.1214/16-AOAS962SUPP.

LEHMANN, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*, 1st ed. Springer, New York. MR2279708

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2016). Balancing covariates via propensity score weighting. Working paper.

LI, Y. P., PROPERT, K. J. and ROSENBAUM, P. R. (2001). Balanced risk set matching. *J. Amer. Statist. Assoc.* **96** 870–882. MR1946360

LINDEROTH, J. T. and LODI, A. (2010). MILP software. In *Wiley Encyclopedia of Operations Research and Management Science* (J. J. Cochran, L. A. Cox, P. Keskinocak and J. P. Kharoufeh, eds.). Wiley, New York.

LU, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics* **61** 721–728. MR2196160

NEMHAUSER, G. L. (2013). Integer programming: Global impact. EURO INFORMS, July 2013.

NIKOLAEV, A. G., JACOBSON, S. H., CHO, W. K. T., SAUPPE, J. J. and SEWELL, E. C. (2013). Balance optimization subset selection (BOSS): An alternative approach for causal inference with observational data. *Oper. Res.* **61** 398–412. MR3046118

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244

ROSENBAUM, P. R. (1987). Model-based direct adjustment. *J. Amer. Statist. Assoc.* **82** 387–394.

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. (1991). Discussing hidden bias in observational studies. *Arch. Intern. Med.* **115** 901–905.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. MR3265687

ROSENBAUM, P. R. (2015). How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application* **2** 21–48.

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSENBAUM, P. R. and SILBER, J. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics* **2** 217–232.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–840. MR2516795

SILBER, J. H., ROSENBAUM, P. R., KELZ, R. R., GASKIN, D. J., LUDWIG, J. M., ROSS, R. N., NIKNAM, B. A., HILL, A., WANG, M., EVEN-SHOSHAN, O. and FLEISHER, L. A. (2015). Examining causes of racial disparities in general surgical mortality: Hospital quality versus patient risk. *Med. Care* **53** 619–629.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

TRASKIN, M. and SMALL, D. (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach. *Statistics in Biosciences* **3** 94–118.

TUKEY, J. W. (1986). Sunset salvo. *Amer. Statist.* **40** 72–76.

WESTON, S. and CALAWAY, R. (2014). Getting Started with `doParallel` and `foreach`.

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. MR2959630

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672

ZUBIZARRETA, J. R. and KILCIOGLU, C. (2016). *designmatch*: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design R package version 0.2.0.

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* **8** 204–231. MR3191988

ZUBIZARRETA, J. R., REINKE, C. E., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2011). Matching for several sparse nominal variables in a case–control study of readmission following surgery. *Amer. Statist.* **65** 229–238. MR2867507

ZUBIZARRETA, J. R., SMALL, D. S., GOYAL, N. K., LORCH, S. and ROSENBAUM, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* **7** 25–50. MR3086409

DIVISION OF DECISION, RISK, AND OPERATIONS
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
USA
E-MAIL: ckilcioglu16@gsb.columbia.edu
zubizarreta@columbia.edu
URL: http://www.columbia.edu/~ck2560/
http://www.columbia.edu/~jz2313/