MAXIMIZING THE INFORMATION CONTENT OF A BALANCED MATCHED SAMPLE IN A STUDY OF THE ECONOMIC PERFORMANCE OF GREEN BUILDINGS*

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 the study of Eichholtz et al. (2010) 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 previously found by Eichholtz et al. (2010).

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.[1] For the same year, the U.S. Energy Information Administration reported that about 40% of total U.S. energy consumption was from these types of buildings.[2] 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.[3] 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 et al. (2010) analyzed a large sample of commercial green- and non-green-rated buildings in the United States. Using linear regression and propensity score methods, they found that buildings with green ratings have 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 en-

---

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

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

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

ergy used meets certain criteria.[4] 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 et al. (2010), we consider a building to be green if it is certified as Energy Star or LEED and focus our analysis on commercial buildings.

To estimate the effect of energy efficiency and sustainability on the economic returns of buildings, we compare green-rated buildings to similar non-green-rated buildings in the same market. For this, we use multivariate matching methods and find matches of green and non-green 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 7,411 non-green buildings, organized in 694 geographic clusters. In each of these clusters, there is one green building and one or more non-green buildings not further apart than one quarter mile from the green building. While some clusters have only one non-green building, others have as many as 83 non-green buildings. As a result of this structure, pair matching (or matching with a 1 : 1 ratio) would result in many non-green buildings not being used in the analysis, and matching with a fixed 1 : κ ratio (where κ is an integer greater than 1) would result in some clusters not being used at all. Naturally, for our analyses 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, to 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 of causal effects, matching methods are often used in an attempt to compare like with like; i.e., 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

---

[4]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, number of occupants.

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 [for instance, see 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 form of adjustments that better suits 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 curse of dimensionality in exact matching: as the number of observed covariates increases, there is a combinatorial explosion in the resulting types of units. In fact, with two binary covariates there are 2^2 or four types of units, but with twenty binary covariates there are 2^{20} or over a million 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) for a detailed exposition of this argument]. 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 (this 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 Zubizarreta et al. (2011) and Yang et al. (2012) for related discussions]. 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 the marginal distributions (Rosenbaum, Ross, and Silber, 2007), and this can be very difficult by matching on the propensity score [for a related argument in the context of weighting see, for instance, Zubizarreta (2015)]. It is for these reasons that matching on the propensity score involves a considerable amount of guesswork in practice.

A recent method that addresses these limitations is optimal cardinality matching, or cardinality matching for short (Zubizarreta, Paredes, and Rosenbaum, 2014). Cardinality matching solves an integer programming problem to maximize the cardinality or size of a matched sample subject to constraints on covariate balance. These constraints allow the investigator to balance the covariates directly and in a very precise manner. 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 et al., 2007) and strength-k matching (Hsu et al., 2015).[5] In this way cardinality matching directly balances covariates.

Other interesting matching methods that aim at covariate balance include coarsened exact matching (Iacus, King, and Porro, 2012), balance optimization subset selection (Nikolaev, Jacobson, Cho et al., 2013), genetic

---

[5]Fine balance forces the marginal distributions of a nominal variable to be identical, but without constraining units to be matched within each of the categories of a nominal variable (see Chapter 10 of Rosenbaum, 2010 for details); whereas strength-k matching is a stronger form of balance in which low dimensional joints are forced to be identical: out of K nominal covariates, each of the $\binom{K}{k}$ possible interactions of covariates is finely balanced, so the joint distributions of each of the $\binom{K}{k}$ combinations of covariates is perfectly balanced.

matching (Diamond and Sekhon, 2013), and refined covariate balance via network flows (Pimentel, Kelz, Silber et al., 2015). Other related weighting methods include inverse probability tilting (Graham, de Xavier Pinto, and Egel, 2012), entropy balancing (Hainmueller, 2012), stable balancing weights (Zubizarreta, 2015), calibration weighting (Chan, Yam, and Zhang, 2016), and the overlap weights (Li, Morgan, and Zaslavsky, 2016).
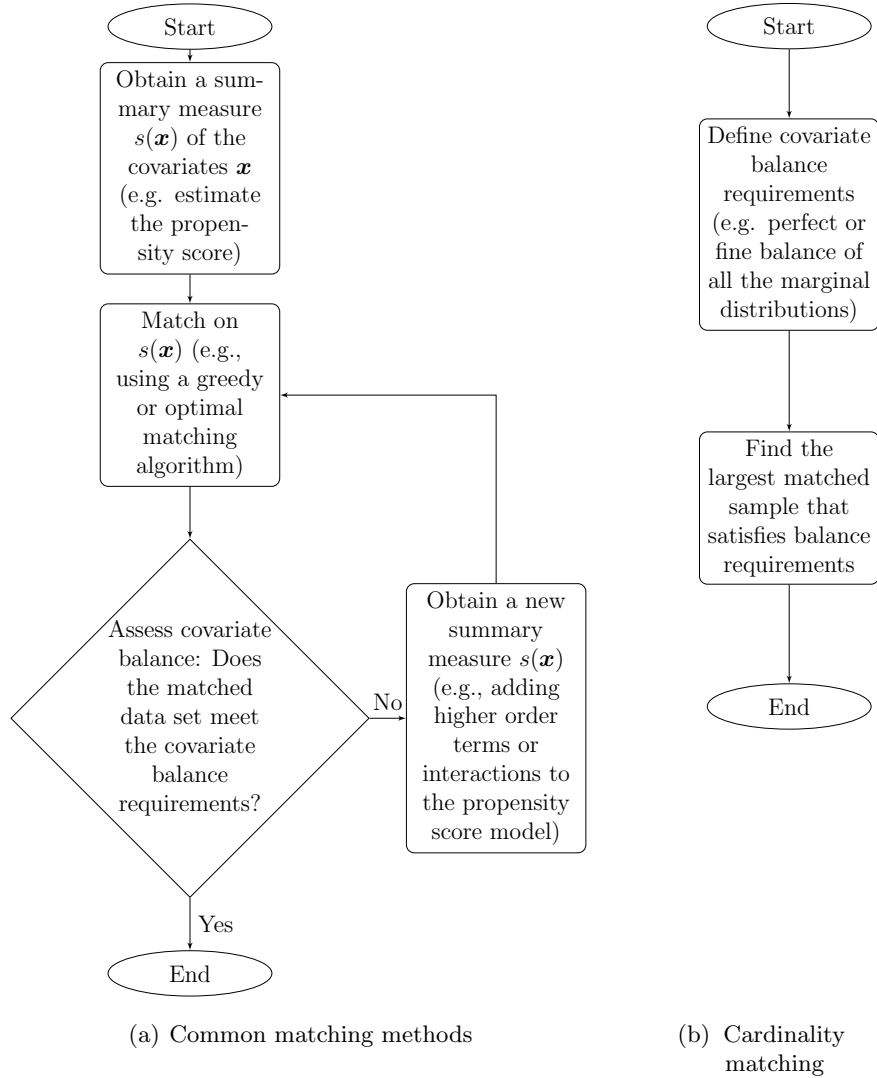

(a) Common matching methods   (b) Cardinality matching

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

The flowcharts in Figure 1 compare the basic steps involved in cardinality matching and in standard matching methods based on the propensity score or other summary measures of the observed covariates (such as the Mahalanobis distance). While 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. For simplicity, in Figure 1(a) we omit the decisions involved in propensity score matching about overlap, but typically additional steps would be present [for example, see Chapter 15 of Imbens and Rubin (2015) for an extensive discussion]. In contrast, with cardinality matching the possibility of covariate 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 investigator. 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 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 new matching methods proposed. In Section 7 we close with a summary and remarks.

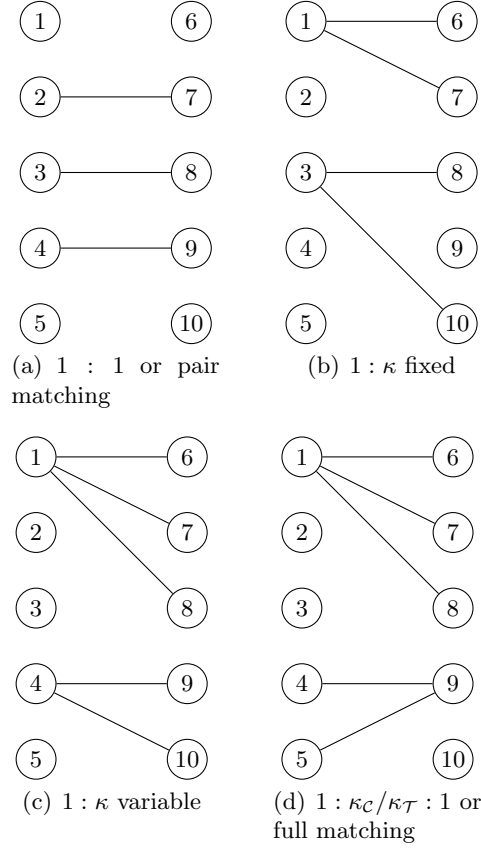## 2. Review: cardinality matching; matching structures; information content.

2.1. *Cardinality matching.*   As described above, most matching methods target covariate balance indirectly, by matching treated and control units (green and non-green buildings) that are close on a summary measure of the covariates such as the propensity score. Unlike these matching methods, cardinality matching uses the original covariates to match units and directly balance their covariate distributions (Zubizarreta et al., 2014). Specifically, cardinality matching finds the *largest* matched sample that satisfies the investigator's specifications for covariate balance. Following Zubizarreta (2012), these specifications for covariate balance may not only require mean balance, but perhaps also other forms of distributional balance such as fine balance (Rosenbaum et al., 2007), x-fine balance (Zubizarreta et al., 2011), strength-k matching (Hsu et al., 2015), and exact matching [Rosenbaum (2010), Section 9.3], all this on several covariates simultaneously. 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 focuses on covariate balance in aggregate, allowing the investigator to re-match the treated and control units in the balanced

matched sample to emphasize covariates that are strongly correlated with the outcome. As illustrated in Zubizarreta et al. (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 a polynomial time algorithm to solve 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). At the present, cardinality matching is solved with the optimization solvers CPLEX, GLPK, Gurobi and Symphony via the statistical package `designmatch` for R [Zubizarreta (2012), Zubizarreta and Kilcioglu (2015)].

2.2. *Matching structures.* In its simplest form, a matched sample is assembled by pairs of treated and control units selected from larger reservoirs of both types of units. 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 (Rosenbaum, 1989; Hansen, 2004). [In rigor, the term full match refers not only to a one-to-many/many-to-one structure but 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). Also, by matching without replacement it is straightforward to conduct inference with existing methods ((Rosenbaum, 1993, 2001).] 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

(a) 1 : 1 or pair matching

(b) 1 : κ fixed

(c) 1 : κ variable

(d) 1 : $\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}$: 1 or full matching

FIG 2. *Different matching structures*

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 non-green buildings (exemplified in Figure 3(a)) should count more than 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; so, subject to the same constraints on covariate balance, the first matching should be 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.

(a) 5 matches with 10 units

(b) 5 matches with 8 units

(c) 5 matches with 6 units
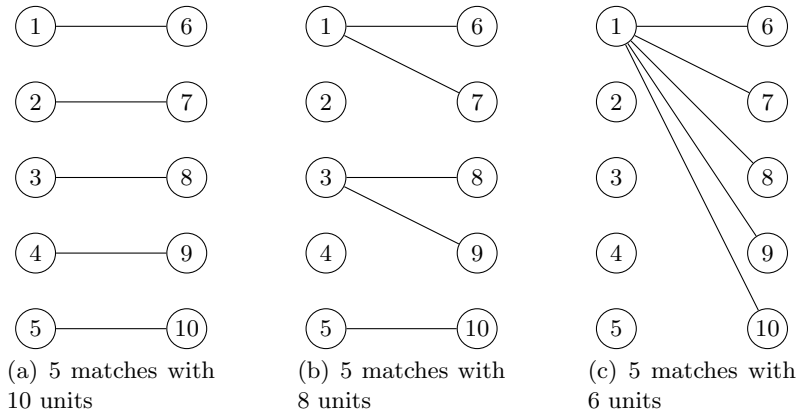
FIG 3. *Different matching structures with the same number of matches*

2.3. *Information content of a matched sample.* Let $i \in \mathcal{I} = \{1, 2, ..., I\}$ index the set of matched groups and $j \in \mathcal{J}_i = \{1, 2, ..., J_i\}$ 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.

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

$$(2.1) \qquad Y_{ij} = \alpha_i + \beta Z_{ij} + \varepsilon_{ij}$$

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 assume the outcome variance is constant across units. Consider the matched group difference in outcomes

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

where κ_i is the number of controls units in matched group i. We can calculate

(a) 6 units in two matches

(b) 6 units in one match
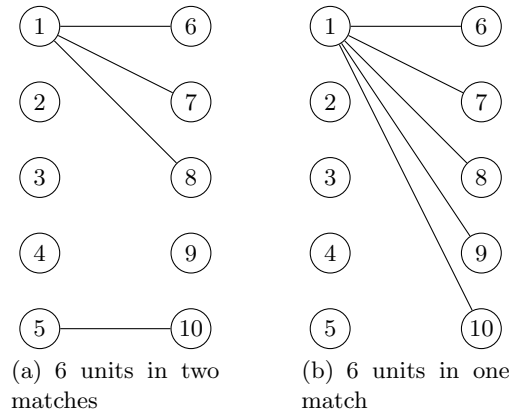
FIG 4. *Different matching patterns with the same number of matched units*

the variance of this difference and find that

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

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

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

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 informations 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 of the average of the group differences

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

However, we find that this particular definition is somewhat restrictive, as 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 et al. (2014), 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 informations of the matched groups. Building upon this notion of information content, in the next section we present a general matching framework and specific matching formulations 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 (non-green 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 $\boldsymbol{x}_t = \{x_{t,p_1}, \ldots, x_{t,p_P}\}$, and each control $c \in \mathcal{C}$ has a similar vector $\boldsymbol{x}_c = \{x_{c,p_1}, \ldots, x_{c,p_P}\}$. 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

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

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 (in our study, what drives buildings' rent). Ideally 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

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

(3.3)

$$\mathcal{M} = \left\{ \sum_{c \in \mathcal{C}} m_{tc} = \kappa, \ t \in \mathcal{T} \text{ if } \kappa > 1 \text{ and } \sum_{c \in \mathcal{C}} m_{tc} \leq \kappa, \ t \in \mathcal{T} \text{ if } \kappa = 1; \right.$$

$$\left. \sum_{t \in \mathcal{T}} m_{tc} \leq 1, \ c \in \mathcal{C}; m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C} \right\},$$

(3.4)

$$\mathcal{B} = \left\{ -\varepsilon_p \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc} \leq \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc}(f(x_{t,p}) - f(x_{c,p})) \leq \varepsilon_p \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc}, \right.$$

$$\left. m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; p \in \mathcal{P} \right\},$$

where $\varepsilon_p \geq 0$ 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.[6] Similar ideas can be used to balance the interactions of several nominal covariates. See Zubizarreta (2012) and Zubizarreta et al. (2014) for more balancing examples.

---

[6]This technique is called fine balance (Rosenbaum et al., 2007) and it has the effect of exactly balancing the mean of every linear combination of the categories of the covariates finely balanced.

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

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

(3.6)
$$\mathcal{M} = \left\{ \sum_{c \in \mathcal{C}} m_{tc} = n_t, \ t \in \mathcal{T}; n_t \leq \kappa_{\mathcal{C}}, \ t \in \mathcal{T}; \sum_{t \in \mathcal{T}} m_{tc} \leq 1, \ c \in \mathcal{C}; \right.$$
$$\left. m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 0, \ t \in \mathcal{T} \right\},$$

(3.7)
$$\mathcal{B} = \left\{ -\varepsilon_p \sum_{t \in \mathcal{T}} h^{(n_t)} \leq \sum_{t \in \mathcal{T}} h^{(n_t)} x_{t,p} - \sum_{c \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} m_{tc} \frac{h^{(n_t)}}{n_t} \right) x_{c,p} \leq \varepsilon_p \sum_{t \in \mathcal{T}} h^{(n_t)}, \right.$$
$$\left. p \in \mathcal{P}, m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 0, \ t \in \mathcal{T} \right\}.$$

Here, we let $f(x) = x$ for mean balance. Note that by using transformations of the covariates, it is possible to balance statistics other than 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 ($t \in \mathcal{T}, r \in \{1, \ldots, \kappa_{\mathcal{C}-1}\}$). This new decision variable can be written using linear constraints as

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

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

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

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

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. Define

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

where $\frac{h^{(0)}}{0}$ is set to 0. The expression of interest becomes $m_{tc} w_t^{(2)}$ which is still not linear. Therefore, 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. It can be written using linear constraints as

(3.12)$\qquad\qquad q_{tc} \;\le\; m_{tc}, \; t \in \mathcal{T}, c \in \mathcal{C}$

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

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

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

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

This program is no longer a pure integer programming (IP) problem, as 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 non-green buildings with largest information content (3.5) that satisfies the matching structure (3.6) and that balances the original covariates in the sense of (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 non-green 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 et al. (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).

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

Covariate	Standardized difference in means	
	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 ≤ 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, whereas with a variable 1 : 4 ratio it is 941. In other words, matching with a variable 1 : 4 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.

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

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 somewhat smaller than with our method (730 versus 940.6) and there are substantial 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 $1:\kappa_{\mathcal{C}}$ ratio, (3.5)-(3.15), as in our study, and also matching with a flexible $1:\kappa_{\mathcal{C}}/\kappa_{\mathcal{T}}:1$ 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 $1:\kappa_{\mathcal{C}}$ ratio with harmonic means has additional $T \times (\kappa_{\mathcal{C}} + C)$ continuous decision variables and $T \times (2 \times \kappa_{\mathcal{C}} + 3 \times C - 1)$ 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 256000 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 1990's to one second today (Nemhauser, 2013). These major improvements have been made possible by a combina-

tion 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 15-minute time limit. Among these subproblems, one gives the optimal solution within the time limit, and the others give solutions with about 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 non-green buildings. The 95% confidence interval associated to this estimate is [1.3%, 5.5%] [obtained using the inferential procedures in Hansen, Rosenbaum, and Small (2014)]. For comparison, this estimate is moderately larger than the one of Eichholtz et al. (2010), who reported that green buildings have rental rates 2.8% higher per square foot than similar non-green buildings (with 95% confidence interval of [1%, 4.6%]).

In principle, our estimand is not the same as the one of Eichholtz et al. (2010), since our approach restricts the analysis to the sample with largest information that is balanced, usually discarding some treated units (in our study, these are 19 out of the 694 green buildings available before matching). To get a better understanding 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, 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, so the results are only moderately insensitive to hidden biases [see Rosenbaum and Silber (2009) and Hansen et al. (2014) for details of this analysis].

To interpret these results, let us remember that about 30% of building operating costs are driven by energy consumption and that green buildings typically have 25% less energy use and in aggregate 19% lower operating costs. Therefore, in rough 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 et al. (2010) mention, even a small improvement on the energy use of existing buildings has a big impact not only on the economy but also on the environment. In this way, companies are also willing to pay more to "go green" for a sustainable environment.

6. Discussion of the proposed matching methods. 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 for this 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 et al., 2007). For these reasons, in view of the bias-variance —or, stated differently, the balance-precision— tradeoff 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. Interestingly, 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 non-constant variances across units, the formulations in Section 3.3 and Appendix C for matching with a $1 : \kappa_{\mathcal{C}}$ variable ratio and the more flexible $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ matching ratio, respectively, may be more appropriate.[7] As discussed in Section 2.3, these formulations are not only easier to implement but also more intuitive as they maximize the sum of the Fisher informations 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 such as fine balance (Rosenbaum et al., 2007) and strength-k matching (Hsu et al., 2015); for a related discussion, see Zubizarreta (2012).

Assessing overlap or lack of common support in covariate distributions is a widespread practice undertaken in observational studies in order 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 (Crump et al., 2009) 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

---

[7] In model (2.1) we assumed that the variance is constant across units. One way to relax this assumption is to suppose instead that the variance in the treated group is f times bigger than the variance in the control group. Then $h^{(\kappa)}$ becomes the sum of the harmonic means of 1 (treated unit) and κ_i/f ("control" units) for each matched group (as opposed to 1 and κ_i, as before). As another example, suppose that the variance in one category of a binary covariate is f times bigger than in the other category. Then the weighting becomes $h^{(\kappa_i)}/f$ for the matched group with greater variance and emphasizing to match f times as many groups from the strata with smaller variance. Extending on 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.

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 average treatment effect on the treated and it will become a more local estimand, the average treatment effect on the matched treated units. In view of this limitation of 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) and Fogarty et al. (2015)]. Another way to proceed is by weighting the matched samples to a target population of greater policy interest perhaps by using the method in Zubizarreta (2015).

In cardinality matching, finding the largest balanced matched sample is followed by re-matching the pairs or groups that constitute the matched sample to minimize their total sum of covariate distances. If these covariates are predictive of the outcome, this re-matching 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 et al. (2001), Lu (2005)]. Under weaker identification assumptions than those of "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 et al., 2015).

7. Summary. In this paper we revisited the study of Eichholtz et al. (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 et al. (2010)]. Thus, besides being environmentally responsible it is also an economically sound decision to pursuit environmentally sustainable building practices.

References.

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, University of Pennsylvania.*

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," *Journal of the American Statistical Association*, 105, 1285–1296.

Bertsekas, D. P. (1981), "A New Algorithm for the Assignment Problem," *Mathematical Programming*, 21, 152–171.

Bertsimas, D. (2014), "Statistics and Machine Learning via a Modern Optimization Lens," *The 2014-2015 Philip McCord Morse Lecture.*

Bixby, R. E. and Rothberg, E. (2007), "Progress in Computational Mixed Integer Programming—A Look Back From the Other Side of the Tipping Point," *Annals of Operations Research*, 149, 37–41.

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," *Journal of the Royal Statistical Society*, forthcoming.

Cochran, W. and Rubin, D. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Cochran, W. G. (1965), "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society*, 128, 234–266.

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.

Diamond, A. and Sekhon, J. S. (2013), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," *The Review of Economics and Statistics*, 95, 932–945.

Eichholtz, P., Kok, N., and Quigley, J. M. (2010), "Doing Well by Doing Good? Green Office Buildings," *American Economic Review*, 2492–2509.

Fogarty, C., Mikkelsen, M., Gaieski, D., and Small, D. (2015), "Discrete Optimization for Interpretable Study Populations and Randomization Inference in an Observational Study of Severe Sepsis Mortality," *Journal of the American Statistical Association*, forthcoming.

Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies*, 79, 1053–1079.

Hainmueller, J. (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis*, 20, 25–46.

Hansen, B. B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

— (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," *Statistical Science*, 23, 219–236.

Hansen, B. B. and Klopfer, S. O. (2006), "Optimal Full Matching and Related Designs Via Network Flows," *Journal of Computational and Graphical Statistics*, 15, 609–627.

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014), "Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies," *Journal of the American Statistical Association*, 109, 133–144.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007), "Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study," *Psychological Methods*, 12, 247.

Hill, J. L. (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," *Statistics in Medicine*, 27, 2055–2061.

Hsu, J. Y., Zubizarreta, J. R., Small, D. S., and Rosenbaum, P. R. (2015), "Strong Control of the Family-Wise Error Rate in Observational Studies that Discover Effect Modification by Exploratory Methods," *Biometrika*, forthcoming.

Iacus, S. M., King, G. K., and Porro, G. (2012), "Causal Inference Without Balance Checking: Coarsened Exact Matching," *Political Analysis*, 20, 1–24.

Imbens, G. W. (2015), "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50, 373–419.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kalton, G. (1968), "Standardization: A Technique to Control for Extraneous Variables," *Applied Statistics*, 17, 118–136.

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," *Journal of the Royal Statistical Society: Series A*, 178, 223–239.

Kuhn, H. (1955), "The Hungarian Method for the Assignment Problem," *Naval research logistics quarterly*, 2, 83–97.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2016), "Balancing Covariates via Propensity Score Weighting," *Working Paper*.

Li, Y. P., Propert, J. K., and Rosenbaum, P. R. (2001), "Balanced Risk Set Matching," *Journal of the American Statistical Association*, 96, pp. 870–882.

Linderoth, J. T. and Lodi, A. (2010), "MILP Software," in *Wiley Encyclopedia of Operations Research and Management Science*, eds. Cochran, J. J., Cox, L. A., Keskinocak, P., Kharoufeh, J. P., and Smith, J. C., Wiley.

Lu, B. (2005), "Propensity Score Matching with Time-Dependent Covariates," *Biometrics*, 61, 721–728.

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." *Operations Research*, 61, 398–412.

Papadimitriou, C. (1994), *Computational Complexity*, Reading (Mass.): Addison-Wesley.

Papadimitriou, C. H. and Steiglitz, K. (1982), *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall.

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," *Journal of the American Statistical Association*, 110, 515–527.

Rosenbaum, P. (1991), "Discussing Hidden Bias in Observational Studies," *Archives of Internal Medicine*, 115, 901– 905.

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

— (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (1993), "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies," *Journal of the American Statistical Association*, 88, 1250–1253.

— (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot," *Biometrika*, 88, 219–231.

— (2002), *Observational Studies*, Springer.

— (2005), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *The American Statistician*, 59, 147–152.

— (2010), *Design of Observational Studies*, Springer.

— (2014), "Weighted M-statistics With Superior Design Sensitivity in Matched Observational Studies With Multiple Controls," *Journal of the American Statistical Association*, 109, 1145–1158.

— (2015), "How to See More in Observational Studies: Some New Quasi-Experimental Devices," *Annual Review of Statistics and Its Application*, 2, null.

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," *Journal of the American Statistical Association*, 102, 75–83.

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.

— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 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 Observational Studies," *Journal of the American Statistical Association*, 104, 1398–1405.

Rubin, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

— (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1–21.

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.

Weston, S. and Calaway, R. (2014), *Getting Started with `doParallel` and `foreach`*.

Yang, D., Small, D., 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.

Yoon, F. (2009), "Entire Matching and its Application in an Observational Study of Treatments for Melanoma," *Dissertation, Department of Statistics, The Wharton School, University of Pennsylvania*.

Zubizarreta, J. R. (2012), "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

— (2015), "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data," *Journal of the American Statistical Association*, 110, 910–922.

Zubizarreta, J. R. and Kilcioglu, C. (2015), "`mtchdsgn`: Optimal Matched Designs for Randomized Experiments and Observational Studies in `R`," `R` package version 0.0-1.

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," *Annals of Applied Statistics*, 8, 204–231.

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," *The American Statistician*, 65, 229–238.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. A., and Rosenbaum, P. R. (2013), "Stronger Instruments Via Integer Programming in an Observational Study of Late Preterm Birth Outcomes," *Annals of Applied Statistics*, 7, 25–50.

APPENDIX A: COMPUTATIONAL COMPLEXITY OF CARDINALITY MATCHING

In complexity theory, computational problems are categorized in terms of their inherent difficulty, usually in connection with the time it takes to find a solution (Papadimitriou, 1994). While some problems can be solved quickly, with algorithms that run in polynomial time, other problems cannot. Problems that can be solved with polynomial time algorithms are considered tractable in the sense that the number of arithmetic steps it takes to solve a problem instance increases as a polynomial function of the size of the problem. For instance, the assignment problem of matching treated and controls units to minimize the total sum of covariate distances between matched units [as in Rosenbaum (1989)] is considered tractable because it has a worst-case time bound of $O(U^3)$ where U is the number of units available before matching [Kuhn (1955), Bertsekas (1981), Papadimitriou and Steiglitz (1982)]. General IP and MIP problems are NP-hard in the sense that no polynomial time algorithm has been found to solve any problem in their general class so far.

Cardinality matching (3.2)-(3.4) is an IP problem and, although a polynomial time algorithm has not been found to solve this specific problem, from a user standpoint the time it takes in practice to solve a typical instance of this problem is comparable to the time it takes to solve the assignment problem. In the cardinality matching problem the constraint matrix defined by (3.3)-(3.4) is not totally unimodular (meaning that the feasible region it defines is not an integral polyhedron, so the problem cannot be solved by relaxing the original problem and solving a linear program as in the assignment problem), however there is much structure in the constraints (3.3)-(3.4) so it can be solved in reasonable time with modern optimization solvers.

APPENDIX B: MATCHING TO MINIMIZE THE VARIANCE OF A DIFFERENCE-IN-MEANS EFFECT ESTIMATOR

Consider the effect estimator (2.5) and calculate its variance

(B.1)
$$\sum_{i \in \mathcal{I}} \text{Var}(\hat{\delta}) = \frac{\sigma^2}{I^2} \sum_{i \in \mathcal{I}} \left(1 + \frac{1}{\kappa_i} \right).$$

Ideally, within the matching framework of (3.1), we would define the information content $\mathbb{I}$ as the inverse of this variance; however, since the number of matched pairs I is also a decision variable, the resulting optimization problem is very complicated. A simplification is to fix I by matching all the treated units with a variable $1 : \kappa_{\mathcal{C}}$ ratio. For fixed I, the variance of the effect estimator is proportional to

(B.2)
$$\sum_{i \in \mathcal{I}} \text{Var}(\hat{\delta}) \propto \sum_{i \in \mathcal{I}} \frac{1}{\kappa_i}.$$

Put $\ell^{(\kappa_i)} = \frac{1}{\kappa_i}$. Since maximizing the inverse of the variance is equivalent to minimizing the variance, the problem we want to solve can be written as

(B.3)
$$\min_{\boldsymbol{m}, \boldsymbol{n}} \{ \mathbb{V}(\boldsymbol{m}, \boldsymbol{n}) : (\boldsymbol{m}, \boldsymbol{n}) \in \mathcal{M} \cap \mathcal{B} \}$$

where

(B.4)
$$\mathbb{V}(\boldsymbol{m}, \boldsymbol{n}) = \sum_{t \in \mathcal{T}} \ell^{(n_t)},$$

(B.5)
$$\mathcal{M} = \left\{ \sum_{c \in \mathcal{C}} m_{tc} = n_t, \ t \in \mathcal{T}; n_t \leq \kappa_{\mathcal{C}}, \ t \in \mathcal{T}; \sum_{t \in \mathcal{T}} m_{tc} \leq 1, \ c \in \mathcal{C}; \right.$$
$$\left. m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 1, \ t \in \mathcal{T} \right\},$$

(B.6)
$$\mathcal{B} = \left\{ -\varepsilon_p T \leq \sum_{t \in \mathcal{T}} x_{t,p} - \sum_{c \in \mathcal{C}} \left(\sum_{t \in \mathcal{T}} m_{tc} \ell^{(n_t)} \right) x_{c,p} \leq \varepsilon_p T, \right.$$
$$\left. p \in \mathcal{P}, m_{tc} \in \{0, 1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; n_t \geq 1, \ t \in \mathcal{T} \right\}.$$

Note that the set $\mathcal{B}$ above is only written for mean balancing constraints. In a similar way to the model described in the paper, $\ell^{(n_t)}$ and $m_{tc}\ell^{(n_t)}$ have to be linearized using $m_t^{(r)}$s ($t \in \mathcal{T}, r \in \{2, 3, \ldots, \kappa_{\mathcal{C}}-1\}$). The only difference is we do not need $m_t^{(1)}$, it is set to 1 since all treated units are forced to be matched with at least one control unit in this formulation. Therefore,

$$(\text{B.7}) \qquad m_t^{(r)} \quad \leq \quad n_t - \sum_{s=2}^{r-1} m_t^{(s)} - 1, \ t \in \mathcal{T}, r \in \{2, \ldots, \kappa_{\mathcal{C}-1}\}$$

$$(\text{B.8}) \qquad \kappa_{\mathcal{C}} m_t^{(r)} \quad \geq \quad n_t - \sum_{s=2}^{r-1} m_t^{(s)} - 1, \ t \in \mathcal{T}, r \in \{2, \ldots, \kappa_{\mathcal{C}-1}\},$$

and

(B.9)
$$w_t := \ell^{(n_t)}$$
$$= 1 + \sum_{s=2}^{\kappa_{\mathcal{C}}-1} \left(\frac{1}{s} - \frac{1}{s-1} \right) m_t^{(s)} + \left(\frac{1}{\kappa_{\mathcal{C}}} - \frac{1}{\kappa_{\mathcal{C}}-1} \right) \left(n_t - \sum_{s=2}^{\kappa_{\mathcal{C}}-1} m_t^{(s)} - 1 \right).$$

To linearize $m_{tc}\ell^{(n_t)}$, define $q_{tc} = m_{tc}\ell^{(n_t)}$ which can be formulated as

$$(\text{B.10}) \qquad q_{tc} \quad \leq \quad m_{tc}, \ t \in \mathcal{T}, c \in \mathcal{C}$$
$$(\text{B.11}) \qquad q_{tc} \quad \leq \quad w_t, \ t \in \mathcal{T}, c \in \mathcal{C}$$
$$(\text{B.12}) \qquad q_{tc} \quad \geq \quad w_t - (1 - m_{tc}), \ t \in \mathcal{T}, c \in \mathcal{C}.$$

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

$$(\text{B.13}) \qquad -\varepsilon_p T \leq \sum_{t \in \mathcal{T}} x_{t,p} - \sum_{c \in \mathcal{C}} w_c x_{c,p} \leq \varepsilon_p T, p \in \mathcal{P}.$$

APPENDIX C: MATCHING WITH A FLEXIBLE $1 : \kappa_{\mathcal{C}}/\kappa_{\mathcal{T}} : 1$ RATIO

First, let us define g_t and g_c

$$(\text{C.1}) \qquad g_t = \begin{cases} h^{(n_t)} & \text{if } n_t \geq 2 \\ \sum_{c \in \mathcal{C}} m_{tc} \dfrac{h^{(n_c)}}{n_c} & \text{if } n_t \leq 1 \end{cases} \qquad t \in \mathcal{T},$$

$$(\text{C.2}) \qquad g_c = \begin{cases} h^{(n_c)} & \text{if } n_c \geq 2 \\ \sum_{t \in \mathcal{T}} m_{tc} \dfrac{h^{(n_t)}}{n_t} & \text{if } n_c \leq 1 \end{cases} \qquad c \in \mathcal{C}.$$

Then, the problem can be formulated as

$$(C.3) \qquad \mathbb{I}(\boldsymbol{m}, \boldsymbol{n}) = \sum_{t \in \mathcal{T}} g_t,$$

$$(C.4) \qquad \mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3$$

where
(C.5)

$$\mathcal{M}_1 = \left\{ \sum_{c \in \mathcal{C}} m_{tc} = n_t, \ t \in \mathcal{T}; n_t \le \kappa_{\mathcal{C}}, \ t \in \mathcal{T}; m_{tc} \in \{0,1\}, n_t \ge 0, \ t \in \mathcal{T} \right\},$$

(C.6)

$$\mathcal{M}_2 = \left\{ \sum_{t \in \mathcal{T}} m_{tc} = n_c, \ c \in \mathcal{C}; n_c \le \kappa_{\mathcal{T}}, \ c \in \mathcal{C}; m_{tc} \in \{0,1\}, n_c \ge 0, \ c \in \mathcal{C} \right\},$$

(C.7)

$$\mathcal{M}_3 = \{ (n_t - 1)(n_c - 1)m_{tc} = 0, \ t \in \mathcal{T}, c \in \mathcal{C}; m_{tc} \in \{0,1\}, \ t \in \mathcal{T}, c \in \mathcal{C};$$
$$n_t \ge 0, \ t \in \mathcal{T}; n_c \ge 0, \ c \in \mathcal{C} \},$$

and

(C.8)

$$\mathcal{B} = \left\{ -\varepsilon_p \sum_{t \in \mathcal{T}} g_t \le \sum_{t \in \mathcal{T}} g_t x_{t,p} - \sum_{c \in \mathcal{C}} g_c x_{c,p} \le \varepsilon_p \sum_{t \in \mathcal{T}} g_t, \ p \in \mathcal{P}; \right.$$
$$\left. m_{tc} \in \{0,1\}, \ t \in \mathcal{T}, c \in \mathcal{C}; n_t \ge 0, \ t \in \mathcal{T}; n_c \ge 0, \ c \in \mathcal{C} \right\}.$$

Note that g_t, g_c and the constraint set $\mathcal{M}_3$ have to be written in linear form. Let us define $m_c^{(r)}$, $w_c^{(1)}$ and $w_c^{(2)}$ analogous to $m_t^{(r)}$, $w_t^{(1)}$, and $w_t^{(2)}$. The decision variable $m_c^{(r)}$ is equal to 1 if control unit c is matched with at least r number of treated units, and 0 otherwise ($c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\}$). With linear constraints

$$(C.9) \qquad m_c^{(r)} \ \le \ n_c - \sum_{s=1}^{\kappa_{\mathcal{T}} - 1} m_c^{(s)}, \qquad c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\}$$

$$(C.10) \quad \kappa_{\mathcal{T}} m_c^{(r)} \ \ge \ n_c - \sum_{s=1}^{\kappa_{\mathcal{T}} - 1} m_c^{(s)}. \quad c \in \mathcal{C}, r \in \{1, \ldots, \kappa_{\mathcal{T}} - 1\}$$

Using $m_c^{(r)}$,

(C.11)

$$w_c^{(1)} := \sum_{s=1}^{\kappa_\mathcal{T}-1} \left(h^{(s)} - h^{(s-1)} \right) m_c^{(s)} + \left(h^{(\kappa_\mathcal{T})} - h^{(\kappa_\mathcal{T}-1)} \right) \left(n_c - \sum_{s=1}^{\kappa_\mathcal{T}-1} m_c^{(s)} \right),$$

(C.12)

$$w_c^{(2)} := \sum_{s=1}^{\kappa_\mathcal{T}-1} \left(\frac{h^{(s)}}{s} - \frac{h^{(s-1)}}{s-1} \right) m_c^{(s)} + \left(\frac{h^{(\kappa_\mathcal{T})}}{\kappa_\mathcal{T}} - \frac{h^{(\kappa_\mathcal{T}-1)}}{\kappa_\mathcal{T}-1} \right) \left(n_c - \sum_{s=1}^{\kappa_\mathcal{T}-1} m_c^{(s)} \right),$$

where $\frac{h^{(0)}}{0}$ is set to 0.

Now, we can rewrite g_t and g_c as

(C.13)
$$g_t = \begin{cases} w_t^{(1)}, & \text{if } m_t^{(2)} = 1 \\ \sum_{c \in \mathcal{C}} m_{tc} w_c^{(2)}, & \text{if } m_t^{(2)} = 0 \end{cases}, \qquad t \in \mathcal{T}$$

(C.14)
$$g_c = \begin{cases} w_c^{(1)}, & \text{if } m_c^{(2)} = 1 \\ \sum_{t \in \mathcal{T}} m_{tc} w_t^{(2)}, & \text{if } m_c^{(2)} = 0 \end{cases}. \qquad c \in \mathcal{C}$$

The expressions $m_{tc} w_c^{(2)}$ and $m_{tc} w_t^{(2)}$ are still not in linear form; therefore, we define two new sets of decision variables $u_{tc} := m_{tc} w_c^{(2)}$ and $v_{tc} := m_{tc} w_t^{(2)}$, and formulate them in the following way:

(C.15)	$u_{tc} \leq m_{tc},$	$t \in \mathcal{T}, c \in \mathcal{C}$
(C.16)	$u_{tc} \leq w_c^{(2)},$	$t \in \mathcal{T}, c \in \mathcal{C}$
(C.17)	$u_{tc} \geq w_c^{(2)} - (1 - m_{tc}),$	$t \in \mathcal{T}, c \in \mathcal{C}$
(C.18)	$v_{tc} \leq m_{tc},$	$t \in \mathcal{T}, c \in \mathcal{C}$
(C.19)	$v_{tc} \leq w_t^{(2)},$	$t \in \mathcal{T}, c \in \mathcal{C}$
(C.20)	$v_{tc} \geq w_t^{(2)} - (1 - m_{tc}).$	$t \in \mathcal{T}, c \in \mathcal{C}$

As the last step, we write g_t and g_c using conditional constraints

$$(C.21) \qquad g_t \leq h^{(\kappa_C)} m_t^{(2)} + \sum_{c \in \mathcal{C}} u_{tc}, \qquad\qquad t \in \mathcal{T}$$

$$(C.22) \qquad g_t \leq 1 - m_t^{(2)} + w_t^{(1)}, \qquad\qquad t \in \mathcal{T}$$

$$(C.23) \qquad g_t \geq \sum_{c \in \mathcal{C}} u_{tc} - \kappa_{\mathcal{C}} m_t^{(2)}, \qquad\qquad t \in \mathcal{T}$$

$$(C.24) \qquad g_t \geq w_t^{(1)} - \left(1 - m_t^{(2)}\right) h^{(\kappa_C)}, \qquad\qquad t \in \mathcal{T}$$

$$(C.25) \qquad g_c \leq h^{(\kappa_{\mathcal{T}})} m_c^{(2)} + \sum_{t \in \mathcal{T}} v_{tc}, \qquad\qquad c \in \mathcal{C}$$

$$(C.26) \qquad g_c \leq 1 - m_c^{(2)} + w_c^{(1)}, \qquad\qquad c \in \mathcal{C}$$

$$(C.27) \qquad g_c \geq \sum_{t \in \mathcal{T}} v_{tc} - \kappa_{\mathcal{T}} m_c^{(2)}, \qquad\qquad c \in \mathcal{C}$$

$$(C.28) \qquad g_c \geq w_c^{(1)} - \left(1 - m_c^{(2)}\right) h^{(\kappa_{\mathcal{T}})}. \qquad\qquad c \in \mathcal{C}$$

Finally, constraint set $\mathcal{M}_3$ can be written in linear form as

$$(C.29) \qquad m_t^{(2)} + m_c^{(2)} \leq 2 - m_{tc}. \qquad\qquad t \in \mathcal{T}, c \in \mathcal{C}$$

APPENDIX D: RUNNING TIMES

Here we present the results of a small simulation study to provide a sense of the running times of the proposed methods. In the original data, we have 694 clusters with one treated (green) building and multiple control buildings. In the simulation study, we randomly selected 100, 500 and 2000 clusters with replacement, and for each of these number of clusters we tested our method with different number of covariates: 5, 10, 20, and 50. For covariate sizes 5 and 10 we randomly selected the covariates from our covariate set, and for covariate sizes 20 and 50 (since there are not that many covariates to begin) we included interactions of covariates that are fairly independent of each other. As in the actual study, we divided each of the optimization problem into 10 subproblems using exact matching constraints (as explained in Appendix E) and solved each of them in parallel. We gave a time limit of 60 minutes to each of the problems. The results are presented in the following tables.

Table 3(a) presents the running times and optimality gaps for the method used in the actual study; this is, matching with a variable 1 : 4 ratio with the weighted balancing constraints (3.7). In each cell of the table, the first row shows the running time of the optimization problem, which is the maximum running time of the 10 subproblems. (Since these problems are run in

parallel, the total running time of the optimization portion is the maximum running time of all the subproblems. A running time greater than 60 minutes indicates that an optimal solution would be found after the reported duration.) The second row shows the optimality gaps in terms of the maximum effective sample size reached in the given time limit and the tightest upper bound found by the solver after the branch and bound procedure also within the time limit. One can evaluate how close the provided solution is to the theoretical solution from these numbers. In the table, we observe that one can obtain relatively small optimality gaps within the time limit for samples of size $(n_t, n_c) \approx (500, 5000)$, and for larger sample sizes if the number of covariates is smaller than 10.

As described in Appendix E, one way to decrease the complexity of the problem, and therefore to reduce computing times is by omitting the weights in the balancing constraints. Table 3(b) presents the results for this approach with a time limit of 15 minutes for each of the optimization problems. Within this time limit, it is possible to find solutions with a small optimality gap for all the instances in the table.

Table 3

Running times and optimality gaps for matching for different combinations of sample sizes and number of covariates. The running times are reported in minutes and the optimality gaps appear in terms of two numbers: the best solution found within the given time limit and the bounding (perhaps infeasible) solution found also within the time limit.

(a) *Matching with a variable $1:4$ ratio with the weighted balancing constraints* (3.7)

Number of units	Number of covariates			
(n_t, n_c)	5	10	20	50
(100, 1228)	0.1	60.0	60.0	0.1
	130.3–130.3	123.1–123.8	77.2–77.5	62.2–62.2
(500, 5237)	60.0	60.0	60.0	60.0
	673.4–690.6	655.2–678.4	617.8–646.8	535.7–574.9
(1000, 10806)	60.0	60.0	60.0	60.0
	1355.5–1405.8	1332.0–1391.4	1274.4–1365.1	1165.8–1286.8
(2000, 21190)	60.0	60.0	60.0	60.0
	2700.3–2876.7	2630.0–2869.2	2468.7–2864.6	2244.4–2750.7

(b) *Matching with a variable $1:4$ ratio with the unweighted balancing constraints* (3.4)

Number of units	Number of covariates			
(n_t, n_c)	5	10	20	50
(100, 1228)	0.1	0.1	0.1	0.1
	132.4–132.4	127.0–127.0	78.9–78.9	66.2–66.2
(500, 5237)	0.1	0.2	15.0	15.0
	700.5–700.5	692.1–692.1	658.4–658.4	585.8–592.5
(1000, 10806)	1.7	0.3	15.0	15.0
	1405.3–1405.3	1397.2–1397.2	1366.0–1367.9	1288.6–1293.1
(2000, 21190)	0.5	15.0	15.0	15.0
	2841.1–2841.1	2827.8–2827.9	2781.0–2782.7	2677.0–2683.5

APPENDIX E: DEVICES FOR SPEED

One tactic for more quickly solving the previous matching problems is exact matching for nominal covariates of prognostic relevance or which are to be used for subgroup analyses. Let $x_{\cdot,p}$ be a nominal covariate taking integer values $\tilde{n} \in \mathcal{N} \subset \mathbb{N}$. To match exactly for $x_{\cdot,p}$, one possibility is to include the constraint

$$(E.1) \qquad \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc} \left| \mathbb{1}_{x_{t,p}=\tilde{n}} - \mathbb{1}_{x_{c,p}=\tilde{n}} \right| = 0, \ \forall \tilde{n} \in \mathcal{N},$$

where $\mathbb{1}$ is the indicator function. Exact matching constraints reduce the feasible region considerably and therefore the optimal solution is found faster. Another possibility to match exactly for $x_{\cdot,p}$ is to divide the dataset into smaller, mutually exclusive and collectively exhaustive pieces based on the categories of $\mathcal{N}$ and solve a matching problem for each piece in parallel. If

the problem is run on a machine with multiple processors and/or multiple cores, each subproblem can be assigned to be solved independently by a processing unit. The default settings in R do not use all the cores available in the machine running the code; however, there are some packages available to create a parallel backend so that independent subproblems can be solved simultaneously on different processing units [see, for instance, Weston and Calaway (2014) for the R packages doParallel and foreach].

Other tactics that can be used to attain computational speedups include simplifying the matching problem by eliminating the harmonic mean weights from the balancing constraints (but not the objective function) or using Yoon's (2009) entire number to determine the matching ratio for each unit before matching (Zubizarreta, 2012). However, we do not recommend the first of these approaches because it results in an inconsistency between the balance criteria used to assess the quality of the match and the balance criteria needed for unbiased estimation with an estimator that uses the harmonic mean weights. Also, we are not enthusiastic about the second approach because it requires that one estimate the propensity score in order to calculate the entire number.

APPENDIX F: DESCRIPTION OF THE MATCHED SAMPLE

Table 4 below describes the samples of green buildings before matching, after matching and of those green buildings that were left out from the matched analyses due to lack of good controls. We observe that the sample of matched green buildings is very similar to that of all the green buildings (after all, only 19 green buildings were unmatched and left out from the analyses). Among others, the unmatched buildings are larger on average, have better quality (are all in class A and have a higher proportion of amenities), are not very old, less of them are renovated, and have high stories.

TABLE 4

Means and sizes of the samples of green buildings before matching ("All"), after matching ("Matched") and of those green buildings that were left out from the analyses due to lack of good controls ("Unmatched").

Covariate	Sample		
	All	Matched	Unmatched
Building size	0.324	0.327	0.520
Building class A	0.794	0.780	1.000
Building class B	0.195	0.207	0.000
Building class C	0.012	0.012	0.000
Net contract	0.058	0.059	0.053
Employment growth	0.035	0.037	-0.028
Employment growth missing	0.009	0.009	0.000
Age ≤ 10 years	0.143	0.140	0.158
Age 11-20 years	0.241	0.234	0.316
Age 21-30 years	0.434	0.425	0.526
Age 31-40 years	0.111	0.120	0.000
Age >40 years	0.059	0.066	0.000
Age missing	0.013	0.014	0.000
Renovated	0.210	0.213	0.158
Stories low	0.463	0.455	0.211
Stories intermediate	0.267	0.264	0.263
Stories high	0.271	0.281	0.526
Stories missing	0.000	0.000	0.000
Amenities	0.718	0.711	0.895
Sample size	694	675	19

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

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