Haphazard intentional allocation and rerandomization to improve covariate balance in experiments

Marcelo S. Lauretto, Rafael B. Stern, Kari L. Morgan, Margaret H. Clark, and Julio M. Stern

Citation: AIP Conference Proceedings **1853**, 050003 (2017); doi: 10.1063/1.4985356

View online: http://dx.doi.org/10.1063/1.4985356

View Table of Contents: http://aip.scitation.org/toc/apc/1853/1

Published by the American Institute of Physics

Haphazard Intentional Allocation and Rerandomization to Improve Covariate Balance in Experiments

Marcelo S. Lauretto[1], Rafael B. Stern[2], Kari L. Morgan[3], Margaret H. Clark[4] and Julio M. Stern[1,a)]

[1]*Universidade de S˜ao Paulo*, Brazil
[2]*Universidade Federal de S˜ao Carlos, Brazil*
[3]*Penn State University, USA*
[4]*University of Central Florida, USA*

[a)]Corresponding author: jstern@ime.usp.br

Abstract. In randomized experiments, a single random allocation can yield groups that differ meaningfully with respect to a given covariate. Furthermore, it is only feasible to use classical control procedures of allocation for a very modest number of covariates. As a response to this problem, Morgan and Rubin [11, 12] proposed an approach based on *rerandomization* (repeated randomization) to ensure that the final allocation obtained is balanced. However, despite the benefits of the rerandomization method, it has an exponential computational cost in the number of covariates, for fixed balance constraints. Here, we propose the use of *haphazard intentional allocation*, an alternative allocation method based on optimal balance of the covariates extended by random noise, see Lauretto et al. [7]. Our proposed method can be divided into a randomization and an optimization step. The randomization step consists of creating new (artificial) covariates according a specified distribution. The optimization step consists of finding the allocation that minimizes a linear combination of the imbalance in the original covariates and the imbalance in the artificial covariates. Numerical experiments on real and simulated data show a remarkable superiority of haphazard intentional allocation over the rerandomization method, both in terms of balance between groups and in terms of inference power.

INTRODUCTION

This paper addresses the problem of allocation in the design of experiments, which is illustrated with the following example: Consider a research laboratory which develops a new drug. In order to test the effectiveness of this drug, the laboratory may treat some patients with the new drug and some with a placebo. The problem of allocation consists of determining, for each patient in the trial, whether he will be treated with the new drug or the placebo. Often, the main interest is in understanding the effectiveness of the drug according to some covariates, such as gender, age, blood type, etc. In order to obtain meaningful conclusions from the study, the researchers often wish the allocation to be balanced, in the sense that the distribution of the covariates be the same among the two groups of patients (new drug and placebo), and at the same time to be free of ad-hoc interferences in allocation decisions. The standard solution for this problem involves a random allocation.

The role of randomization is a controversial subject in Bayesian statistics. This controversy stems from the following result in Decision Theory: there exists no randomized decision which is more desirable than the optimal deterministic decision. Based on this result, Lindley [8] argues that randomization is not a necessary condition for an adequate experiment. Indeed, according to Lindley, the experiment is adequate as long it is well-balanced according to the relevant covariates. In this sense, a randomized experiment might not be adequate, since it can yield groups that differ meaningfully with respect to a covariate. Lindley concludes from this approach based on Decision Theory that an adequate allocation can be obtained by deterministically selecting a balanced allocation.

However, one can argue that Decision Theory is not the adequate framework for the design of experiments. While Decision Theory is concerned with the preferences of a single agent, the design of experiments should often consider the preferences of several agents simultaneously. For instance, the goal of experiments is often to prove an hypothesis to others [6]. These situations can be covered by frameworks such as non-cooperative and cooperative game theory

Bayesian Inference and Maximum Entropy Methods in Science and Engineering
AIP Conf. Proc. 1853, 050003-1–050003-8; doi: 10.1063/1.4985356
Published by AIP Publishing. 978-0-7354-1527-0/$30.00

[1]. In these frameworks, good experimental designs often require a trade-off between purposive balance of covariates and randomization.

Morgan and Rubin [11, 12] propose an approach to allocation that satisfies these conditions. This approach is based on rerandomization (repeated randomization) to ensure that the allocation that is obtained is balanced. One can divide the algorithm in Morgan and Rubin [11] into two processes. On the base level, one obtains proposed allocations from a simple random sampling. This process guarantees the stochastic behavior of the chosen allocation. On the upper level, one rejects the proposals until the allocation that is obtained is sufficiently balanced. This process optimizes the allocation with respect to its balance.

Despite the benefits of the above algorithm, it can be hard to use it in a way that yields a highly balanced allocation at a low computational cost. The following two examples illustrate this idea in a problem of allocation into two groups of the same size:

1. The probability that a simple random sampling generates an allocation that is significantly unbalanced (at level α) for at least one out of d covariates is proportinal to $(1-\alpha)^d$. As a result, the expected number of rerandomizations that are required in order for the sample to be balanced in every covariate grows exponentially with the number of covariates.
2. Consider a group of $2n$ people such that n are male and n are female. Although there exists approximately $2^{2n}/\sqrt{n}$ allocations that are perfectly balanced with respect to gender, random sampling obtains, with high probability, allocations with an imbalance of the order of $n^{\frac{1}{2}}$ individuals. Furthermore, the expected number of simple random allocations that must be performed until perfect balance is obtained is of the order of $n^{\frac{1}{2}}$. If one were to control for d binary covariates, then the expected number of such allocations is of the order of $n^{\frac{d}{2}}$.

Here, we propose the use of *haphazard intentional allocation*, an alternative allocation process that is an adaptation of the method described in Lauretto et al. [7], see also Fossaluza et al. [2]. Similarly to the allocation process in Morgan and Rubin [11], our proposal can be divided into a randomization and an optimization step. The randomization step consists of creating new covariates that are distributed according to a standard multivariate normal. We say that these new covariates are *artificial* and that the original covariates are *of interest*. The optimization step consists of finding the allocation that minimizes a linear combination of the imbalance in the covariates of interest and the imbalance in the artificial covariates.

HAPHAZARD ALLOCATION

Let $\mathbf{X}$ denote the covariates of interest. $\mathbf{X}$ is a matrix in $\mathbb{R}^{n \times d}$, where n is the number of individuals to be allocated and d is the number of covariates of interest. An allocation consists of assigning to each individual a group chosen from a set of possible groups, $\mathcal{G}$. For simplicity, we assume that $\mathcal{G} = \{0, 1\}$. We denote an allocation, $\mathbf{w}$, by a $1 \times n$ vector in $\mathcal{G}^n$. The goal of the allocation problem is to generate an allocation that, with high probability, is close to the infimum of the imbalance between groups with respect to individuals covariate values, measured by a loss function, $L(\mathbf{w}, \mathbf{X})$.

The haphazard allocation consists of finding the minimum of a noisy version of the loss function. Let $\mathbf{Z}$ be an artificially generated matrix in $\mathbb{R}^{n \times k}$, with elements that are independent and identically distributed according to the standard normal distribution. For a given tuning parameter, $\lambda \in [0, 1]$, the haphazard allocation finds a feasible allocation, w^*, that minimizes $(1-\lambda)L(\mathbf{w}, \mathbf{X}) + \lambda L(\mathbf{w}, \mathbf{Z})$. λ controls the amount of perturbation that is added to the original loss, $L(\mathbf{w}, \mathbf{X})$. If $\lambda = 0$, then w^* is the deterministic minimizer of $L(\mathbf{w}, \mathbf{X})$. If $\lambda = 1$, then w^* is the minimizer of the unrelated random loss, $L(\mathbf{w}, \mathbf{Z})$. By choosing an intermediate value of λ, one can obtain w^* to be a random allocation such that, with a high probability, $L(\mathbf{w}^*, \mathbf{X})$, is close to the infimum loss.

For example, Morgan and Rubin [11] discusses the case in which the loss function is the Mahalanobis distance between the covariates of interest in each group. In order to define this distance, let $\mathbf{A}$ be an arbitrary matrix in $\mathbb{R}^{n \times m}$. Furthermore, define $\mathbf{A}^* := \mathbf{AL}$, where $\mathbf{L}$ is the Cholesky decomposition [3] of $\text{Cov}(\mathbf{A})^{-1} = \mathbf{L}^t\mathbf{L}$. For an allocation $\mathbf{w}$, let $\overline{\mathbf{A}^*}_1$ and $\overline{\mathbf{A}^*}_0$ denote the averages of each column of $\mathbf{A}^*$ over individuals allocated to, respectively, groups 1 and 0. That is,

$$\overline{\mathbf{A}^*}_1 := \frac{\mathbf{w}}{\mathbf{1} \cdot \mathbf{w}^t}\mathbf{A}^* \quad \text{and} \quad \overline{\mathbf{A}^*}_1 := \frac{(\mathbf{1} - \mathbf{w})}{\mathbf{1} \cdot (\mathbf{1} - \mathbf{w})^t}\mathbf{A}^* \ . \tag{1}$$

The Mahalanobis distance between the average of the column values of $\mathbf{A}$ in each group defined by $\mathbf{w}$ is defined as:

$$M(\mathbf{w}, \mathbf{A}) := m^{-1}\|\overline{\mathbf{A}^*}_1 - \overline{\mathbf{A}^*}_0\|_2^2 \ . \tag{2}$$

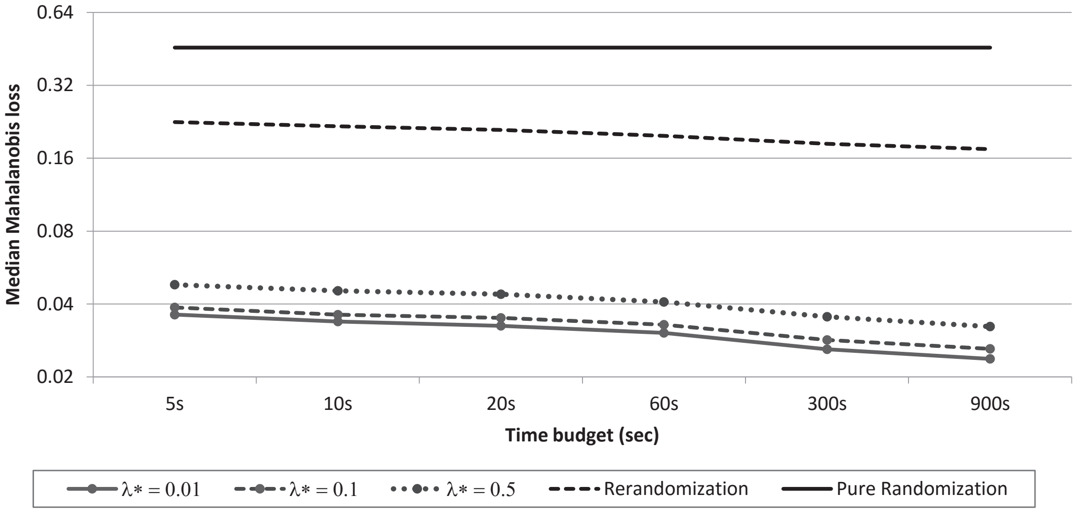

FIGURE 1. Median Mahalanobis loss for each allocation method and time budget

Morgan and Rubin [11] takes the loss function to be $M(\mathbf{w}, \mathbf{X})$.

Commonly, one wishes to obtain an allocation with a fixed number of individuals assigned to each group. That is, there exist integers n_1 and n_0 such that $n_1 + n_0 = n$, $\mathbf{1} \cdot \mathbf{w}^t = n_1$ and $\mathbf{1} \cdot (\mathbf{1} - \mathbf{w})^t = n_0$. One can take these restrictions into consideration by taking the haphazard allocation with the Mahalanobis distance as the solution to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & (1-\lambda)\, M(\mathbf{w},\mathbf{X}) + \lambda\, M(\mathbf{w},\mathbf{Z}) \\ \text{subject to} & \mathbf{1}\cdot\mathbf{w}^t = n_1 \\ & \mathbf{1}\cdot(\mathbf{1}-\mathbf{w})^t = n_0 \\ & \mathbf{w} \in \{0,1\}^n \end{array} \tag{3}$$

The description above is a mixed-integer quadratic programming problem (MIQP) [13, 9, 10, 18] and can be solved by the use of standard optimization software. Instead of directly solving the problem defined by Equation 3, one can approximate this solution through a mixed-integer linear programming problem (MILP). This is highly desirable, since a quadratic programming problem is computationally much more expensive than a linear programming problem.

Ward and Wendell [16] define a surrogate loss function that approximates $M(\mathbf{A}, \mathbf{w})$, as a linear combination of the norms l_1 and l_∞.

$$H(\mathbf{w},\mathbf{A}) := m^{-1}\left(\|\overline{\mathbf{A}^*}_1 - \overline{\mathbf{A}^*}_0\|_1 + \sqrt{m}\ \ \|\overline{\mathbf{A}^*}_1 - \overline{\mathbf{A}^*}_0\|_\infty\right) \tag{4}$$

The minimization of this *hybrid* norm yields a mixed-integer linear programming problem, see Murtagh [13] and Wolsey and Nemhauser [18]

$$\begin{array}{ll} \text{minimize} & (1-\lambda)\, H(\mathbf{w},\mathbf{X}) + \lambda\, H(\mathbf{w},\mathbf{Z}) \\ \text{subject to} & \mathbf{1}\cdot\mathbf{w}^t = n_1 \\ & \mathbf{1}\cdot(\mathbf{1}-\mathbf{w})^t = n_0 \\ & \mathbf{w} \in \{0,1\}^n \end{array} \tag{5}$$

NUMERICAL EXPERIMENTS

In this section we describe a numerical experiment conducted in order to evaluate the performance of the haphazard allocation method as defined in problem 5 vis-à-vis the rerandomization method, as described by Morgan and Rubin [11, 12]. Our empirical analysis was based on the dataset described at Shadish et al. [15], the same dataset used in the paper of Morgan and Rubin [12]. This dataset describes volunteer students from introductory psychology classes at a public university, that were assigned randomly to two experimental groups of cardinality $n_0 = 210$ and $n_1 = 235$.

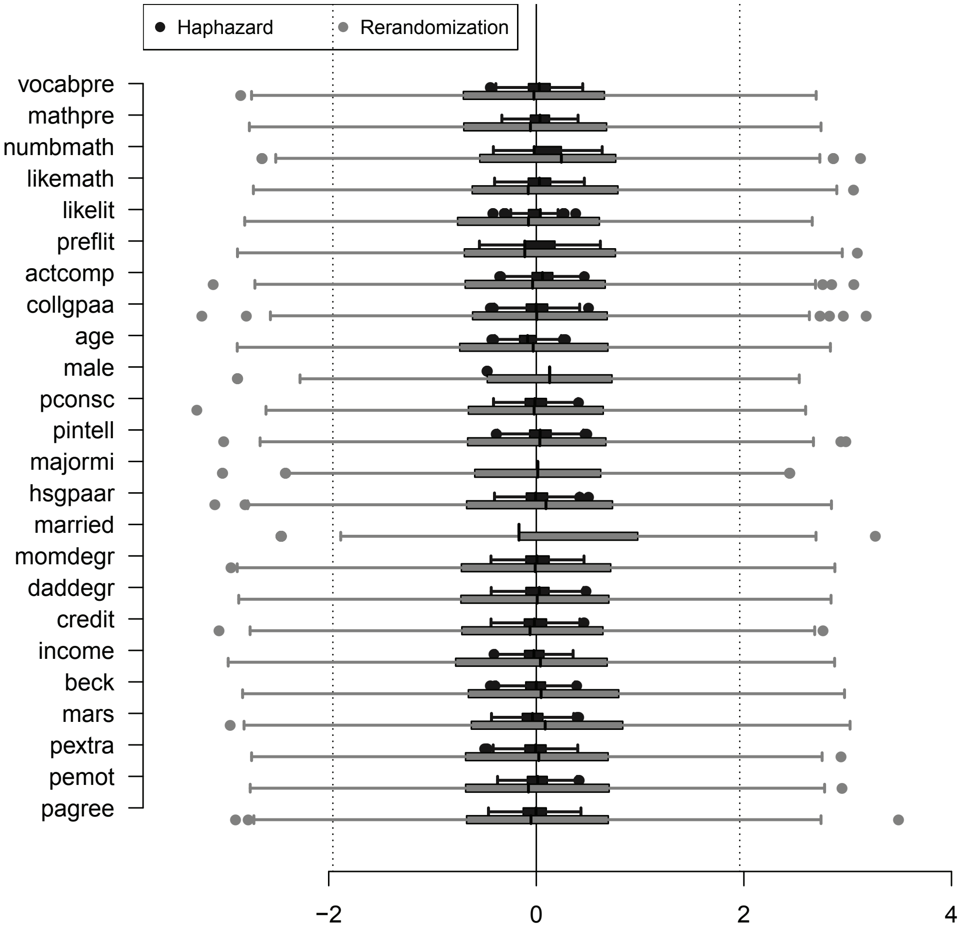

FIGURE 2. Difference between groups 0 and 1 with respect to average of standardized covariate values for each type of allocation. Time budget is 300s/allocation.

The dataset covariates relate to demographic characteristics, previous knowledge in vocabulary and mathematics, personality, mathematics anxiety, depression level and posttest scores on vocabulary and mathematics. This same dataset had also been used by Morgan and Rubin [12] for the empirical analysis of rerandomization method. From the 31 original covariates, we kept the maximum subset (24 covariates) with a non-singular covariance matrix and Cholesky factors.

In our empirical study, we explore the trade-off between randomization and optimization by using well calibrated values for the parameter λ, as defined in the next equation. The transformation between parameters λ and λ^* is devised to equilibrate the weights given to the terms of Equation 5 corresponding to the covariates of interest and artificial, which have distinct dimensions, d and k.

$$\lambda = \lambda^* \, / \left[\lambda^*(1 - k/d) + k/d\right] \, , \quad \text{where} \;\; \lambda^* \in \{0.01, 0.1, 0.5\}. \tag{6}$$

The performance of the haphazard randomization method was evaluated on a bi-dimensional grid were the parameter λ took the values defined by Equation 6, and processing time budget was set to $5, 10, 20, 60, 300$ and 900 seconds. For each point of this grid, the haphazard allocation method was repeated 500 times (each time with a fresh random matrix of artificial covariates, $\mathbf{Z}$). For comparison, we drew 500 allocations using the rerandomization method, in the slightly modified *fixed-time* version, that chooses the best allocation obtained with a given processing time budget. Finally, as a benchmark, we also drew 500 allocations using the standard (pure) randomization.

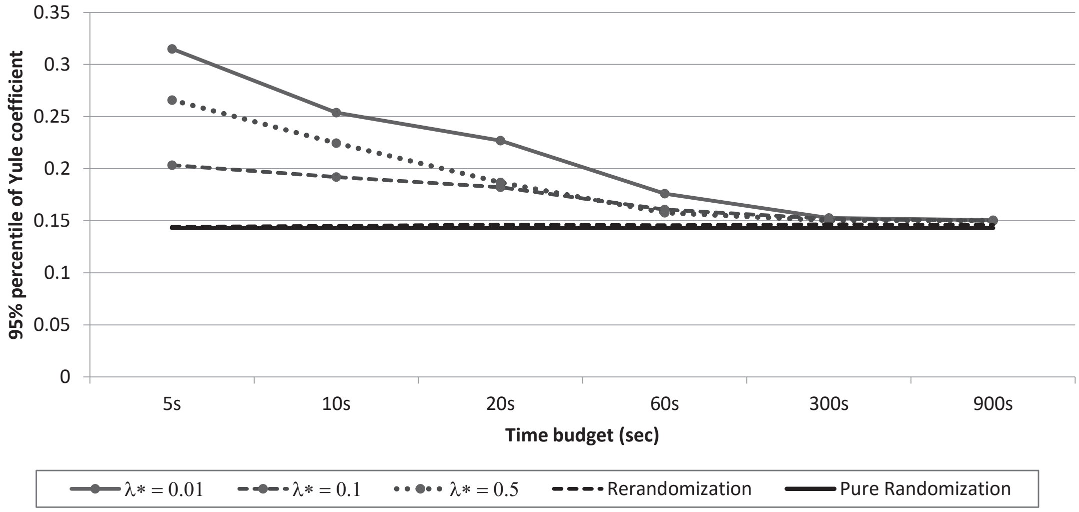

FIGURE 3. 95% percentile of the Yule correlation between allocations for each allocation method and time budget

Computational tests were conducted on a desktop computer with a processor Intel I7-4930K (3.4Ghz, 6 cores, 2 threads/core), Motherboard ASUS P9X79 LE, 24Gb RAM DDR3 and Linux Ubuntu Desktop v.14.04. The MILP problems were solved using Gurobi v.6.5.2 [5], a high performance solver that allows us to easily control all parameters of interest. Each allocation problem – among the batch of 500 allocations per allocation method, time budget and λ value – was distributed to one of the 12 logical cores available.

Figure 1 presents the median of the Mahalanobis loss function for the allocations obtained by the haphazard, rerandomization and pure randomization methods. For haphazard and rerandomization methods, the larger the time budget, the smaller the median value of the loss function. However, not only the absolute medians of the loss function yielded by haphazard allocations are much smaller (by an average factor 5.9), but also their decrease rate with time budget is considerably higher: the median loss in haphazard method decreases 33% from time budgets $5s$ to $900s$, against 23% in rerandomization. In the haphazard allocation, the smaller the value of λ, the less noise is added to the optimization problem and, therefore, the smaller the median value of the loss function. Choosing, e.g., $\lambda^* = 0.1$, haphazard allocation obtains a median loss that is almost $1/7$ of the one that is obtained using rerandomization, and $1/17$ of that obtained using pure randomization.

Figure 2 illustrates the difference in covariate balance between haphazard ($\lambda^* = 0.1$) and fixed-time rerandomization allocations ($900s$ for both methods). It can be easily seen that standardized differences on covariates between groups 0 and 1 are remarkably smaller in haphazard allocations than in rerandomization method that, in turn, are remarkably smaller than using pure randomization.

Figure 3 presents the 95% percentile of the Yule coefficient between pairs of observations. For each pair of individuals, for $(i, j) \in \{0, 1\}^2$, let z_{ij} denote the number of allocations such that the first individual is assigned to group i and the second individual is assigned to group j. The Yule coefficient for pairs of observation is computed as

$$Y = (z_{00}z_{11} - z_{01}z_{10}) \,/\, (z_{00}z_{11} + z_{01}z_{10}). \tag{7}$$

This coefficient ranges in the interval $[-1, 1]$ and measures how often the individuals under consideration are allocated to the same or to different groups. It equals zero when the numbers of agreement and disagreement pairs is equal; and is maximum (-1 or $+1$) in the presence of total negative (complete disagreement) or positive (complete agreement) association. The Yule coefficient for each pair of individuals was computed over the 500 allocations, and Figure 3 displays the 95% percentile taken over all pairs of individuals. The pure randomization method provides the lowest benchmark for Yule coefficient. As expected, the fixed-time rerandomization method attains the next lowest value of the Yule coefficient. Indeed, it attains almost the same coefficient as simple random allocation. In the grid set in this study, the haphazard allocations method was less sensitive to the choice of λ, and the major role was played by

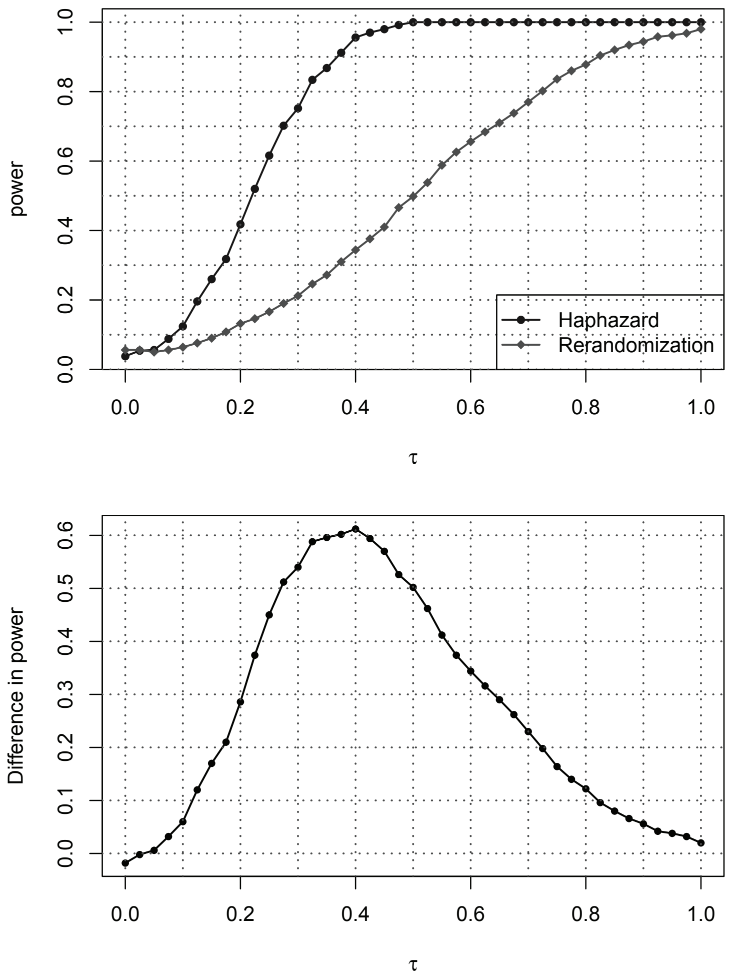

FIGURE 4. Power curves for each allocation method for testing $\tau = 0$ using a rerandomization test. Difference between haphazard and rerandomization allocations with respect to the power curves for testing $\tau = 0$ using a rerandomization test.

the time budget. The importance of the time budget parameter can be intuitively understood as follows: A large time budget implies a very precise solution that, in turn, is highly sensitive even to small perturbations. For the maximum time budget on the experimental grid, 900 seconds, haphazard allocation obtains a Yule coefficient of 0.15, only 3% higher than fixed-time rerandomization that, in turn, is 2% higher that pure randomization.

Although the above measures are relevant, they can also be seen as a proxy for optimizing other statistical properties. For instance, one might be interested in testing the existence of a causal effect of the group assignment on a given response variable. For example, consider that, for each $j \in \{0, 1\}$, Y^j is a vector of observations of a response variable when all individuals are assigned to group j. Assume that $Y_i^0 = \epsilon_i + \sum_j (X_{i,j} - \bar{X}_{,j}) / \text{Var}(X_{,j})$, where ϵ are independent standard normals. Also, $Y_i^1 = Y_i^0 + \tau$. Figures 4 illustrates the difference of power in the allocations obtained by the haphazard and the rerandomization methods for a rerandomization test for the hypothesis $\tau = 0$. The tests obtained using the haphazard allocations are uniformly more powerful over τ than the ones obtained using the rerandomization allocations. Figure 4 shows that the difference in power between these allocation methods can be as high as 0.6 (at $\tau = 0.4$).

FINAL COMMENTS

Results presented in this paper indicate that the haphazard intentional allocation method is a promising tool for design of experiments. In numerical experiments performed on a real dataset used for a two-arms study, the haphazard allocations method outperformed the alternative fixed-time rerandomization method by a factor 6.7 concerning the loss function of imbalance between the allocated groups. At the same time, measures of association related to possible systematic bias in non-random allocation had a degradation of less than 3%. Besides, permutation tests using haphazard allocations are uniformly more powerful than those obtained using the rerandomization allocations.

In future works, we shall explore the use of the haphazard intentional allocation and rerandomization methods in the application fields of Clinical Trials and Jurimetrics. Future works shall also consider the use of alternative surrogate Loss functions for balance performance, such as CVaR norms, Deltoidal norms and Block norms [14, 4, 17].

ACKNOWLEDGMENTS

The authors are grateful for the support of IME-USP and EACH-USP, the Institute of Mathematics and Statistics and the School of Arts, Sciences and Humanities of the University of São Paulo; UFSCar, the Federal University of São Carlos; FAPESP - the State of São Paulo Research Foundation (grants CEPID 2013/07375-0 and 2014/50279-4); and CNPq - the Brazilian National Counsel of Technological and Scientific Development (grant PQ 301206/2011-2). Finally, the authors are grateful for advice and comments received from anonymous referees, from participants of MaxEnt 2016, the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, and to its organizing committee, specially Prof. Geert Verdoolaege.

REFERENCES

[1] F. V. Bonassi, R. Nishimura, R. B. Stern, P. M. Goggans, and C.-Y. Chan. In defense of randomization: A subjectivist bayesian approach. In *Aip Conference Proceedings*, volume 1193, page 32, 2009.

[2] V. Fossaluza, M. S. Lauretto, C. A. B. Pereira, and J. M. Stern. Combining optimization and randomization approaches for the design of clinical trials. In *Interdisciplinary Bayesian Statistics*, pages 173–184. Springer, 2015.

[3] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[4] J. Y. Gotoh and S. Uryasev. Two pairs of polyhedral norms versus l_p-norms: proximity and applications in optimization. *Mathematical Programming A*, 156(1):391–431, 2016.

[5] Gurobi Optimization Inc. *gurobi: Gurobi Optimizer 6.5 interface*, 2015. URL `http://www.gurobi.com`. R package version 6.5-0.

[6] J. B. Kadane and T. Seidenfeld. Randomization in a bayesian perspective. *Journal of statistical planning and inference*, 25(3):329–345, 1990.

[7] M. S. Lauretto, F. Nakano, C. A. B. Pereira, and J. M. Stern. Intentional sampling by goal optimization with decoupling by stochastic perturbation. In *AIP Conf. Proc*, volume 1490, pages 189–201, 2012.

[8] D. V. Lindley. The role of randomization in inference. In *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, pages 431–446. JSTOR, 1982.

[9] R. F. Love, J. G. Morris, and G. O. Wesolowsky. Facilities location. *Chapter*, 3:51–60, 1988.

[10] R. K. Martin. *Large scale linear and integer optimization: a unified approach*. Springer Science & Business Media, 2012.

[11] K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

[12] K. L. Morgan and D. B. Rubin. Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, 110(512):1412–1421, 2015.

[13] B. A. Murtagh. *Advanced linear programming: computation and practice*. McGraw-Hill International Book Co., 1981.

[14] K. Pavlikov and S. Uryasev. Cvar norm and applications in optimization. *Optimization Letters*, 8:1999–2020, 2014.

[15] W. R. Shadish, M. H. Clark, and P. M. Steiner. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344, 2008.

[16] J. Ward and R. Wendell. Technical note-a new norm for measuring distance which yields linear location problems. *Operations Research*, 28(3-part-ii):836–844, 1980.

[17] J. Ward and R. Wendell. Using block norms for location modeling. *Operations Research*, 33(5):1074–1090, 1985.

[18] L. A. Wolsey and G. L. Nemhauser. *Integer and combinatorial optimization*. John Wiley & Sons, 2014.