Haphazard Intentional Allocation: Case Study in Air Quality Monitoring

Julio M. Stern[⋆], **Marcelo S. Lauretto** [⋆], **Rafael B. Stern**[†], **Celma O. Ribeiro**[⋆]

[⋆] Universidade de São Paulo
[†] Universidade Federal de São Carlos

Sustainable Gas Research & Innovation (2018)

Introduction I

- In randomized experiments, a simple random allocation can yield groups that differ meaningfully with respect to a given covariate. Furthermore, it is unfeasible to control the allocation with respect to more than a moderate number of covariates.
- Morgan and Rubin (2012, 2015) propose an approach based on *Rerandomization* (repeated randomization) to ensure that the final allocation obtained is well balanced.
- Levels of the Rerandomization method:
  1. Lower level: Random samplings for obtaining proposed allocations (Guarantees stochastic behavior of proposed allocations)
  2. Upper level: Rejection of proposals that do not satisfy balance criteria ("Optimizes" balance of final allocation)
- However, despite the benefits of the Rerandomization method, it has an exponential computational cost in the number of covariates (for fixed balance constraints).

Introduction II

- 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. (2012).
- Similarly to the allocation process in Morgan and Rubin (2012), our method can be divided into a randomization and an optimization step.

  1. Randomization step: consists of creating new (artificial) covariates according to a specified distribution.
  2. Optimization step: consists of finding the allocation that (approximately) minimizes a linear combination of:
     - the imbalance in the original covariates; and
     - the imbalance in the artificial covariates.

Haphazard intentional allocation I

- Let X denote the covariates of interest.
  - X: 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, treatment or arm index, $g \in \mathcal{G} = \{0, 1, 2, \ldots\}$.
- We represent an allocation by w, a $1 \times n$ vector in $\mathcal{G}^n$.
- Our goal is to generate an allocation with a low value for a specified inbalance loss function, $L(w, X)$.
- The Haphazard Intentional Allocation consists of finding the approximate minimum of $L(w, [X, Z])$, where Z is a matrix containing random noise.

Haphazard intentional allocation II

- Let 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 Intentional Allocation finds a feasible allocation, w^* minimizing

$$\begin{aligned} w^* &= \arg\min_{w\in\mathcal{G}^n} L(\lambda, w, X, Z) \\ &= \arg\min_{w\in\mathcal{G}^n} (1-\lambda)L(w,X) + \lambda L(w,Z). \end{aligned}$$

- λ: controls the amount of perturbation that is added to the original loss function, $L(w,X)$.
  - $\lambda = 0 \Rightarrow w^* =$ deterministic minimizer of $L(w,X)$;
  - $\lambda = 1 \Rightarrow w^* =$ minimizer of the unrelated random loss, $L(w,Z)$.
  - Intermediate values of λ render intermediary characteristics.
- From now on, we consider the case of two groups, $\mathcal{G} = \{0,1\}$, and Normal distributed random variables.

Haphazard intentional allocation III

- Morgan and Rubin (2012) discusses the case in which the loss function is based on the Mahalanobis distance between the covariates of interest in each group.
- In order to define this loss function, let A be an arbitrary matrix in $\mathbb{R}^{n \times d}$. Furthermore, define $\widetilde{A} := A\, L$, where L is the lower triangular Cholesky factor: $\text{Cov}(A)^{-1} = L\, L^t$, see [1].
- For an allocation w, let a^1 and a^0 denote the averages of each column of $\widetilde{A}$ over individuals allocated to, respectively, groups 1 and 0. That is,

$$a^1 := \frac{w}{n_1}\widetilde{A} \quad \text{and} \quad a^0 := \frac{(\mathbb{1} - w)}{n_0}\widetilde{A}, \text{ where } \begin{cases} n_1 = w^t\, \mathbb{1} \\ n_0 = (\mathbb{1} - w)^t\, \mathbb{1} \end{cases}$$

- The Mahalanobis loss between the groups is computed as:

$$M(w, A) = \sqrt{n_1\, n_0 / n}\; \| a^1 - a^0 \|_2 \tag{1}$$

Haphazard intentional allocation IV

- We want to allocate a fixed number of individuals to each group, that is, $w^t \mathbb{1} = n_1$ and $(\mathbb{1} - w)^t \mathbb{1} = n_0 = n - n_1$.
- We can take all these restrictions into consideration by choosing a haphazard intentional allocation with minimal Mahalanobis loss function according to the following optimization problem:

$$\begin{aligned} \text{minimize}(w) \quad & M(\lambda, w, X, Z) \\ & = \lambda M(w, Z) + (1-\lambda) M(w, X) \\ \text{such that} \quad & w^t \mathbb{1} = n_1 \\ & w \in \{0,1\}^n \end{aligned} \tag{2}$$

- This is a mixed-integer *Quadratic Programming* problem, that is difficult to solve relative to the mixed-integer *Linear Programming*.

Haphazard intentional allocation V

- Hence, we use the following *Linear Programming* approximation, based on the *hybrid norm*:

$$H(w, A) = \|a^1 - a^0\|_1 \; + \; \sqrt{d}\|a^1 - a^0\|_\infty .$$

  The hybrid norm is a surrogate loss function for the quadratic norm, based on the extreme cases of the L_p norms for $p = 1$ and $p = \infty$, see [12].

- Furthermore, the resulting optimization problem has the form of Linear Programming:

$$\begin{aligned}
\text{minimize}(w) \quad & H(\lambda, w, X, Z) \\
& = \lambda H(w, Z) + (1 - \lambda) H(w, X) \\
\text{such that} \quad & w^t \, \mathbb{1} = n_1 \\
& w \in \{0, 1\}^n
\end{aligned} \tag{3}$$

Case Study I

- We consider the problem of selecting air quality monitoring stations in the State of Sao Paulo
- Problem: given 54 candidate stations, select $n_1 = 20$ stations for installation of additional pollutant sensors
- Station variables:
  - Medians of one-year atmospheric & pollutant indicators
    Weight: 70%
    - Rainy and dry seasons
  - Geolocation (latitude / longitude)
    Weight: 30%

Case Study II

Table 1a: Stations

Abbrev	Name	Abbrev	Name
AMERIC	Americana	CONGON	Congonhas
ARACAT	Aracatuba	GRU-PI	Guarulhos-Pimentas
ARARAQ	Araraquara	GRU-PM	Guarulhos-Paco Municipal
BAURU	Bauru	GUARAT	Guaratingueta
CAPRED	Capao Redondo	IBIRAP	Ibirapuera
CARAPI	Carapicuiba	INTERL	Interlagos
CATAND	Catanduva	ITAIMP	Itaim Paulista
CB-CEN	Cubatao-Centro	JACAR	Jacarei
CB-VMO	Cubatao-Vale do Mogi	JAU	Jau
CB-VPA	Cubatao-V.Parisi	JUNDIA	Jundiai
CERQCE	Cerqueira Cesar	LIMEIR	Limeira
CM-TAQ	Campinas-Taquaral	MARIL	Marilia
CM-VUN	Campinas-V.Uniao	MAUA	Maua

Case Study III

Table 1b: Stations

Abbrev	Name	Abbrev	Name
MGCRUZ	Mogi das Cruzes	SAN-PP	Santos-Ponta da Praia
MOOCA	Mooca	SANTOS	Santos
MT-REM	Marg.Tiete-Pte Remedios	SB-CEN	S.Bernardo-Centro
OSASCO	Osasco	SCAETA	Sao Caetano do Sul
PARELH	Parelheiros	SJCAMP	S.Jose Campos
PAULIN	Paulinia	SJC-JS	S.Jose Campos-Jd.Satelite
PAUL-S	Paulinia-Sul	SJC-VV	S.Jose Campos-Vista Verde
PINHEI	Pinheiros	SJRPRE	Sao Jose do Rio Preto
PIRACI	Piracicaba	SOROC	Sorocaba
PJARAG	Pico do Jaragua	STGERT	Santa Gertrudes
PQDPED	Parque D.Pedro II	TABSER	Taboao da Serra
PRESPR	Presidente Prudente	TATUI	Tatui
RP-CEN	Ribeirao Preto-Centro	TAUBAT	Taubate
SA-CAP	S.Andre-Capuava	USP	Cid.Universitaria-USP-Ipen

Case Study IV

Table 2: Station-related variables

Code	Parameter Description
MP10	Partículas Inaláveis
NO	Monóxido de Nitrogênio
NO2	Dióxido de Nitrogênio
NOx	Óxidos de Nitrogênio
O3	Ozônio
TEMP	Temperatura do Ar
UR	Umidade Relativa do Ar
VV	Velocidade do Vento
LAT	Latitude
LON	Longitude

Case Study V

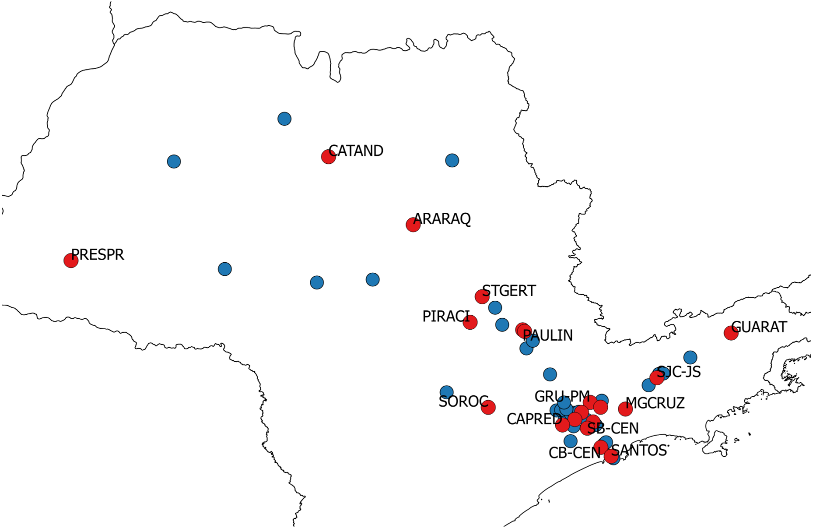

Figure 1a. Selected (red) and unselected (blue) stations

Case Study VI

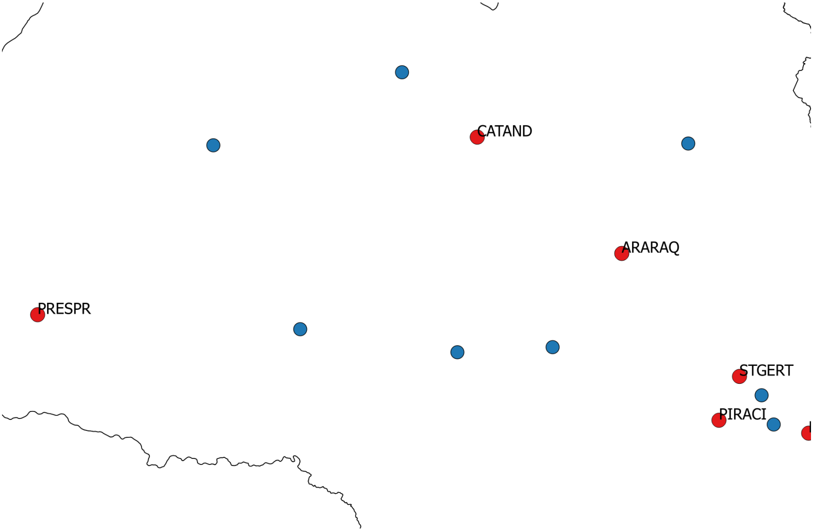

Figure 1b. Selected (red) and unselected (blue) stations

Case Study VII

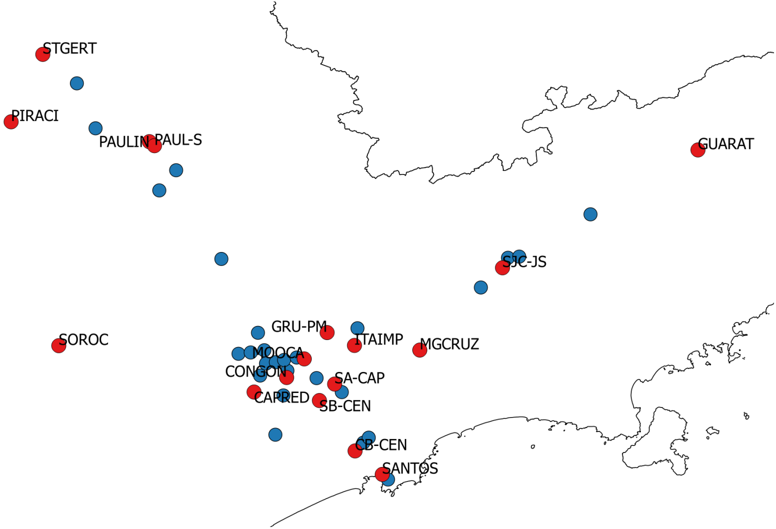

Figure 1c. Selected (red) and unselected (blue) stations

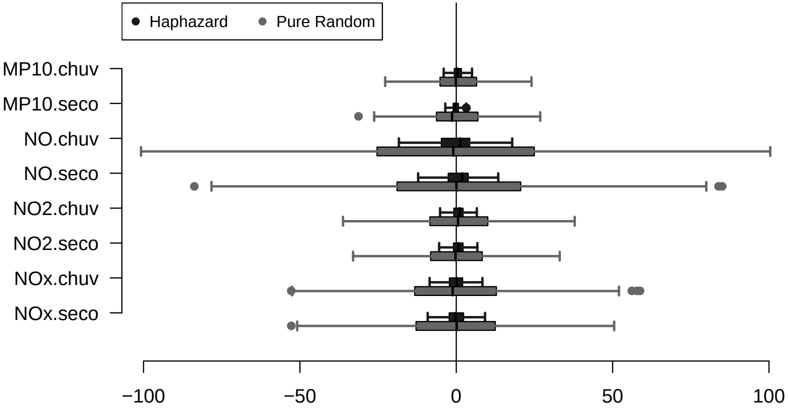

Figure 1a. Percentual differences between groups in each covariate (200 allocations).

Case Study IX

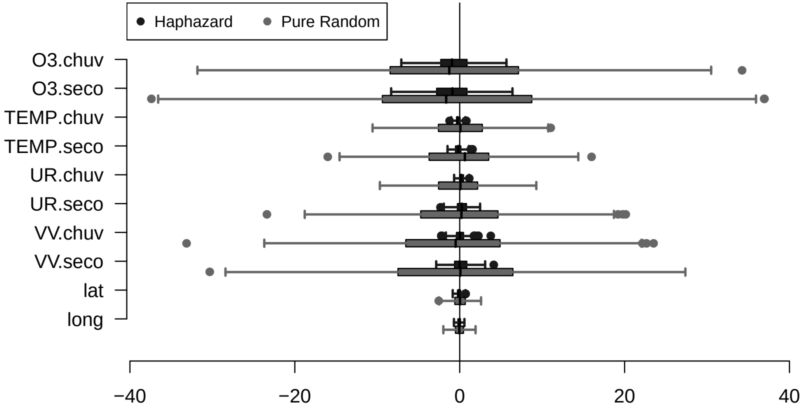

Figure 1b. Percentual differences between groups in each covariate (200 allocations).

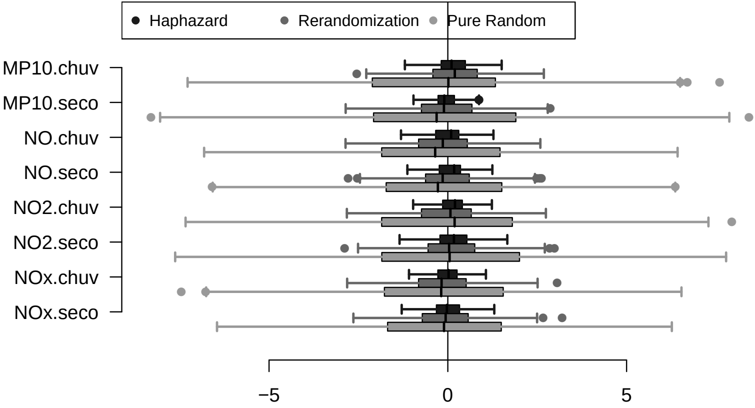

Figure 1c. Percentual differences between groups in each covariate (200 allocations).

Case Study XI

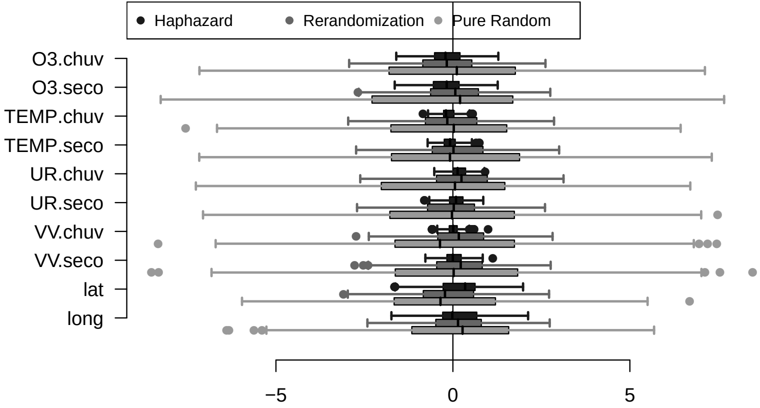

Figure 1d. Percentual differences between groups in each covariate (200 allocations).

Case Study XII

- An alternative interpretation for our experiments is to see them as a proxy for other relevant 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. Ex:
  - For each $j \in \{0, 1\}$, we simulate Y^j as the response variable when all individuals are assigned to group j.
  - We follow the procedure:
    1. $Y_i^0 = \epsilon_i + \sum_j \frac{X_{i,j} - \overline{X}_{\bullet,j}}{\text{std}(X_{\bullet,j})}$, where $\epsilon \sim N(0, \mathbb{I})$.
    2. $Y_i^1 = Y_i^0 + \tau$.

Case Study XIII

- Figure 2 illustrates the difference of power in the allocations obtained by the Haphazard, Rerandomization and Pure Randomization procedures for a permutation test for the hypothesis $\tau = 0$.
- The tests obtained using the Haphazard Intentional Allocation method are uniformly more powerful over τ than the ones obtained using the Rerandomization and Pure Randomization methods.

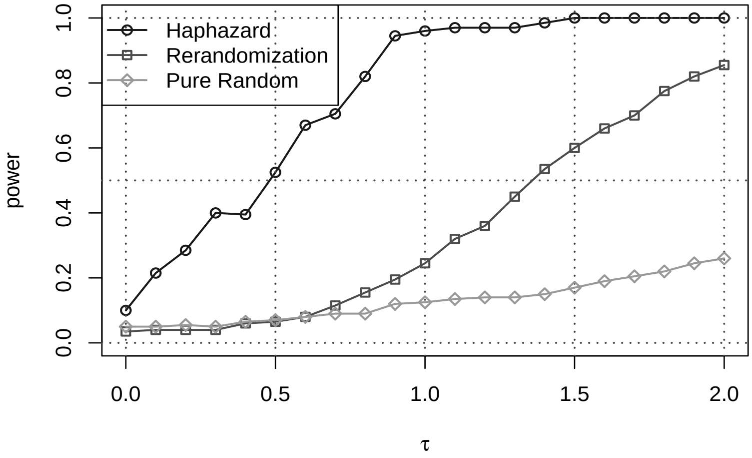

Figure 2. Power curves for each allocation procedure for testing $\tau = 0$ using a permutation test.

Future Research

- Explore the use of the Haphazard Intentional Allocation method and the Rerandomization method in applied problems in the field of:
  - Clinical trials;
  - Jurimetrics.
- Explore the use of alternative surrogate Loss functions for balance performance, such as CVaR norms, Deltoidal norms and Block norms [10, 2, 13].

References I

[1] G.H. Golub and C.F. Van Loan. Matrix Computations. JHU Press, 2012.

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

[3] Gurobi Optimization Inc. gurobi: Gurobi Optimizer 6.5 interface. URL http://www.gurobi.com, 2015.

[4] M. S. Lauretto, F. Nakano, C. A. B. Pereira, J. M. Stern Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation. *AIP Conference Proceedings, 1490, 189-201*, 2012.

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

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

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

References II

[8] K. L. Morgan and D. B. Rubin. Rerandomization to Balance Tiers of Covariates. *JASA*, 110, 512, 1412–1421, 2015.

[9] B. A. Murtag. Advanced linear programming: computation and practice. McGraw-Hill International Book, 1981

[10] K. Pavlikov and S. Uryasev. CVaR norm and applications in optimization. Optimization Letters 8, 1999–2020, 2014.

[11] W. R. Shadish, M. R. Clark, 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.

[12] J. E. Ward and R. E. Wendell. Technical Note-A New Norm for Measuring Distance Which Yields Linear Location Problems. *Operations Research* 28 (3-part-ii), 836–844, 1980.

[13] J. E. Ward and R. E. Wendell. Using Block Norms for Location Modeling. *Operations Research*, 33(5), 1074–1090, 1985.