This paper presents a new approach for designing experiments in which multiple potential clusterings (partitions) of units are considered—each cluster is called a Supergeo—and treatment and control groups are formed by dividing Supergeos in a way that minimizes covariate imbalance, where this optimization is also taken across all considered ways for forming Supergeos.

Overall, I think the idea is pretty interesting. However, I have serious concerns that there is an insufficient of detail and rigor contained in this paper and the efficacy of the proposed method is unclear. I am unsure that these concerns can be addressed with a substantial revision. Thus, I am recommending a rejection from the *Journal of Causal Inference*.

Here are the largest of my concerns:

1. One obvious concern is in the randomization. Once the Supergeos are partitioned into two groups, I can only really think about two possible randomizations—where the first group receives treatment and the second group receives control, and vice versa. So ultimately, the experiment produces two experimental units in total, thereby making the experiment severely underpowered.

   I will admit, this type of issue is not uncommon in the optimal design of experiments, and it seems reasonable that this approach will work heuristically provided that the covariates used in the design are extremely informative. But this approach theoretically cannot protect against the case where there is some important *unmeasured* confounder that is highly associated with the group membership—randomization in general will prevent imbalances on both observable and unobservable covariates. At the very least, this concern needs to be addressed somewhere in this paper.

   I would suggest that the author look into some recent work on Gram Schmidt Random Walks (GSRW) [Harshaw et al., 2024]—which considers a similar objective as this paper—on how to incorporate additional randomness in the treatment assignment. I imagine that a paper which extends GSRWs by incorporating the Supergeo construction would be quite interesting to read.

2. Sections 6–9 overall seem quite incomplete and provide inconclusive evidence (at best) of the efficacy of your method.

   (a) Most concerning is that, aside from a somewhat substantial savings in computational cost, it is not clear from your results that your method is more effective than other methods. In particular, **Table 1, Figure 1, and Figure 2 all show your proposed method to have the largest absolute bias and worst RMSE** relative to your competitors, which does not match what is written in the text. I'm not sure whether this is due to incorrect labeling of the graph or just that it performs poorly, but this issue should be addressed prior to submitting this paper for publication elsewhere.

   (b) The "Mean Balance" term in Table 1 needs to be defined. I'm not sure what that number means or how it's computed.

   (c) I am not sure what the Std Power column is in Table 3. It should not be possible to achieve a standard deviation greater than 0.5 for a variable valued between 0 and 1.

(d) Experiments discussed in Sections 9.8–9.13 seem not to be performed.

3. There is a lack of formalization in the paper, and a lot of results make a lot of unstated assumptions. For example, while the author introduces potential outcomes in Section 3.3.1, the assumption that responses follow the Neyman-Rubin potential outcomes model is not stated:

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$$

However, this lack of formalization makes the proof in Appendix A extremely hard to follow.

(a) The proof of the $(1+\epsilon)$ result in Appendix A seems to assume that the entries in the adjacency matrix representing the network of geos are randomly generated Bernoulli random variables, but this is not made explicit. Additionally, there needs to be additional context to how the observed geo graph could be thought of as an instance of this random adjacency graph. For example, when introducing geos, the paper only states that "E is a weighted adjacency matrix representing inter-geo relationships" and does not give criteria for when an "edge" is formed—I had originally interpreted the setup to mean that the graph of geos to be complete. Additionally, without this formalization, it is not clear, for example, how the Hoeffding bound in Lemma 1 is applied to the "entries in the Adjacency Matrix."

(b) I am having a hard time verifying some of the results, particularly Lemma 2 and Lemma 3. I have to imagine there is some assumption about well-behavedness of the X_i covariates in order for HAC to correctly recover clusters. But it is not clear where to confirm this result solely from the citation—stating Theorem 3 of 24 is not clear enough.

(c) It seems like the results given are asymptotic, but it is unclear what "N going to infinity" means in the context of this study. Oftentimes, when working under the Neyman-Rubin potential outcomes framework, this often implies a sequence of samples of increasing size from an infinite super-population, and I think something similar also needs to hold under whatever random graph model is being assumed in this paper. Perhaps you may be able to get around this issue by finding a sufficiently large N that is a function of the ϵ given in the proof, and starting the proof with "suppose we have $N > f(\epsilon)$ geos..."

4. The optimization problem given on Page 7 could be made more clear.

(a) I would find it significantly helpful if there was more clarity in distinguishing between "baseline outcomes" and responses observed from the experiment—perhaps by incorporating additional notation? Additionally, it is not clear what data is being used to find a baseline outcome—I imagine that it's some kind of historical data—and in this case, some caveat needs to be stated that the prediction of the baseline outcome will only be effective if past performance is indicative of future performance.

(b) Honestly, it may be clearer if you were to use symbols like $\hat{f}(y|x)$ and X_m instead of the words "baseline" and "modifier" in the objective. I found myself spending significant time hunting down the implied definitions of these terms.

(c) The SMD in the objective function needs to be |SMD| instead. Otherwise, the function will form groups to make the differences "as negative as possible."

5. There is a noticeable lack of citations in the text for results mentioned in this paper (for example, see comment 3b in this review).

6. On a final note, I'm not sure what is "adaptive" about this method, as treatments don't seem to change over time. Is there perhaps a more accurate modifier? Perhaps "intelligent" or "learning"?

References

Christopher Harshaw, Fredrik Sävje, Daniel A Spielman, and Peng Zhang. Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *Journal of the American Statistical Association*, 119(548):2934–2946, 2024.