Cluster Assignment and Covariate-Constrained Randomization for Spatially-Correlated Data

Patrick Iben, MS Candidate

Department of Biostatistics and Informatics
University of Colorado Anschutz Medical Campus
Colorado School of Public Health

May 7, 2019

Outline

- Background
  - CRT randomization problem
  - Malaria elimination project in Uganda
- Approach
- Methodology
  - Cluster optimization and assignment
  - Covariate-constrained randomization
- Results
- Conclusions
  - Limitations

CRT randomization problem

- Cluster randomized trials (CRTs) are commonly implemented when treatments cannot be randomized to individuals
- However, randomization of clusters often results in imbalance with respect to important confounders
- To circumvent this imbalance, researchers have applied constraints to the covariate make-up of the clusters so that imbalance is minimized, so-called covariate-based constrained randomization (CCR) Moulton [2004] Dickinson et al. [2015]
  - Constraints applied to all possible randomization schemes
  - Among those that satisfy these constraints one is chosen at random

CRT randomization problem

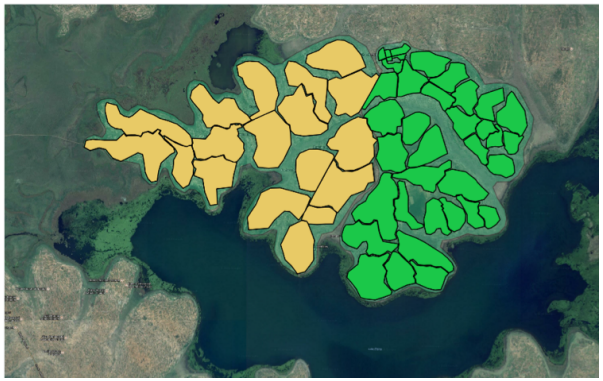
- In some disciplines, such as vector-borne disease research, spatial contamination of the intervention is often of concern, because individuals and pathogens can easily cross into regions receiving the opposite intervention arm
- This introduces the need for an additional randomization constraint

Malaria elimination project in Uganda

- In malaria research, interventions often applied to communities, not individuals
  - Spraying of insecticides, distribution of medications to health centers, or bed net campaigns
- CCR potentially beneficial method for assigning treatments to communities
  - However, spatial heterogeneity needs to be considered, as mosquitoes and people move frequently
    - Potential for treatment contamination high
- Desirable to optimize community assignment to clusters that are visually-contiguous - yet still achieve random assignment of clusters to intervention arms

Malaria elimination project in Uganda - Study area

- Yellow indicates Kapujan County
- Green indicates Toroma County
- The orange district is the location of the study site, bordering lake Bisina
- Total study population: 8,503 households, 50,178 individuals



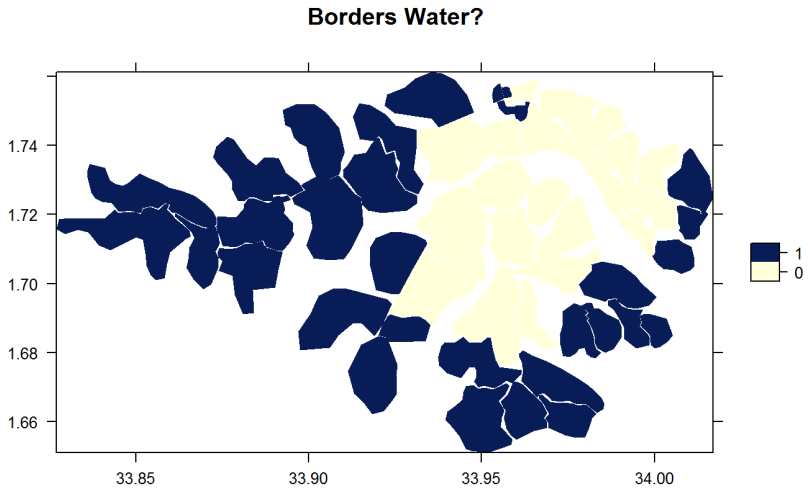
Malaria elimination project in Uganda - CRT Groups

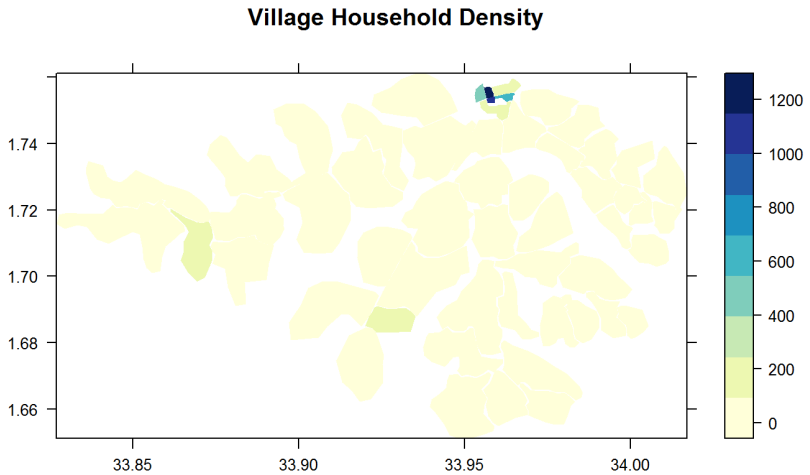
- Treatment group 1: proactive screening and treatment (ProACT) of the community by trained community members (VHT)
  - Trained VHTs will visit each household in their village weekly and evaluate and treat residents with malaria. Residents also allowed to visit the VHTs home at any time in between (intermittent community case management, iCCM)
- Treatment group 2: only passive case detection and treatment is exercised (i.e., residents are allowed to visit the VHTs home at any time - iCCM)

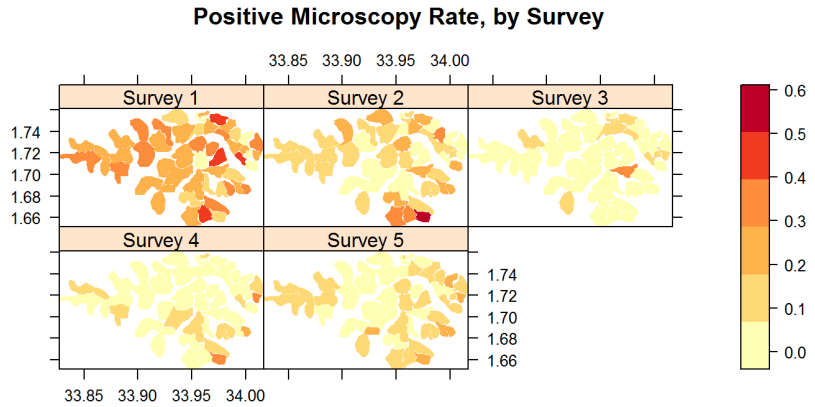
Malaria elimination project in Uganda - Problems

- Tasked with assigning villages to one of the two interventions
- Assignment of villages at random could introduce serious spatial contamination
- Imbalance of important confounders could also contaminate the intervention
  - Malaria prevalence
  - Whether or not the village borders a body of water
  - Population density

Malaria elimination project in Uganda - Covariates







Approach

- Explored and compared clustering methods for assignment of units to clusters - and clusters to treatment arms via CCR
  - Clustering the units for randomization - to mitigate the potential for geographic contamination of treatment arms before performing CCR
  - Clustering on covariates and geographical location, simultaneously, before performing CCR
  - Also explored the use of regression weights in the CCR
- Estimated the extent to which randomized treatment arm clusters mitigate geographic contamination

Approach

- Compared these methodologies using two spatial datasets
  - Aerial dataset of 55 villages in a rural part of Uganda
    - Randomize clusters of villages to intervention or control arms; mitigate spatial contamination of the interventions due to mosquito and population mixing
  - Spatial points dataset of 281 neighborhoods in a section of New York City
    - Randomize clusters of neighborhoods to intervention or control arms; mitigate spatial contamination of the interventions due to industrial pollution
- Chosen because they have mixed covariate data types (e.g. numerical, categorical), and are of very different size

Approach

- Four popular clustering algorithms used and compared
  - K-means
  - Clustering Large Applications (CLARA)
  - Ward-Like Hierarchical Clustering
  - Partitioning Around Medoids (PAM)
- Algorithms compared with respect to their:
  - Resultant degree of cluster covariate imbalance after CCR assignment (both weighted and unweighted)
  - Overall spatial heterogeneity of treatment arms after CCR
  - CPU run times for each major stage of the process - cluster optimization, weighted CCR and unweighted CCR

Approach

- R v.3.5.3 (64-Bit) was used for all data visualization, cluster assignment, and CCR.
  - "sp" and "ggplot2" packages used for spatial visualization Pebesma and Bivand [2005] Bivand et al. [2013] Wickham [2016]
- All calculations and their respective CPU run times derived from a PC with an Intel(R) Xeon(R) i7-6500U CPU E3-1280 v6 @ 3.90 GHz processor and 64.0GB of RAM, running Windows 10 Enterprise on an x64-based processor

Data Scenario 2: New York leukemia data taken from the data sets supporting Waller and A. Gotway [2004]

- Example spatial data from the "spData" package used to develop a hypothetical CRT trial in which CCR could help a larger study design - and where spatial contamination could be inherently present Bivand et al. [2019]
- Evidence that Trichloroethylene (TCE) exposure results from a leak into the ground soil near contamination sites, which is then released into ground water sources or evaporates into the air CDC [2015]
    - TCE is a volatile organic compound (VOC) that has shown some evidence of inducing testicular cancer and leukemia in rats and lymphomas and lung tumors in mice CDC [2015]

Data Scenario 2

- Simple carbon-based filters have proven effective at removing volatile compounds like TCE from water sources EWG [2018]
- In 1989, NASA released their report "Interior Landscape Plants For Indoor Air Pollution Abatement" Wolverton and Bounds [1989], detailing the total micrograms of TCE removed by a wide variety of plants, in a sealed indoor setting during a 24-hour time period
  - One plant that showed particular promise for removing TCE from the air was the Gerbera daisy
- Hypothetical study question: does providing the community homes with Gerbera daisy plants lead to a change in Leukemia rates?
  - Study group 1: carbon-based water filters provided to homes in the study area
  - Study group 2: carbon-based water filters and Gerbera daisy plants provided to homes in the study area

Methodology - Cluster optimization

- CRTs tend to have only two study arms for which to randomize clusters - only even number of clusters considered
- daisy() function in the "cluster" package used to measure covariate and geographic dissimilarity Maechler et al. [2018]
  - K-means and CLARA were used to cluster on geographic location
  - Ward-like Hierarchical and PAM were used to cluster on both geographic location and mixed data type covariates
- To avoid the ambiguous and highly-subjective process of graphically determining the optimal clustering for a given dataset, objective, numerical cluster quality statistics were optimized

Methodology - Cluster optimization - K-means, CLARA, and PAM

- Optimal clustering determined by average silhouette width, via the wcClusterQuality() function from the "WeightedCluster" package Kaufman and Rousseeuw [1990] Studer [2013]
- Average silhouette: coherence of the assignment of an observation to a given group, comparing the average weighted distance of an observation from the other members of its group and its average weighted distance from the closest group
- A value is calculated for each observation, as well as an average value for each group. If average value low, then groups not clearly separated

Methodology - Cluster optimization - Average Silhouette Width

- The authors of the "WeightedCluster" package have defined average silhouette width as the following:
  - Let k be the group of observation i, W_k the sum of the weightings of the observations belonging to group k, w_i the weight of observation i, l one of the other groups
  - a_i the average weighted distance of observation i with the other members of its group and the average weighted distance from the closest group, labeled b_i

Methodology - Cluster optimization - Average Silhouette Width

- Then, the silhouette of an observation, s_i, is computed as follows:

$$a_i = \frac{1}{W_k - 1} \sum_{j \in k} w_j d_{ij} \tag{1}$$

$$b_i = \min_l \frac{1}{W_l} \sum_{j \in l} w_j d_{ij} \tag{2}$$

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{3}$$

- The mean of all s_i observations is called the *average silhouette width*

Methodology - Cluster optimization - Ward-Like Hierarchical

- Optimal clustering determined by proportions of explained pseudo-inertia, via the "ClustGeo" package Chavent et al. [2017] Chavent et al. [2018]:
  - Consider partition $P_K = (C_1, ..., C_K)$ in K clusters; w_i the weight of the i-th observation; w_j the weight of the j-th observation; d_{ij} the dissimilarity measure between observations i and j
  - The pseudo-inertia of a cluster C_k generalizes the inertia to the case of dissimilarity data (Euclidean or not) in the following way:

$$I(C_k) = \sum_{i \in C_k} \sum_{j \in C_k} \frac{w_i w_j}{2\mu_k} d_{ij}^2 \tag{4}$$

where $\mu_k = \sum_{i \in C_k} w_i$ is the weight of C_k

Methodology - Cluster optimization - Pseudo-Inertia

- The smaller the pseudo-inertia $I(C_k)$ is, the more homogeneous are the observations belonging to the cluster C_k
- The pseudo within-cluster inertia of the partition P_K is therefore:

$$W(P_K) = \sum_{k=1}^{K} I(C_k)$$

The smaller this pseudo within-inertia $W(P_K)$ is, the more homogeneous is the partition in K clusters

Methodology - Cluster optimization - Minimum Clusters

- Only even cluster sizes greater than or equal to six were considered
- Two clusters for randomization to treatment arms - number of clusters would equal the number of treatments to be assigned, and there would be no reason to perform CCR
- Four clusters - total of $\binom{4}{2} = 6$ configurations to select randomly from for CCR. 10% cutoff leaves only one treatment allocation scheme to select from
  - Deterministic selection for cluster assignment - no longer qualifies as a CRT

Methodology - Cluster optimization - Maximum Clusters

- To mitigate spatial contamination by creating a buffer between treatment units - for all but the Ward-Like Hierarchical clustering - the maximum number of clusters was determined by the closest even value such that there was an average of at least 3 units per cluster
  - For example: if there was 55 clusters, then the maximum cluster size tested would have been 18
- The "ClustGeo" package limits the number of clusters to 55; thus, the maximum number of clusters that were tested was 54
- Logically, units were not permitted to belong to multiple clusters, and for these examples, all units needed to be assigned to a cluster

K-means

- Briefly, K-means works by first randomly assigning N data points into K disjoint subsets S_j containing N_j data points Bishop [1995]
  - Mean distance point computed for each pair of points in the set
  - Points then assigned to the cluster whose centroid is closest to that point. These two steps are alternated until within-cluster variability between data points is minimized
- K-means is restricted to only numerical data - cannot be used with mixed data types. Clusters determined solely on Euclidean distance between centroids Wol

K-means

- Often, the process of optimizing cluster size for K-means entails graphical interpretation of within sum-of-squares variation UCB
- For automation purposes, and to avoid this subjective approach, K-means was optimized with average silhouette width
- For both studies, arrays were initialized for every cluster size value tested. For the first study, between 6 and 18 clusters were tested; for the second study, between 6 and 92

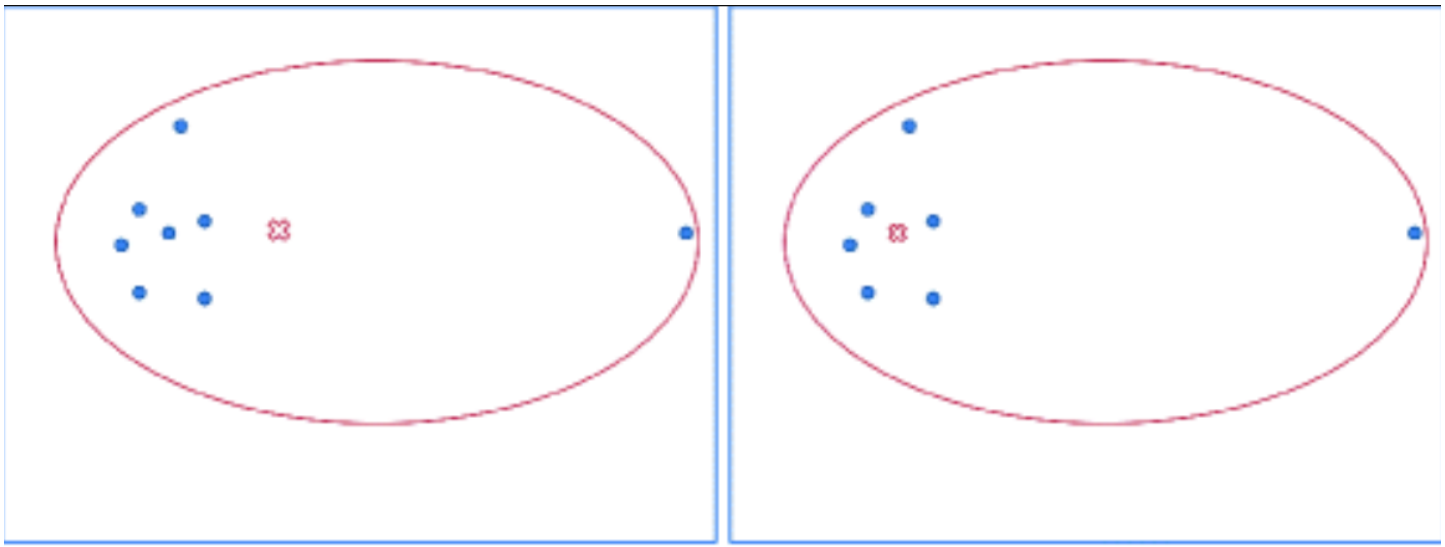
K-means

- Using the kmeans() function from the "stats" package and the wcClusterQuality() function, along with the Euclidean dissimilarity matrix, average silhouette width was iteratively collected for each cluster size with a for-loop R Core Team [2018]
  - The Hartigan and Wong [1979] algorithm, the default for the kmeans() function was used
- After performing sensitivity analysis to determine the optimal cluster size with the wcClusterQuality() function from the "WeightedCluster" package, and saving the value that optimized average silhouette width, the kmeans() clustering algorithm was then re-run with the optimal value
- Clustering scheme then extracted and merged back into the main study data. If several values equal to the optimal statistic, then minimum cluster size selected

CLARA

- CLARA (Clustering LARge Applications) is performed in two major steps Kaufman and Rousseeuw [1990]
  - First, a sample is drawn from the set of objects and clustered into K subsets using the K-medoid method, giving K representative objects Jin and Han [2010]
  - Then, each object not belonging to the sample is assigned to the nearest of the K representative objects
- K-medoid clustering method is a variant of K-means that uses an actual point in the cluster to represent the cluster, rather than a mean point as the center
  - Allows K-medoids to be less sensitive to noise and outliers than K-means

Mean vs. Medoid

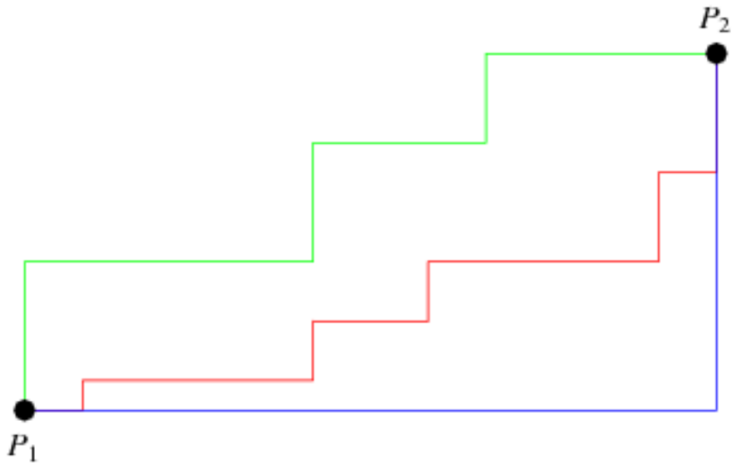


Jin and Han [2010]

CLARA

- clara() function from the "cluster" package was used
- As with K-means, average silhouette width was iteratively collected for each cluster size, and cluster size was selected on maximized average silhouette width
- Geographic dissimilarity was calculated using Manhattan distances Krause [1986]
- If several values were equal to the optimal statistic, then the minimum cluster size was selected

Manhattan Distance



Barile

Ward-Like Hierarchical

- "Clustgeo" allows users to achieve Ward-like hierarchical clustering with non-Euclidean dissimilarity measures and non-uniform weights for units
  - Allows accommodation of mixed data types (e.g., numeric, ordinal, binary), which may often be used for spatial feature data
  - Non-uniform weighting may be of use when the ultimate outcome (e.g., unit prevalence of disease) may be directly proportional to some other factor (e.g., unit population)
- To perform the clustering, hclustgeo() function used
  - Can automatically accommodate non-Euclidean dissimilarities for mixed data types
- For both studies, clusters weighted by unit population

Ward-Like Hierarchical

- Dissimilarity matrices were created for the covariates and geographical location variables separately
- In addition to accommodating non-Euclidean dissimilarities, the "Clustgeo" package allows for weighting of spatial feature and geographical dissimilarity importance during cluster assignment
    - To test the varying weightings of dissimilarity matrices for our mixed data types, Gower dissimilarity was utilized

Ward-Like Hierarchical - Gower Dissimilarity

- Gower dissimilarity uses the Dice coefficient for binary data and range-normalized Manhattan distance for numeric data Gower [1971]
- Let i and j be any two subjects in a dataset; $\delta_{ijk} = 1$ when character k can be compared for i and j, and 0 otherwise. When $\delta_{ijk} = 0$, score s_{ijk} is unknown - but is conventionally set to 0
- The similarity between subjects i and j is defined as:

$$S_{ij} = \frac{\sum_{k=1}^{v} s_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}} \tag{5}$$

Ward-Like Hierarchical - Gower Dissimilarity

- When $\delta_{ijk} = 0$ for all characters k, S_{ij} is undefined. When all comparisons are possible, $\sum_{k=1}^{v} \delta_{ijk} = v$
- Scores s_{ijk} are assigned as the following:
- **Dichotomous characters**
  Let $+$ and - be the possible values of character k, and assume no unknown values of k
  - If $k = +$ for i and $k = +$ for j, then $s_{ijk} = 1$ and $\delta_{ijk} = 1$
  - If $k = +$ for i and $k =$ - for j, then $s_{ijk} = 0$ and $\delta_{ijk} = 1$
  - If $k =$ - for i and $k = +$ for j, then $s_{ijk} = 0$ and $\delta_{ijk} = 1$
  - If $k =$ - for i and $k =$ - for j, then $s_{ijk} = 0$ and $\delta_{ijk} = 0$

Ward-Like Hierarchical - Gower Dissimilarity

- **Qualitative characters** : If i and j agree on character k, then $s_{ijk} = 1$; $s_{ijk} = 0$ if they differ
- **Quantitative characters**
  Let $x_1, x_2, ..., x_n$ be values for character k for the total sample of n cases. Then,

$$s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$$

  where R_k is the range of character k. When $x_i = x_j$, $s_{ijk} = 1$; when x_i and x_j are on opposite ends of range R_k, s_{ijk} is minimized (0 when determined from sample)

Ward-Like Hierarchical - Optimization Problem

- To obtain a new partition P_K in K clusters from a given partition P_{K+1} in $K+1$ clusters, aggregate clusters A and B of P_{K+1} according to:

$$\delta(A, B) := I(A \bigcup B) - I(A) - I(B)$$

and minimize at each step of the hierarchical clustering algorithm

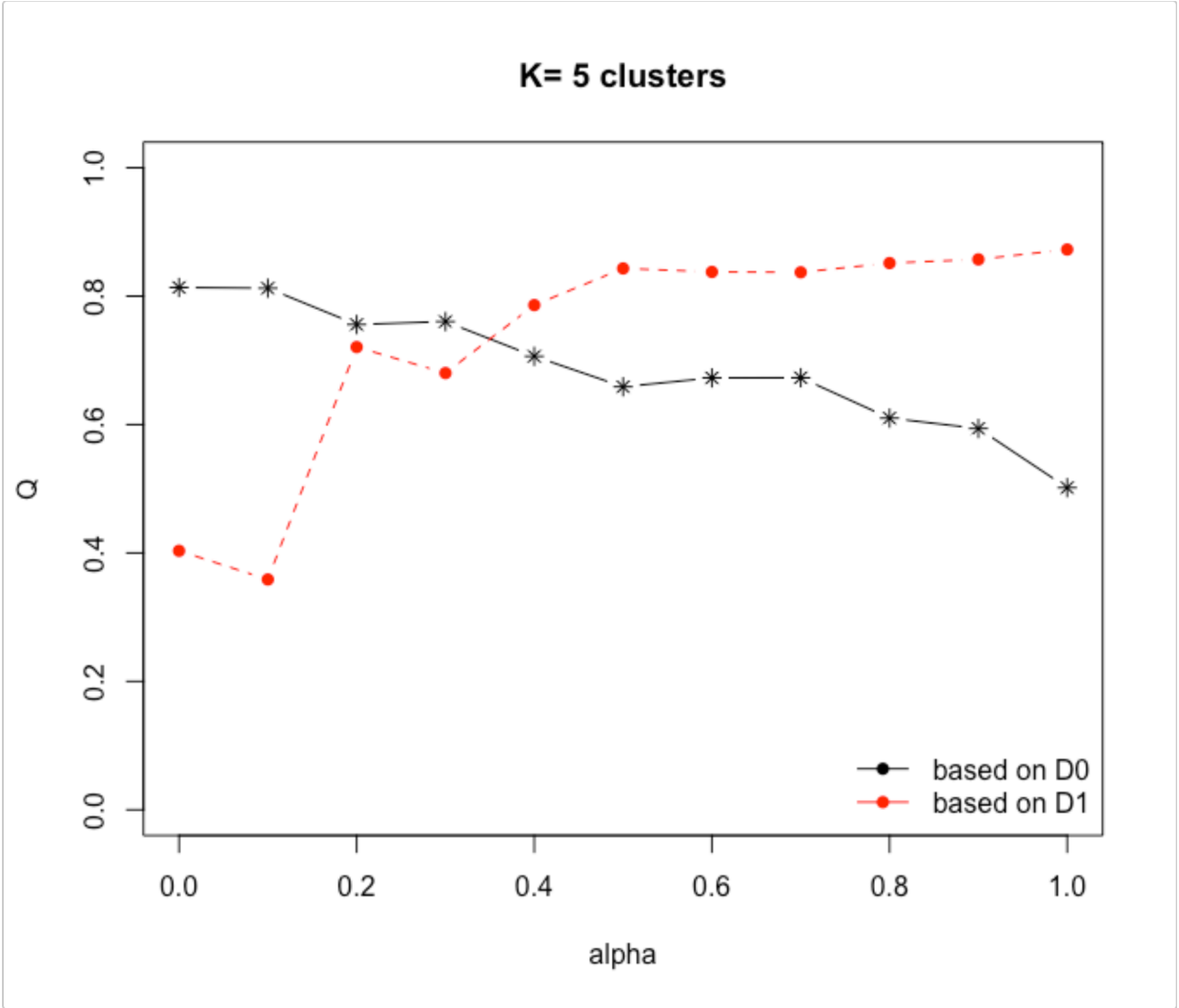
Ward-Like Hierarchical

- The Ward-like hierarchical clustering process for non-Euclidean dissimilarities is further defined by authors as the following:
  1. Step $K = n$: the initial partition P_n in n clusters is unique
  2. Step $K = n - 1, ..., 2$: obtaining the partition in K clusters from the partition in $K + 1$ clusters. At each step K, the algorithm aggregates the two clusters A and B of P_{K+1} according to the optimization problem, such that the increase of the pseudo within-cluster inertia is minimum for the selected partition over the other ones in K clusters
  3. Step $K = 1$: stop. The partition P_1 in one cluster (containing the n observations) is obtained
- The hierarchically-nested set of all partitions is defined as $P_n, ..., P_K, ..., P_1$, where the height of a cluster $C = A \bigcup B$ is $h(C) := \delta(A, B)$

Ward-Like Hierarchical

- The weighting for dissimilarity matrix importance ranges from 0 to 1, in increments of 0.1 - referred to as alpha by developers of the package. 0 represents full weighting on spatial feature data, while 1 represents full weighting on geographical location
- To determine the optimal cluster size and alpha level, the choicealpha() function from the "Clustgeo" package was used to calculate the proportion of explained pseudo-inertia for each cluster size of interest
- Optimal cluster size and alpha level determined by the value at which the difference between the explained pseudo-inertia for covariate and geographical dissimilarity was minimized
    - If several values were equal to the minimized difference, then the minimum cluster and alpha values were selected

Ward-Like Hierarchical - Optimization Example



PAM

- Like CLARA, partitioning around medoids (PAM) creates K clusters about some actual point in the center
- Like the hclustgeo() function in the "ClustGeo" package, the wcKMedoids() function from the "WeightedCluster" package allows users to perform PAM clustering with non-uniform weights
  - For both studies, unit population was again used to weight during clustering
- For use of the weights, the "PAMonce" method is utilized by the package, as defined by Reynolds et al. [2006]

PAM

- First, medoid object i and non-medoid object j are selected that produce the best clustering when their roles are switched. The objective function used is the sum of the distances from each object to the closest medoid
- Starting with an empty set of medoids, objects are then added one at a time until K medoids have been selected
- At each step, the new medoid is selected so as to minimize the objective function, and the cost of each move to a neighboring solution

CCR

- The pseudo-algorithm for CCR is as follows Moulton [2004]:
  1. Form a list of all the possible allocations. For a completely randomized (at the group level) design, there will be $\binom{2m}{m}$ entries, where $2m$ is the total number of groups
  2. Making a pass through all entries, select those allocations that meet the specified criteria. These criteria could mean achieving some level of balance on a given set of covariates
  3. Make a matrix whose elements are the number of times, from among those allocations identified in step 2, each pair is together
  4. Accept the constrained list of possibilities and go to step 5; or relax or tighten criteria and go to step 2; or change the stratification and go to step 1
  5. Randomly select one allocation from among the ones that have been selected as being acceptable in step 2

CCR

- The cvrall() function from the "cvcrand" package was used to conduct CCR, and to compare the overall balancing of subsequent cluster assignment to treatments Yu et al. [2019]
- The $l2$ balance metric was utilized with cvrall(), developed from Raab and Butcher [2001]:
- Suppose n, n_T, and n_C are the total number of clusters, the number of clusters in the treatment arm and the control arm respectively; K cluster-level variables including the continuous covariates as well as the dummy variables created from the categorical covariates; x_{ik} the k-th covariate where $k = 1, \ldots, K$ of cluster i

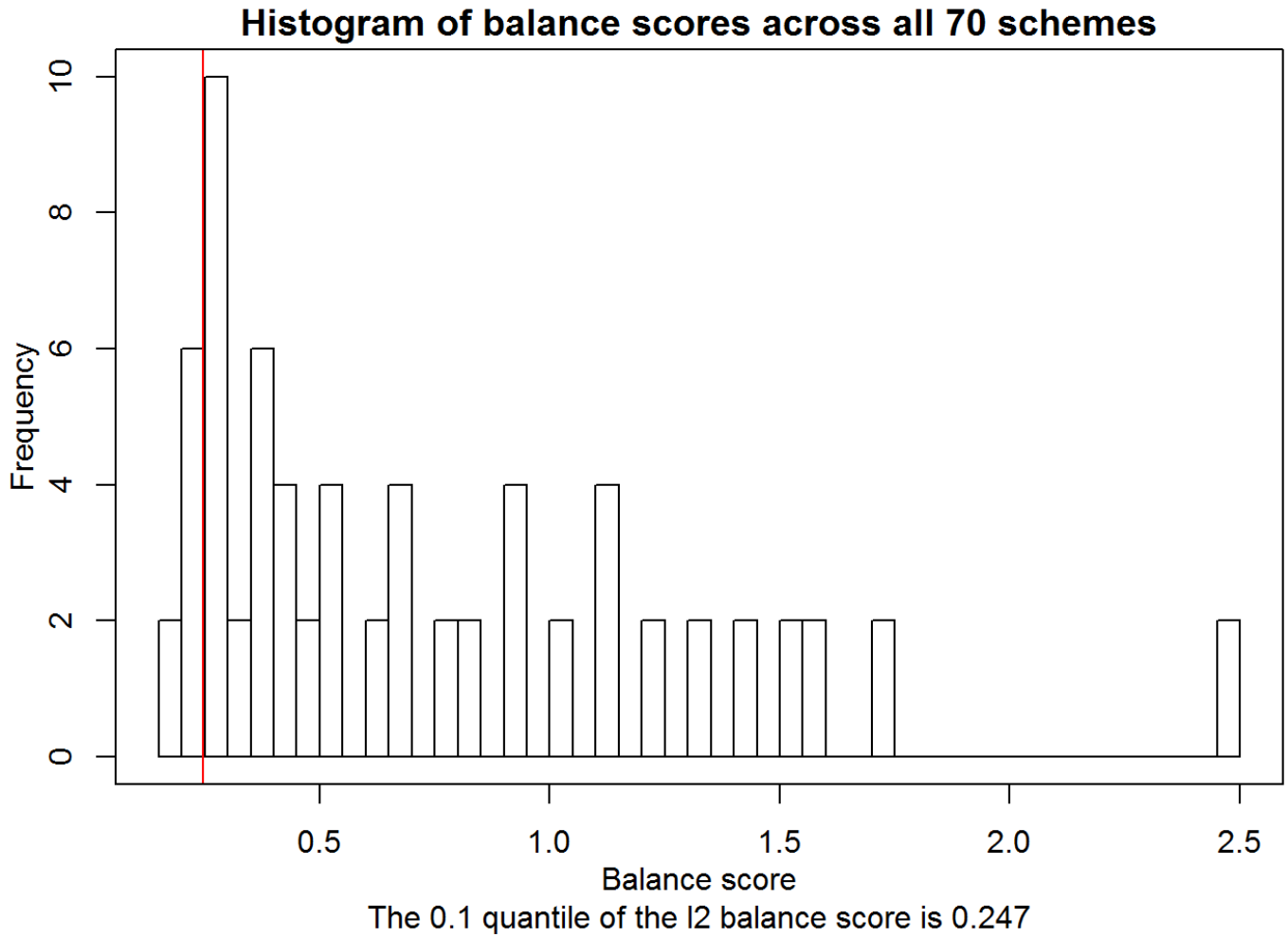
CCR

- $\bar{x}_{Tk} = \sum_{i=1}^{n_T} x_{ik}/n_T$ and $\bar{x}_{Ck} = \sum_{i=n_T+1}^{n} x_{ik}/n_C$ are the means of the k-th cluster-level variable in the treatment arm and the control arm, respectively; $\omega_k = 1/s_k^2 = \frac{n-1}{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$ with $\bar{x}_k = \sum_{i=1}^{n} x_{ik}/n$
- Then, the $l2$ balance metric is defined as:

$$B_{(l2)} = \sum_{k=1}^{K} d_k \omega_k (\bar{x}_{Tk} - \bar{x}_{Ck})^2 \tag{6}$$

where d_k is the user-defined weight for the k-th variable. By default, $d_k = 1$ for all variables. It should be noted that this is technically a measurement of covariate *imbalance*

"cvcrand" Balance Score Example



CCR - Weights

- Standardized regression coefficients from a Poisson cell-means model with an offset, via the glm() function R Core Team [2018]
- Standardized unit-level data used for modeling - scale() function
  - Regression coefficients then standardized such that their absolute values sum to 1
- The effects of weights on model compared with CCR results from un-weighted data (where $d_k = 1$ for all variables)

CCR

- Size set to 25,000,000 - if the total number of possible schemes exceeds 25,000,000, then cvrall() simulates from the complete randomization space and selects 25,000,000 unique schemes for the randomization sample space
- Only the top 10% (in terms of covariate balance) of all possible assignment schemes considered for random selection - the default cutoff for cvrall()
  - Maximum number of assignment schemes: 2,500,000

Spatial clustering of treatment arms

- Mean nearest neighbor distances (k=1) by treatment arm used to quantify - found with the nndist() function from the "spatstat" package Baddeley and Turner [2005]
  - Spatial coordinates (e.g., latitude and longitude) used to calculate mean nearest neighbor distances based on Euclidean distance
- Treatment arm mean nearest neighbor distances then compared with Welch Two-Sample Independent t-tests with unequal variance using the t.test() function from the "stats" package R Core Team [2018] Elizabeth A. Albright [2018]
- No difference in mean nearest neighbor by treatment arms $\rightarrow$ spatial homogeneity within the study space
  - May increase proximity of differing treatment arm units - bad

Results - Study 1 - CCR Balance

Method	Selected Balance Score	10% Cutoff Score
K-Means Weighted	0.248	0.294
K-Means Unweighted	3.403	3.824
CLARA Weighted	0.245	0.247
CLARA Unweighted	2.982	3.229
H-Clust Weighted	0.154	0.178
H-Clust Unweighted	1.618	1.853
PAM Weighted	0.237	0.301
PAM Unweighted	2.367	2.950

Results - Study 1 - CCR Balance

Method	N Clusters/Arm	N Allocations/Space
K-Means Weighted	6	92
K-Means Unweighted	6	92
CLARA Weighted	4	7
CLARA Unweighted	4	7
H-Clust Weighted	6	92
H-Clust Unweighted	6	92
PAM Weighted	9	4862
PAM Unweighted	9	4862

Results - Study 1 - Spatial Clustering of Treatment Arms

Method	Clustering present ($\alpha = 0.05$)?
K-Means Weighted	Yes
K-Means Unweighted	No
CLARA Weighted	No
CLARA Unweighted	No
H-Clust Weighted	Yes
H-Clust Unweighted	No
PAM Weighted	No
PAM Unweighted	No

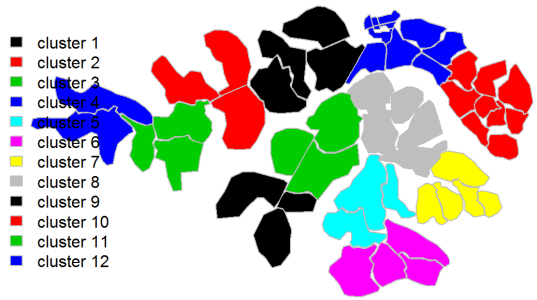
Results - Study 1 - CPU Run Times (Seconds)

Method	Cluster Optimization
K-Means	0.073
CLARA	0.073
H-Clust	5.206
PAM	0.232

Method	Weighted CCR	Unweighted CCR
K-Means	0.063	0.032
CLARA	0.016	0.016
H-Clust	0.016	0.016
PAM	0.197	0.200

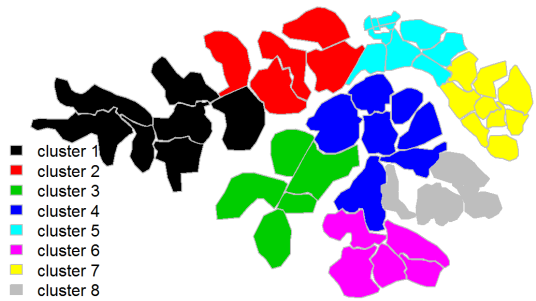
Results - Study 1 - Visualization of Clusters

12-Cluster Partition Obtained with K-Means on Centroid Euclidean Distance



Ward-Like Hierarchical Clustering of Size 12 Obtained with alpha=0.4

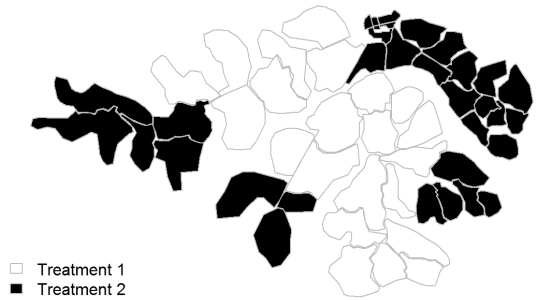
8-Cluster Partition Obtained with CLARA on Centroid Manhattan Distance



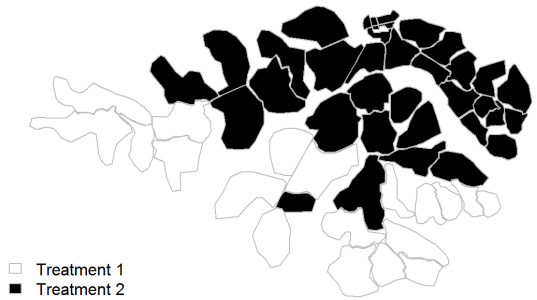
PAM Clustering of Size 18 Obtained with
alpha = 0.0 (Weighting 100% on Covariates and 0% on Geographical)

Results - Study 1 - Visualization of Treatments (Weighted)

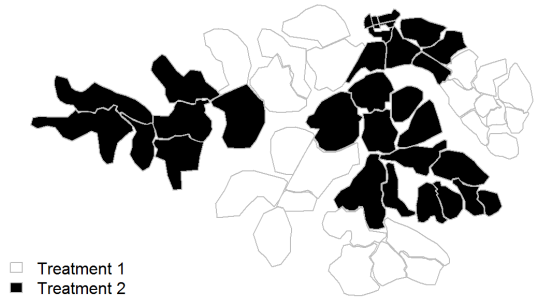
Treatment Allocations Obtained with K-Means and Weighted CCR



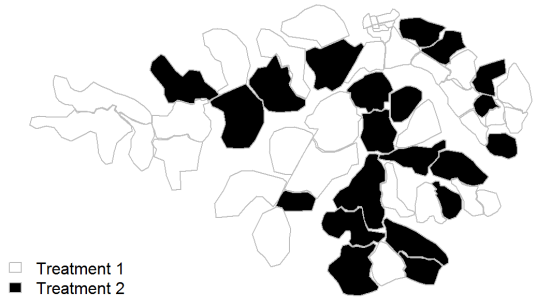
Treatment Allocations Obtained with Ward-Like Hierarchical Clustering, Weighted CCR



Treatment allocations Obtained with CLARA Clustering, Weighted CCR

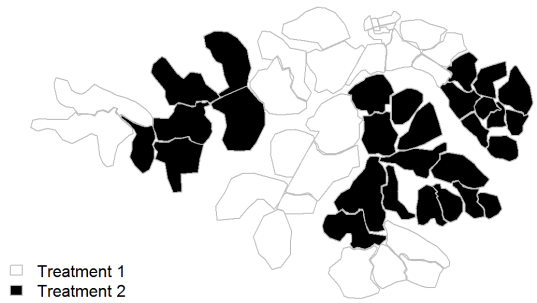


Treatment Allocations Obtained with PAM, Unweighted CCR

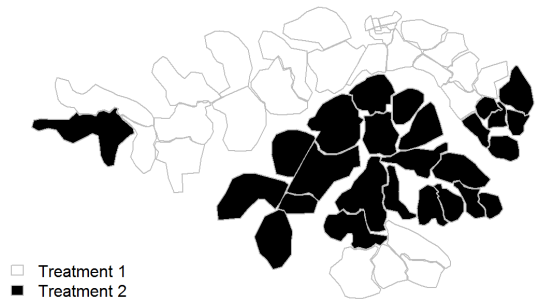


Results - Study 1 - Visualization of Treatments (Unweighted)

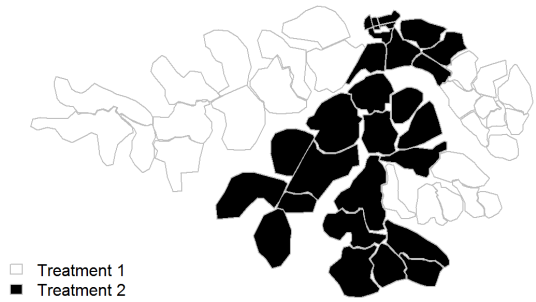
Treatment Allocations Obtained with K-Means and Unweighted CCR



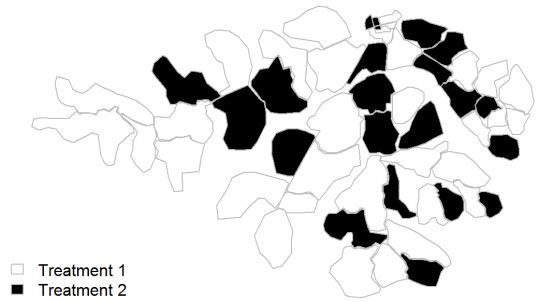
Treatment Allocations Obtained with Ward-Like Hierarchical Clustering, Unweighted CCR



Treatment allocations Obtained with CLARA Clustering, Unweighted CCR



Treatment Allocations Obtained with PAM, Unweighted CCR



Results - Study 2 - CCR Balance

Table: Study 2 CCR Results and CPU Run Times (281 Units).

Method	Selected Balance Score	10% Cutoff Score
K-Means Weighted	0.098	0.098
K-Means Unweighted	0.695	0.695
CLARA Weighted	0.203	0.203
CLARA Unweighted	2.096	2.096
H-Clust Weighted	0.303	0.753
H-Clust Unweighted	2.795	7.872
PAM Weighted	0.708	1.397
PAM Unweighted	6.779	13.247

Results - Study 2 - CCR Balance

Method	N Clusters/Arm	N Allocations/Space
K-Means Weighted	3	2
K-Means Unweighted	3	2
CLARA Weighted	3	2
CLARA Unweighted	3	2
H-Clust Weighted	25	2500000
H-Clust Unweighted	25	2500000
PAM Weighted	46	2500000
PAM Unweighted	46	2500000

Results - Study 2 - Spatial Clustering of Treatment Arms

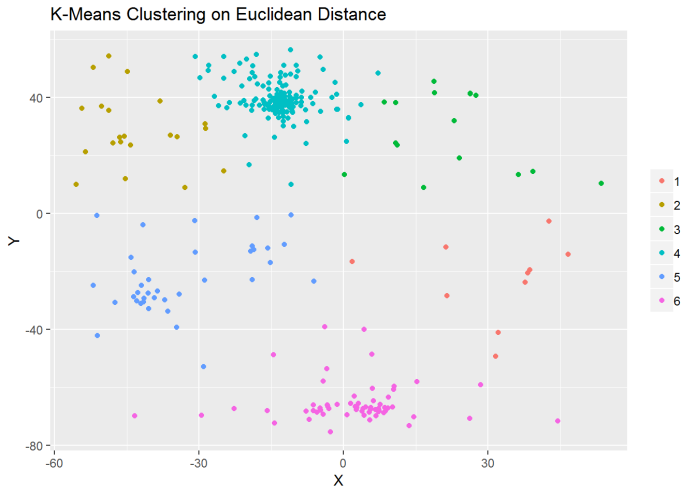
Method	Clustering present ($\alpha = 0.05$)?
K-Means Weighted	Yes
K-Means Unweighted	Yes
CLARA Weighted	Yes
CLARA Unweighted	Yes
H-Clust Weighted	No
H-Clust Unweighted	No
PAM Weighted	Yes
PAM Unweighted	No

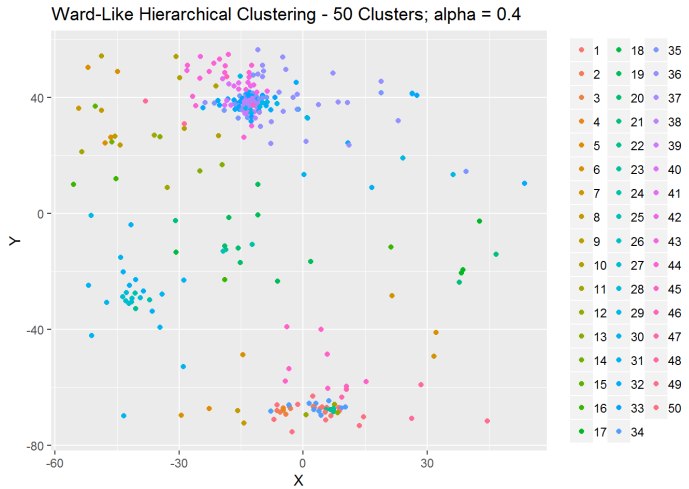
Results - Study 2 - CPU Run Times (Seconds)

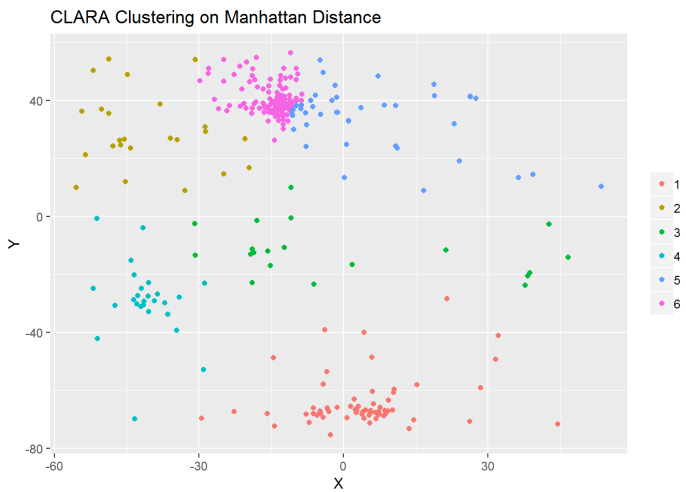
Method	Cluster Optimization
K-Means	1.395
CLARA	11.937
H-Clust	620.302
PAM	68.336

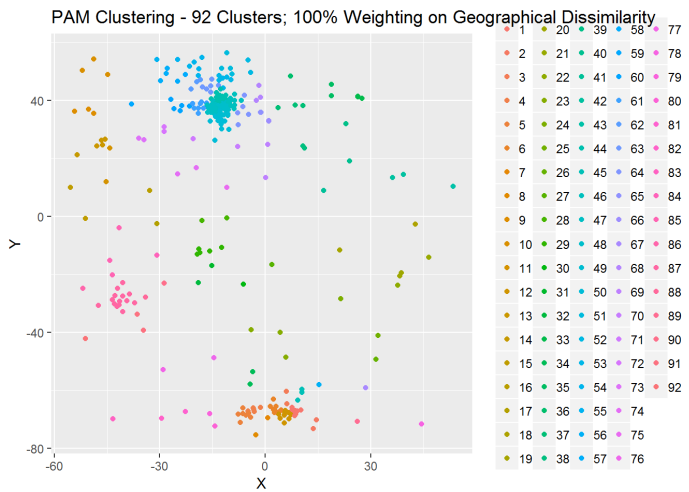
Method	Weighted CCR	Unweighted CCR
K-Means	0.016	0.040
CLARA	0.019	0.071
H-Clust	1205.195	1230.907
PAM	1370.509	1239.719

Results - Study 2 - Visualization of Clusters

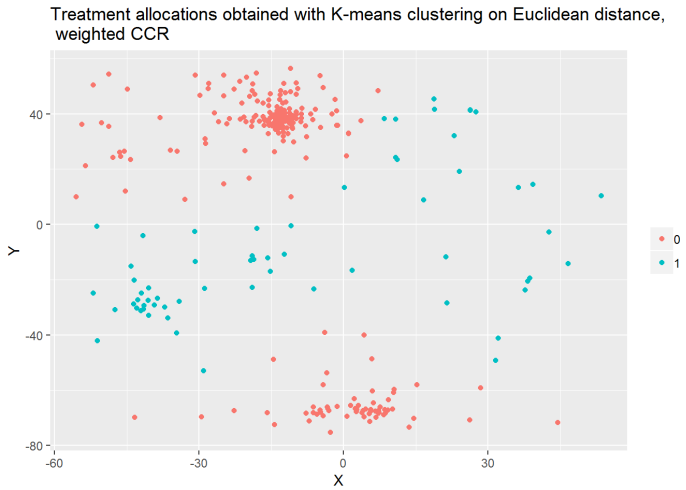


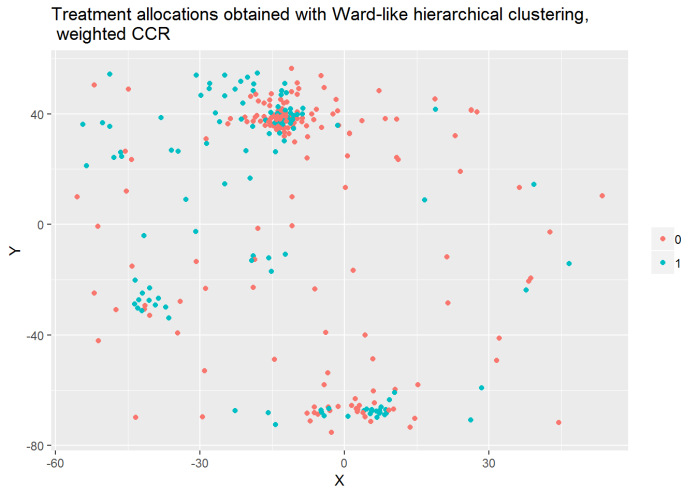


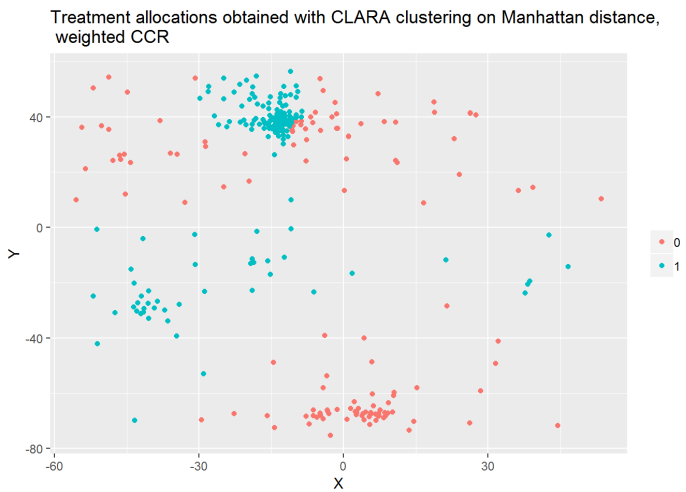


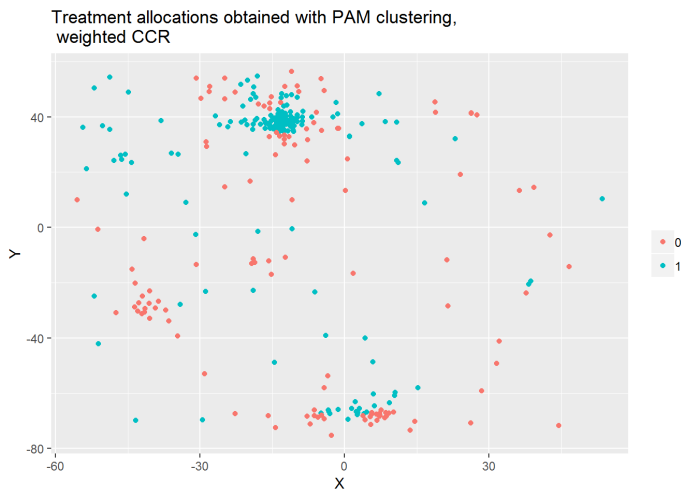


Results - Study 2 - Visualization of Treatments (Weighted)

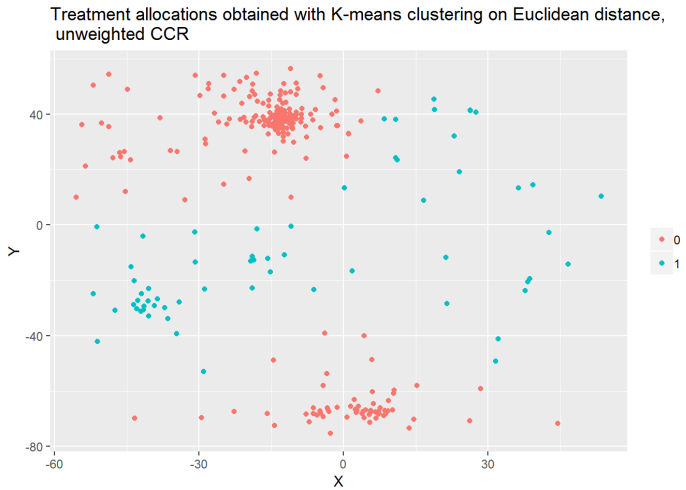


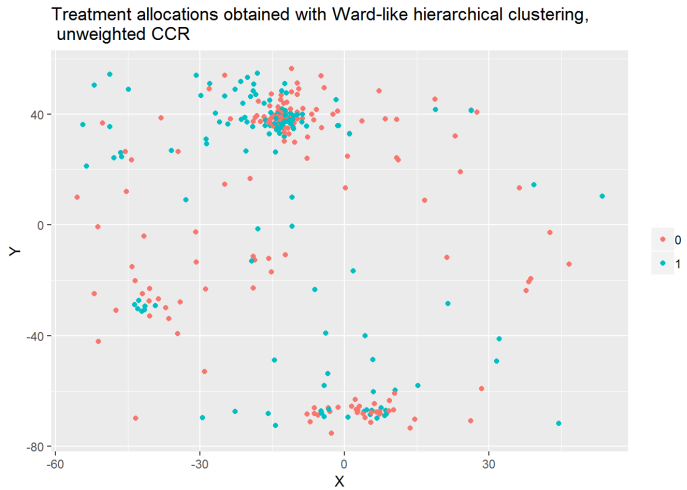


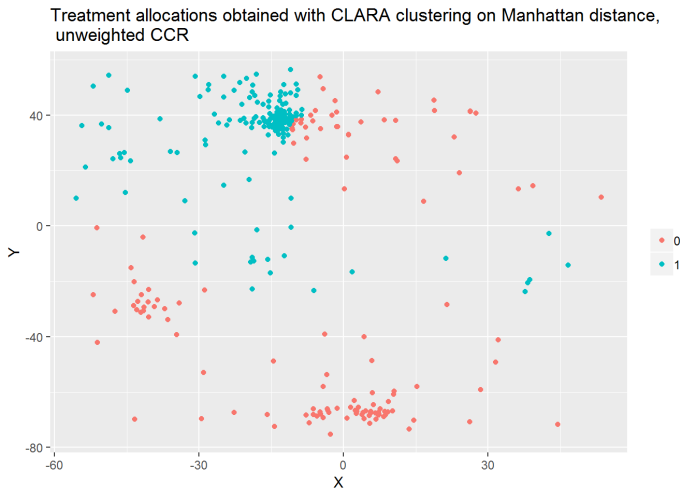


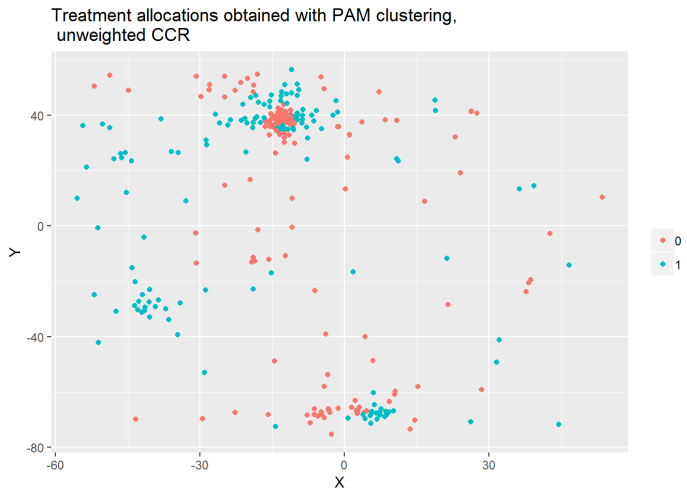


Results - Study 2 - Visualization of Treatments (Unweighted)









Conclusions

- Strategic weighting of covariates during CCR, regardless of clustering method, drastically improved covariate balance during randomization into CRT treatment arms
- For small studies, Ward-Like Hierarchical Clustering with mixed covariate-and-geographical dissimilarities and weighted CCR may be an optimal study design strategy
- However, for the larger second study, K-Means and CLARA performed better than Hierarchical or PAM in terms of selected and cutoff scores, albeit with only two possible allocation schemes with which to randomize

Conclusions

- Weighted K-means also maintained spatial heterogeneity of treatment assignments for both studies
- K-means with Euclidean geographic dissimilarities and covariate weighting during CCR may be the optimal method tested
- However, while simply using geographic dissimilarity may suffice for balancing on spatial feature data, it does not guarantee that the scheme selected at the end of CCR avoids spatial contamination of treatment arms
  - To do so requires additional methodology, in which only allocation schemes that result in spatial heterogeneity of treatment arms are considered during CCR

Limitations

- Only cluster quality statistic used to optimize K-means, CLARA, and PAM was average silhouette width
- "ClustGeo" limits the maximum number of clusters to 55. This may be impractical for some medium-to-large CRT studies. Also took significantly longer to optimize and perform while scaling up in study size within the constraint of 55 clusters - may not be practically scale-able
- Drastically-increased CCR run times for study 2's Hierarchical and PAM clustering methods - "cvcrand" may also be prohibitively time-consuming with larger studies

Limitations

- Did not examine the effects of altering the balance score cutoff point - we simply set it to 10%
  - Minimum number of testable clusters set to 6 and a cutoff of 10%, there can be a minimum of only 2 randomization schemes to select from during CCR - some CRT designers make take issue
- Due to computational power limitations, this study only allowed for a maximum consideration of 25,000,000 randomization schemes with "cvcrand"
  - May be unwise to compare clustering methods when it is not possible to have the entire sample space for consideration during CCR - not fully representative
- "cvcrand" utilizes only one main method of calculating imbalance for a given covariate at the treatment level - difference in treatment means, normalized with overall standard deviation

Limitations

- In both studies, PAM yielded the lowest minimum and the highest maximum CCR scores in the randomization space - was optimized for the greatest number of clusters as well
  - Minimum and maximum CCR scores may be influenced by number of clusters
- Clustering and treatment assignments, as well as conclusions on spatial heterogeneity of treatment assignments, may be influenced by geographical barriers between units (e.g., mountains, canyons, swamps)
  - May not be relevant for a spatially-dependent outcome in which these barriers may not have practical meaning (e.g., a malaria-infected mosquito flying across swamps)
  - May result in a failure to detect spatial heterogeneity when present

References I

K-means cluster analysis. URL https://uc-r.github.io/kmeans_clustering#kmeans.

Distance. URL http://mathworld.wolfram.com/Distance.html.

Toxic substances portal - trichloroethylene (tce), Jan 2015. URL https://www.atsdr.cdc.gov/phs/phs.asp?id=171&tid=30.

a civil action carcinogen pollutes tap water supplies for 14 million americans, 2018. URL https://www.ewg.org/childrenshealth/carcinogen-pollutes-tap-water-supplies-14-million-americ

Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. URL http://www.jstatsoft.org/v12/i06/.

References II

Margherita Barile. Taxicab metric. URL http://mathworld.wolfram.com/TaxicabMetric.html.

Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995. ISBN 0198538642.

Roger Bivand, Jakub Nowosad, and Robin Lovelace. spdata: Datasets for spatial analysis, 2019. dataset retrieved from "spData" package version 0.3.0, https://CRAN.R-project.org/package=spData.

Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL http://www.asdar-book.org/.

References III

Marie Chavent, Vanessa Kuentz, Amaury Labenne, and Jerome Saracco. *ClustGeo: Hierarchical Clustering with Spatial Constraints*, 2017. URL `https://CRAN.R-project.org/package=ClustGeo`. R package version 2.0.

Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, and Jrme Saracco. Clustgeo: an r package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4): 17991822, 2018. doi: 10.1007/s00180-018-0791-1. URL `https://link.springer.com/content/pdf/10.1007/s00180-018-0791-1.pdf`.

References IV

L. Miriam Dickinson, Brenda Beaty, Chet Fox, Wilson Pace, W. Perry Dickinson, Caroline Emsermann, and Allison Kempe. Pragmatic cluster randomized trials using covariate constrained randomization: A method for practice-based research networks (pbrns). *The Journal of the American Board of Family Medicine*, 28(5):663–672, 2015. ISSN 1557-2625. doi: 10.3122/jabfm.2015.05.150001. URL https://www.jabfm.org/content/28/5/663.

PhD Elizabeth A. Albright. Two independent samples unequal variance (welch's test), Aug 2018. URL https://sites.nicholas.duke.edu/statsreview/means/welch/.

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528823.

References V

J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346830.

Xin Jin and Jiawei Han. *K-Medoids Clustering*, pages 564–565. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_426. URL https://doi.org/10.1007/978-0-387-30164-8_426.

Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. 01 1990. ISBN 0-471-87876-6. doi: 10.2307/2532178.

Eugene F. Krause. *Taxicab geometry an adventure in non-Euclidean geometry*. Dover Publ., 1986.

References VI

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2018. R package version 2.0.7-1 — For new features, see the 'Changelog' file (in the package source).

Lawrence H Moulton. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*, 1(3):297–305, 2004. doi: 10.1191/1740774504cn024oa. URL https://doi.org/10.1191/1740774504cn024oa. PMID: 16279255.

Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005. URL https://CRAN.R-project.org/doc/Rnews/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

References VII

Gillian M. Raab and Izzy Butcher. Balance in cluster randomized trials. *Statistics in Medicine*, 20(3):351–365, 2001. doi: 10.1002/1097-0258(20010215)20:3⟨351::AID-SIM797⟩3.0.CO;2-C.

A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504, Dec 2006. ISSN 1572-9214. doi: 10.1007/s10852-005-9022-1. URL https://doi.org/10.1007/s10852-005-9022-1.

Matthias Studer. Weightedcluster library manual: A practical guide to creating typologies of trajectories in the social sciences with r. Technical report, LIVES Working Papers 24, 2013. DOI: 10.12682/lives.2296-1658.2013.24.

References VIII

Lance Waller and Carol A. Gotway. Applied spatial statistics for public health data. *Applied spatial statistics for public health data*, 01 2004. doi: 10.1002/0471662682.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `http://ggplot2.org`.

Johnson A. Wolverton, B. C. and K. Bounds. Interior landscape plants for indoor air pollution abatement, final report, september n.a.s.a. 1989 stennis space centre ms. *Water, Air, and Soil Pollution*, September 1989. URL `https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19930073077.pdf`.

References IX

Hengshi Yu, Fan Li, John A. Gallis, and Elizabeth L. Turner. *cvcrand: Efficient Design and Analysis of Cluster Randomized Trials*, 2019. URL `https://CRAN.R-project.org/package=cvcrand`. R package version 0.0.3.