Spatial design

Applied Spatial Statistics

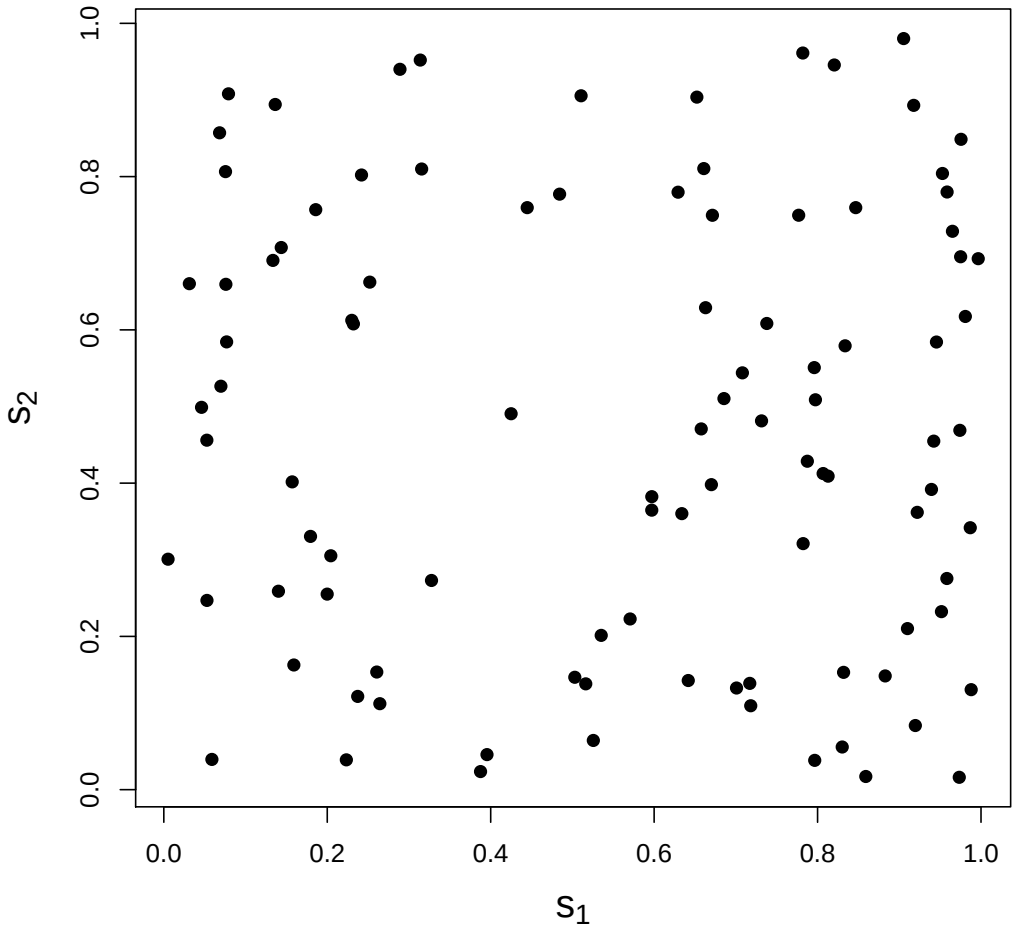
Spatial design

- In some cases, there is an opportunity to design data collection
- Say at the onset of a field season there is time/money to set up n stations, how to select $\mathbf{s}_1, \ldots, \mathbf{s}_n$?
- The optimal sampling locations depends on the objectives of the study
- There is often a balance between the optimal statistical design and logistical constraints

Types of designs

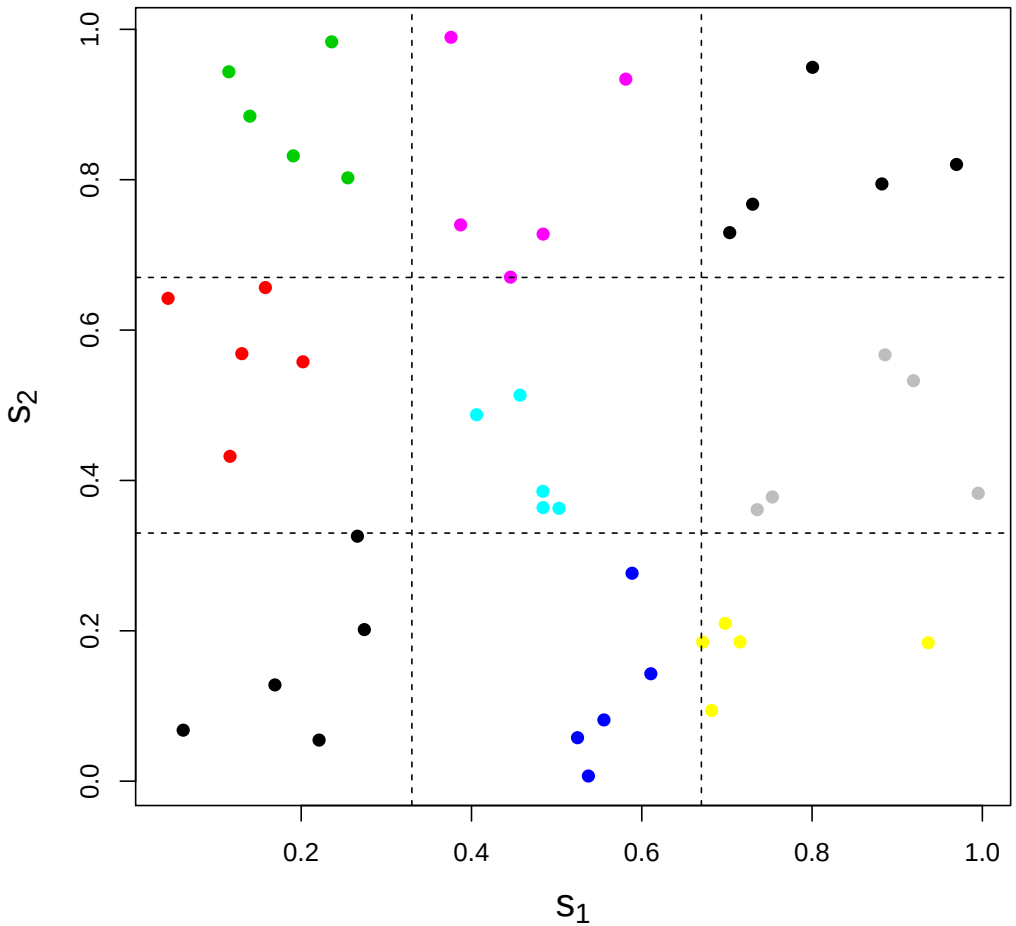
- Completely random sample
- Stratified random sample by **s**
- Space-filling design/regular grid
- Stratified random sample by **X**
- Cluster sample
- Split plots
- Convenience sample

Completely random sample



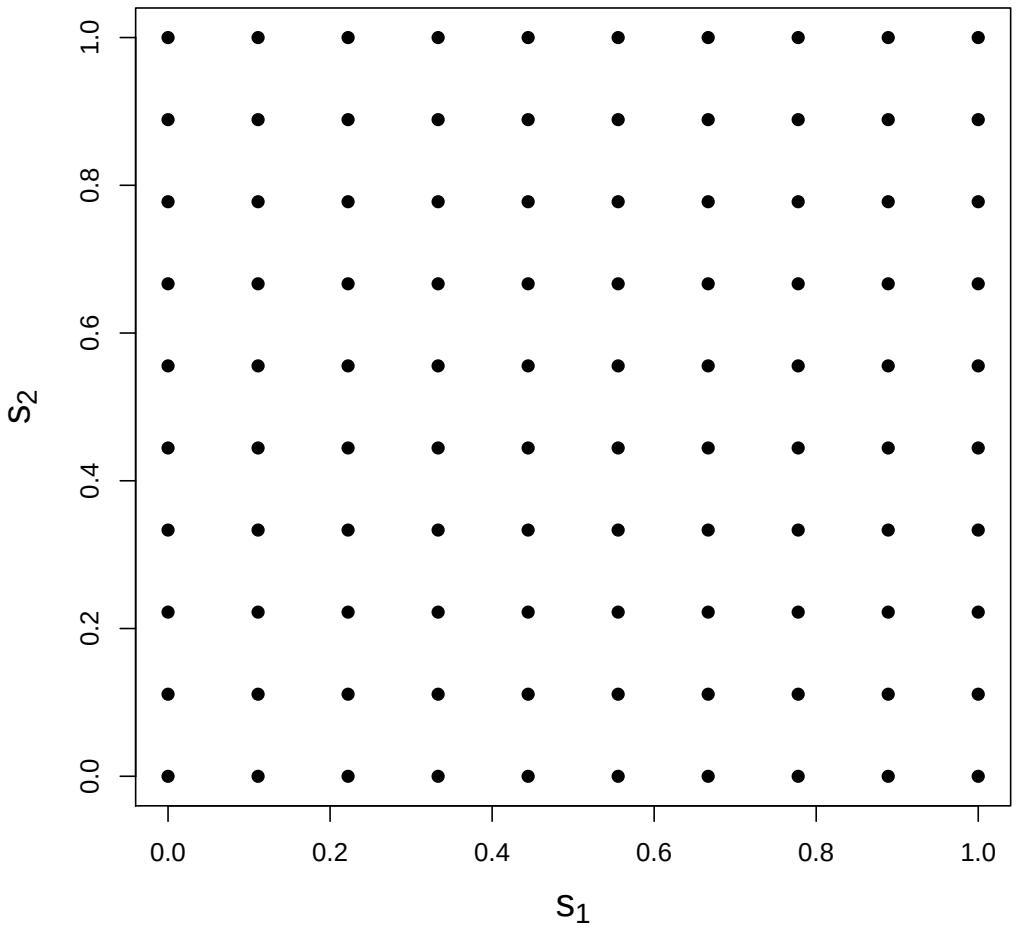
The sample locations are purely random

Stratified by **s**



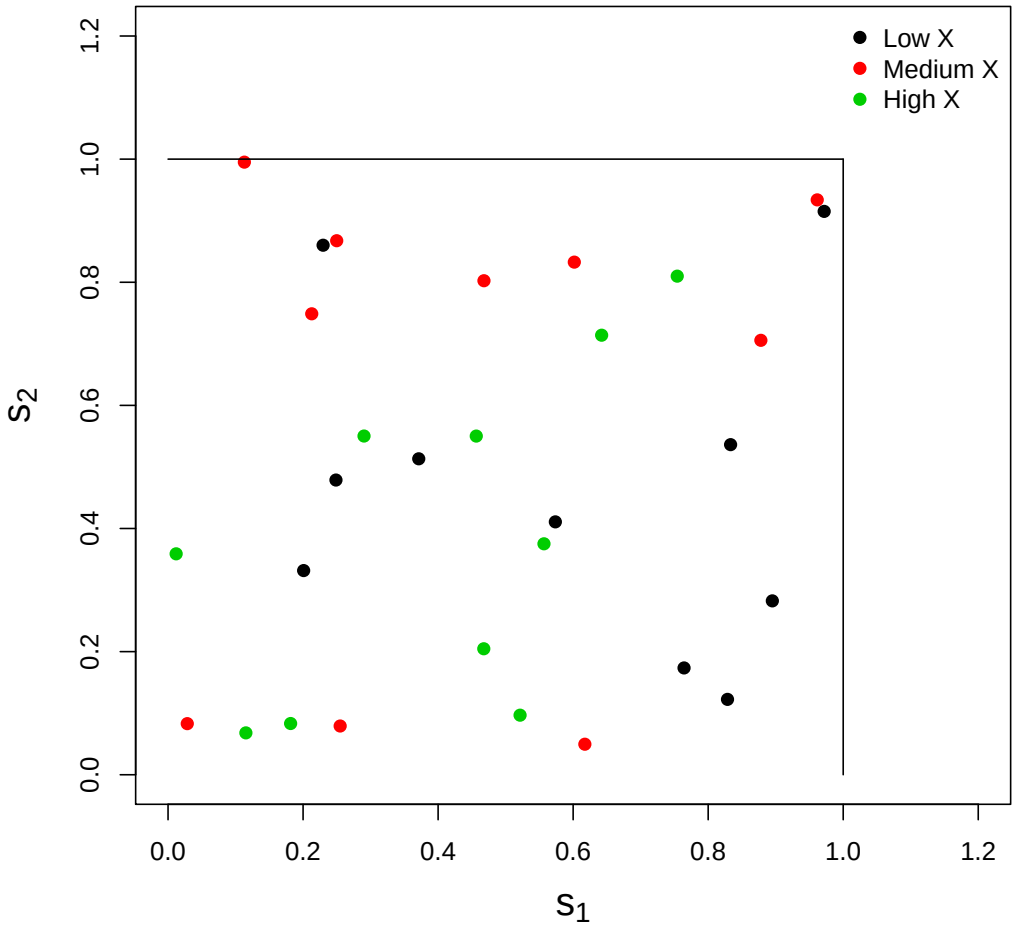
Partition the domain, and take a random sample in each region

Space-filling design/regular grid



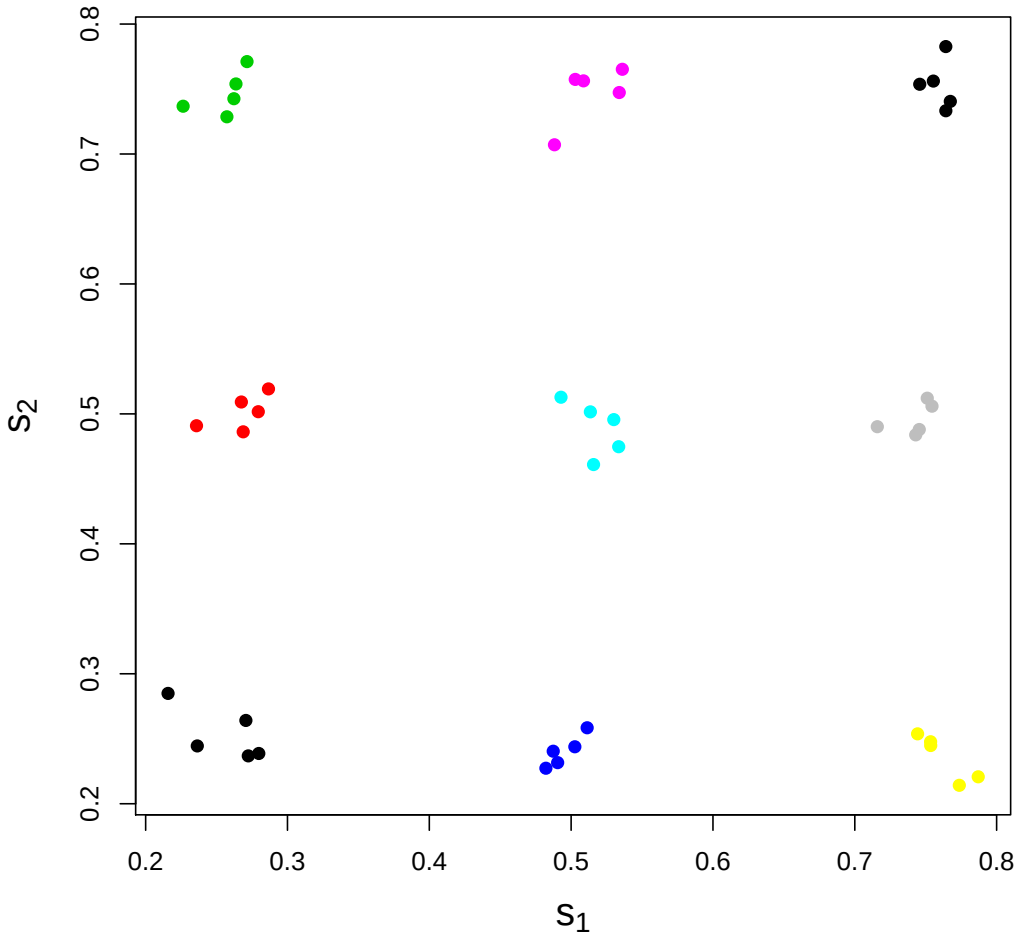
Maximize the minimum distance between points

Stratified by **X**



Partition the observations by a covariate, and take a random sample in each covariate range

Cluster design



Randomly draw samples around cluster centers

Split plots

- Split plots are great when you are evaluating a treatment that you control
- A “plot” is a cluster of locations, e.g., a corn field
- A split plot assigns different treatments to observations within each plot
- Comparing observations within the plot but with different treatment isolates the treatment effect

Convenience sample

- In practice, it is often hard to get to a randomly-generated location
- For convenience, samples are often taken near roads or cities
- Even under this restriction, you can design surveys to balance across space or covariates
- There is usually this balance

Optimal design

- How to pick the “best” design?
- We first need to define “best” mathematically
- Example: minimize prediction mean squared error (MSE)
- Example: minimize MSE for θ
- Example: minimize MSE for β
- Example: minimize MSE for estimating a treatment effect

Optimal design

- With a criteria in mind, you can then compare different candidate for the n locations
- At this stage, there is no data, and so you have to rely on mathematical approximations to the design criteria
- It is rarely possible to find the optimal set of n location exactly
- Be best we have are rules of thumb

Optimal design rules of thumb

- If the goal is prediction and θ and β are known, a regular grid is good
- If the goal is to estimate θ, a cluster sample is good
- If the goal is to estimate β, stratifying by $\mathbf{X}$ is good
- If the goal is prediction and θ and β are unknown, a regular grid with a few clusters is good
- Splits plots are good for estimating treatment effects
- A completely random sample is never too bad