Regularización: Lasso

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Regularization
   - Recap
   - Lasso
   - Familia de regresiones penalizadas
   - Ridge and Lasso: Pros and Cons
   - Elastic Net
2. Spatial Data
   - Motivation
   - Types of Spatial Data
   - Projections
   - Spatial Dependence

Agenda

Agenda

Regularización: Motivación

- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error "*dentro de muestra*", eligiendo β de forma tal que

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 \tag{1}$$

- pero para predicción, no estamos interesados en hacer un buen trabajo dentro de muestra
- Queremos hacer un buen trabajo, **fuera de muestra**

Ridge

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.
- Vamos a proponer modelos del estilo

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{2}$$

- donde R es un regularizador que penaliza funciones que crean varianza
- Explícitamente en la minimización incluimos un termino de sesgo y un termino de varianza.

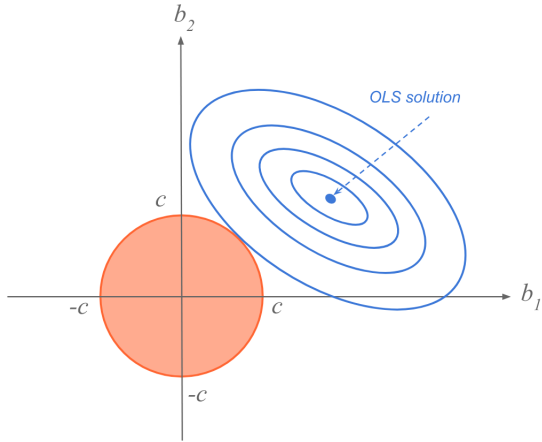
Ridge

- Para un $\lambda \geq 0$ dado, consideremos ahora el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{3}$$

Intuición en 2 Dimensiones (Ridge)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{4}$$



Agenda

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

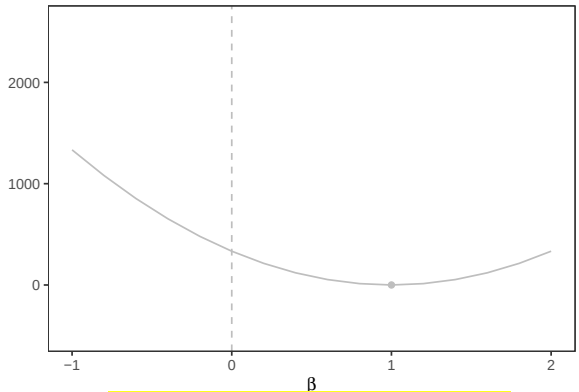
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

  - "LASSO's free lunch": selecciona automáticamente los predictores que van en el modelo ($\beta_j \neq 0$) y los que no ($\beta_j = 0$)
  - Por qué? Los coeficientes que no van son soluciones de esquina
  - $L(\beta)$ es no differentiable

Intuición en 1 Dimensión

$\hat{\beta} > 0$

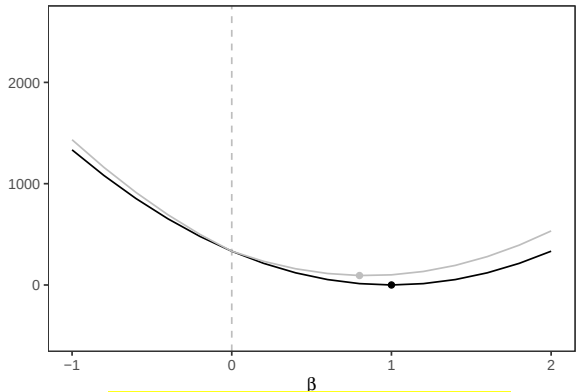
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{6}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

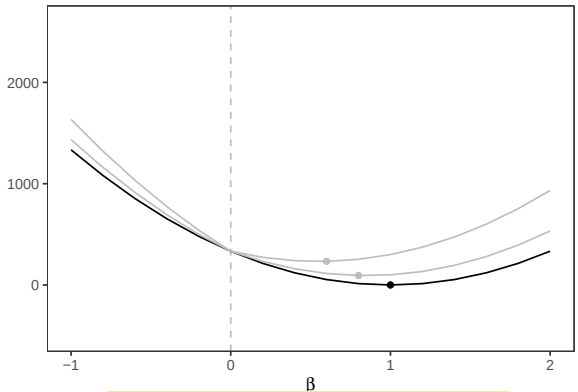
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{7}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

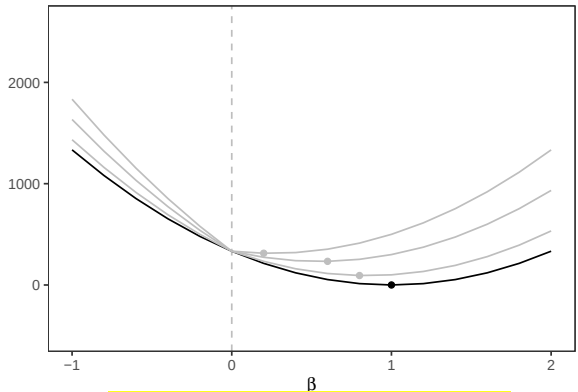
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{8}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

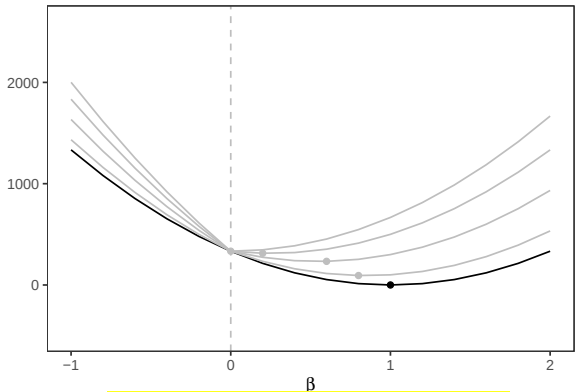
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{9}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

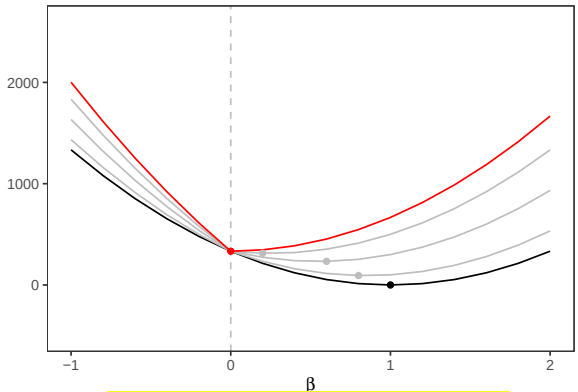
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{10}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

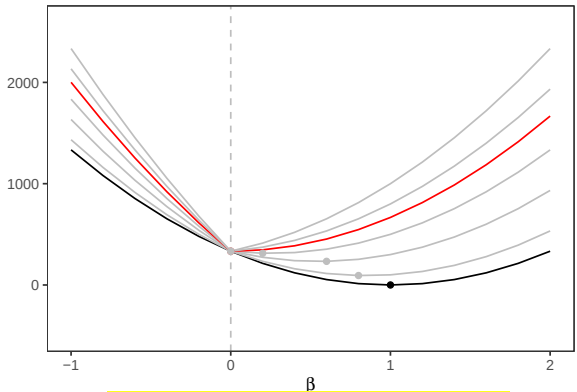
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{11}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{12}$$



Intuición en 1 Dimension

Solución analitica

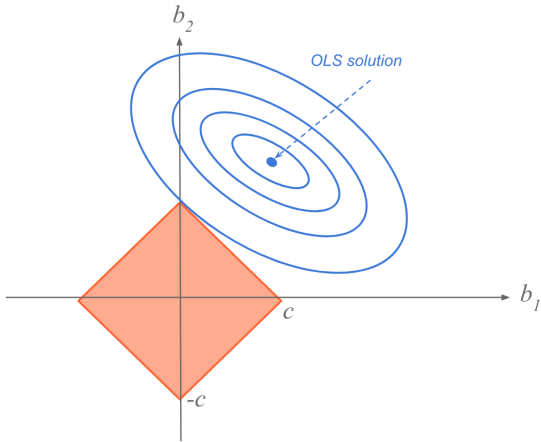
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{13}$$

- la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \text{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \text{si } \lambda < \lambda^* \end{cases} \tag{14}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \tag{15}$$



Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Resumen

- Ridge y Lasso son sesgados, pero las disminuciones en varianza pueden compensar estoy y llevar a un MSE menor
- Lasso encoje a cero, Ridge no tanto
- Importante para aplicación:
  - Estandarizar los datos
  - Como elegimos λ?

Resumen

- Ridge y Lasso son sesgados, pero las disminuciones en varianza pueden compensar estoy y llevar a un MSE menor
- Lasso encoje a cero, Ridge no tanto
- Importante para aplicación:
  - Estandarizar los datos
  - Como elegimos λ? $\rightarrow$ Validación cruzada

Agenda

Family of penalized regressions

$$min_{\beta} R(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|^p \tag{16}$$

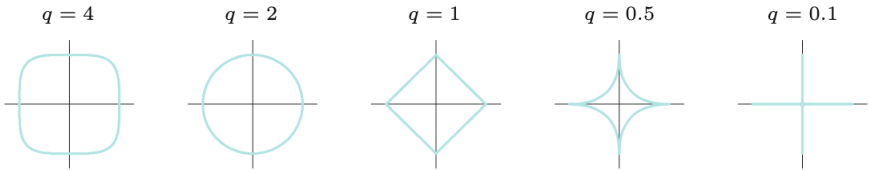


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Agenda

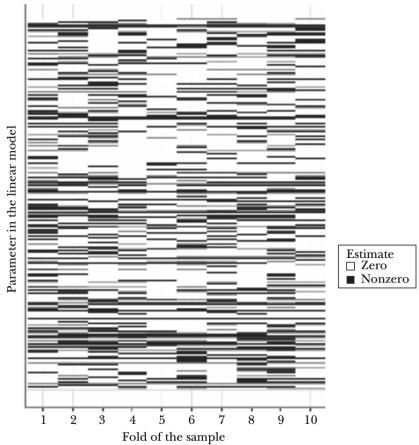
Ridge and Lasso: The good and the bad

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- More predictors than observations ($k > n$)
  - OLS fails
  - Ridge augments data
  - Lasso chooses at most n variables

Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one.

Ridge and Lasso: The good and the bad



Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one. Makes it unstable for prediction.
  - Ridge shrinks the coefficients of correlated variables toward each other. This makes Ridge "work" better than Lasso. "Work" in terms of prediction error

Agenda

Elastic net

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \quad (17)$$

- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Ridge

Elastic Net

- Elastic net: happy medium.
  - Good job at prediction and selecting variables

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{18}$$

- Mixes Ridge and Lasso
- Lasso selects predictors
- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- How to choose (λ, α)? $\rightarrow$ Bidimensional Crossvalidation
- Recomended lecture: Zou, H. & Hastie, T. (2005)
- H.W.: $\beta_{OLS} > 0$ one predictor standarized

$$\hat{\beta}_{EN} = \frac{\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)_+}{1 + \lambda_2} \tag{19}$$

Example

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

1. Regularization
   - Recap
   - Lasso
   - Familia de regresiones penalizadas
   - Ridge and Lasso: Pros and Cons
   - Elastic Net

2. Spatial Data
   - Motivation
   - Types of Spatial Data
   - Projections
   - Spatial Dependence

Agenda

Motivation

- In Big Data volume was only a part of the story
- Big Data are data of high complexity: anarchic and spontaneous
- They are the by product of an action: pay with credit card, tweet, move from point A to point B, buy a house, etc.
- Now we are going to center on spatial data

Agenda

Types of Spatial Data

Spatial data comes in many “shapes” and “sizes”, the most common types of spatial data are:

- Points are the most basic form of spatial data. Denotes a single point location, such as cities, a GPS reading or any other discrete object defined in space.
- Lines are a set of ordered points, connected by straight line segments
- Polygons denote an area, and can be thought as a sequence of connected points, where the first point is the same as the last
- Grid (Raster) are a collection of points or rectangular cells, organized in a regular lattice

Types of Spatial Data: Points

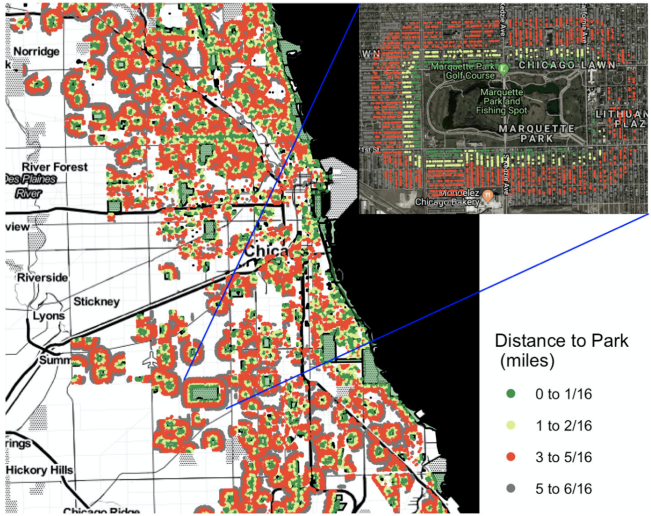


Fig. 1. Housing transactions around parks: neighborhood distance intervals. Notes: The following figure shows transactions within 3/8 miles of the nearest park in Chicago. The

Types of Spatial Data: Lines

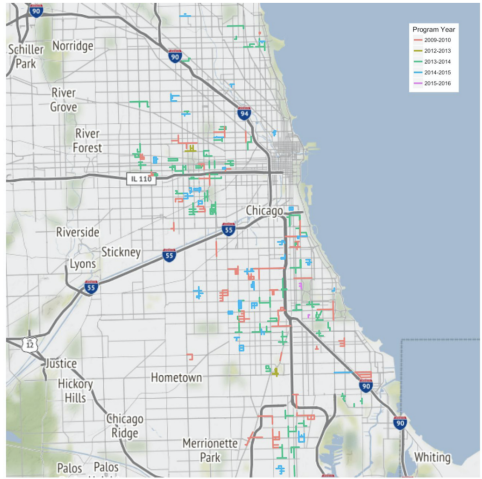
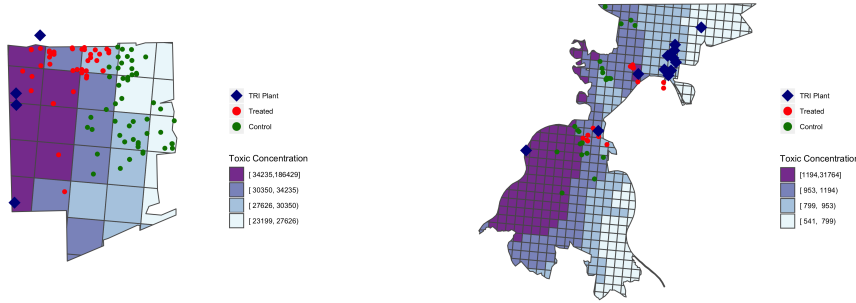


Fig. 1. Safe Passage Routes, by year of program adoption.
Note: Shapefiles with Safe Passage shape and location where obtained from the Chicago Data Portal and year that the program was launched at each location through a FOIA request.

McMillen, Sarmiento-Barbieri & Singh, 2019

Types of Spatial Data: Rasters



Christensen,Sarmiento-Barbieri & Timmins (2022)

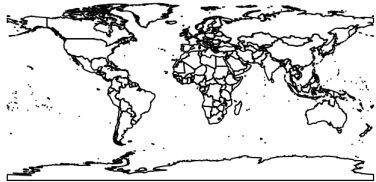
Agenda

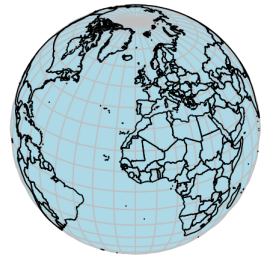
The earth ain't flat

- The world is an irregularly shaped ellipsoid, but plotting devices are flat
- But if you want to show it on a flat map you need a map projection,
- This will determine how to transform and distort latitudes and longitudes to preserve some of the map properties: area, shape, distance, direction or bearing

The earth ain't flat

- Map projections try to portray the surface of the earth or a portion of the earth on a flat piece of paper or computer screen.
- A coordinate reference system (CRS) then defines, with the help of coordinates, how the two-dimensional, projected map in your GIS is related to real places on the earth.





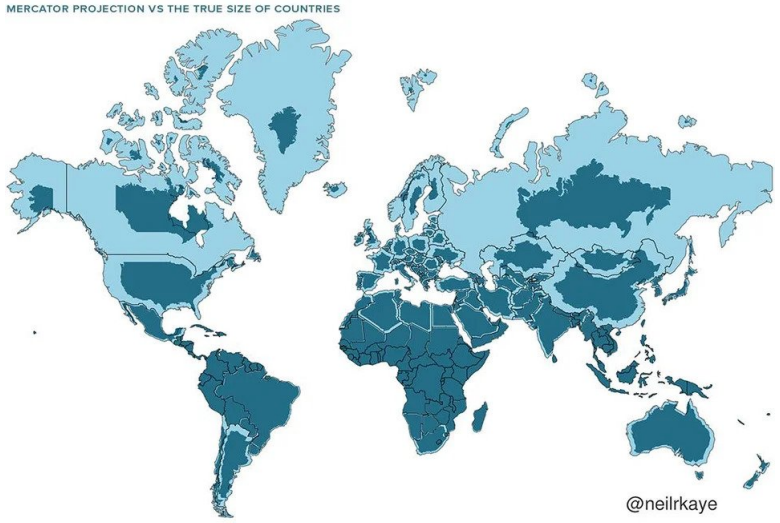
The earth ain't flat

- For example, sailors use Mercator projection where meridians and parallels cross each other always at the same 90 degrees angle.
- It allows to easy locate your self on the line showing direction in which you sail
- But the projection do not preserve distances

Source: https://www.geoawesomeness.com/all-map-projections-in-compared-and-visualized/

The earth ain't flat

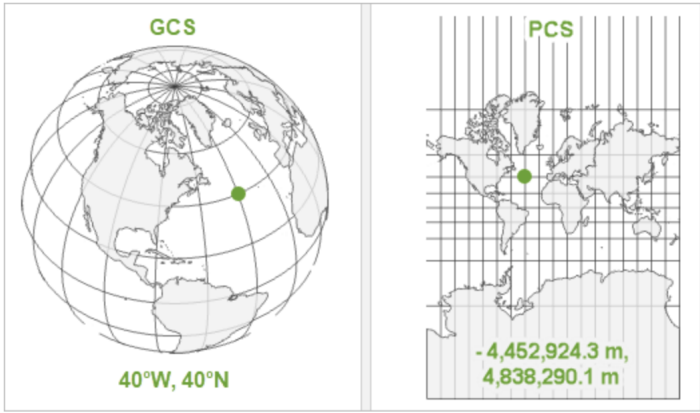
Mercator and the true size of countries



Source: https://www.reddit.com/r/Damnthatsinteresting/comments/xziol9/mercator_projection_vs_the_true_size_of_countries

Coordinate Reference System (CRS)

- With the help of coordinate reference systems (CRS) every place on the earth can be specified by coordinates.



Source: https://www.geoawesomeness.com/all-map-projections-in-compared-and-visualized/

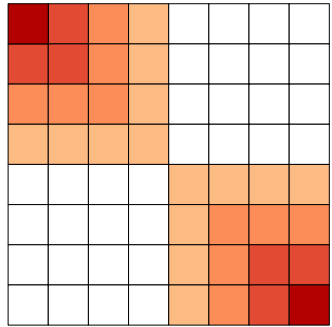
Which projection should I choose?

- "There exist no all-purpose projections, all involve distortion when far from the center of the specified frame" (Bivand, Pebesma, and Gómez-Rubio 2013)
- The decision as to which map projection and coordinate reference system to use, depends on the regional extent of the area you want to work in, on the analysis you want to do and often on the availability of data.
- In some cases, it is not something that we are free to decide: "often the choice of projection is made by a public mapping agency" (Bivand, Pebesma, and Gómez-Rubio 2013).
- This means that when working with local data sources, it is likely preferable to work with the CRS in which the data was provided.

Agenda

Spatial Dependence

- We now take a closer look at spatial dependence, or to be more precise on it's weaker expression spatial (auto)correlation.
- Spatial autocorrelation measures the degree to which a phenomenon of interest is correlated to itself in space (Cliff and Ord (1973)).
- For example, positive spatial correlation arises when units that are *close* to one another are more similar than units that are far apart



Spatial Dependence

- We can express the existence of spatial autocorrelation with the following moment condition:

$$Cov(y_i, y_j) \neq 0 \quad for \quad i \neq j \tag{20}$$

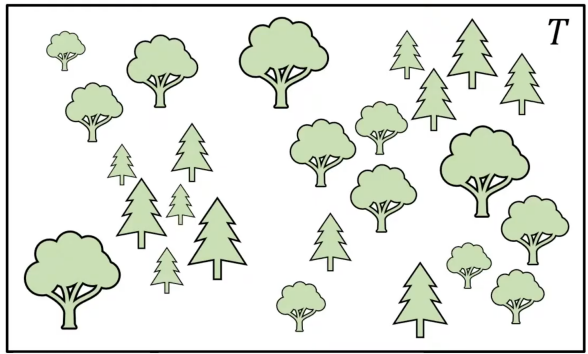
  were y_i and y_j are observations on a random variable at locations i and j.

- This autocorrelation can lead to overfitting of the model and poor generalization to new spatial locations.

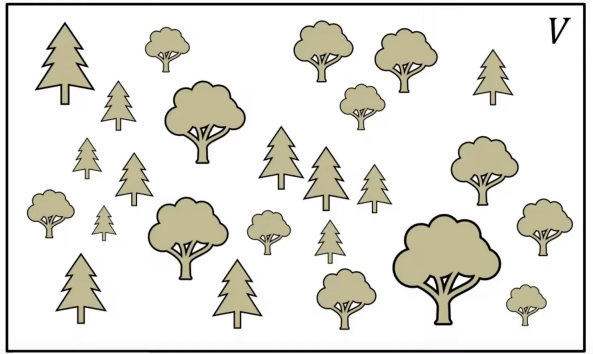
- By using spatial cross-validation, we can ensure that the model is tested on data that is independent of the training data and has similar spatial characteristics.

Spatial Prediction and Cross-Validation

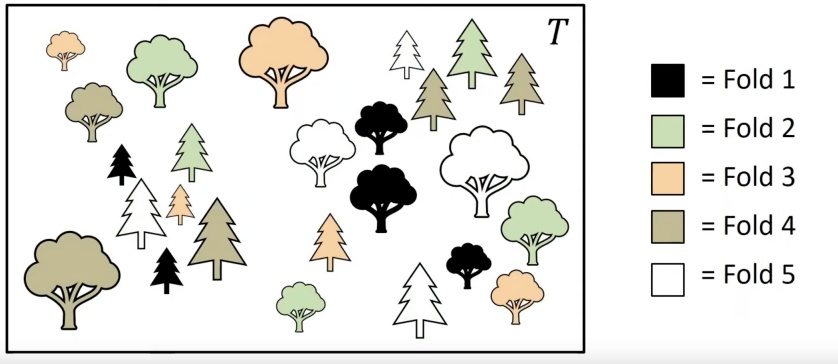
Spatial Prediction and Cross-Validation



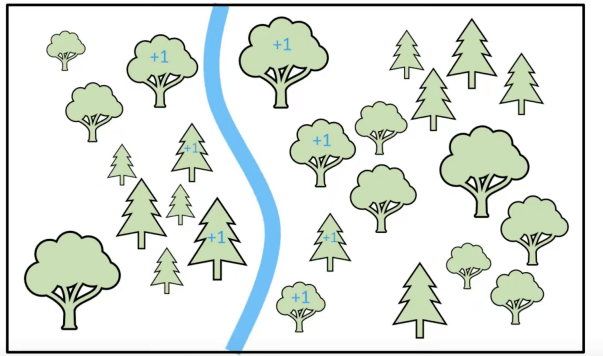
Spatial Prediction and Cross-Validation



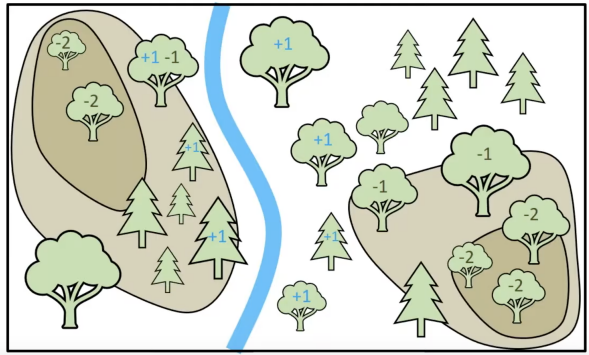
Spatial Prediction and Cross-Validation



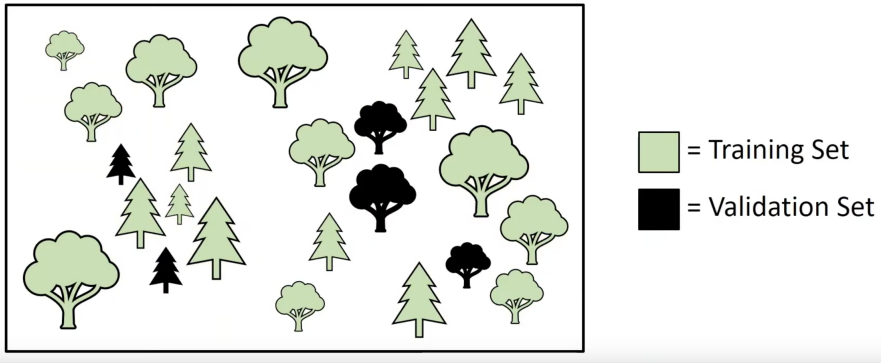
Spatial Prediction and Cross-Validation



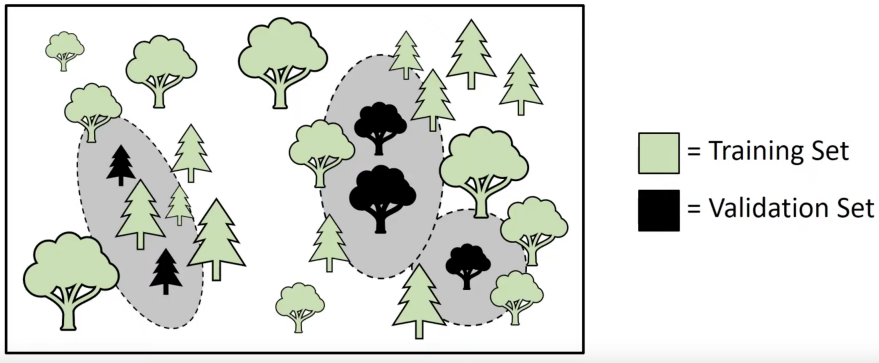
Spatial Prediction and Cross-Validation



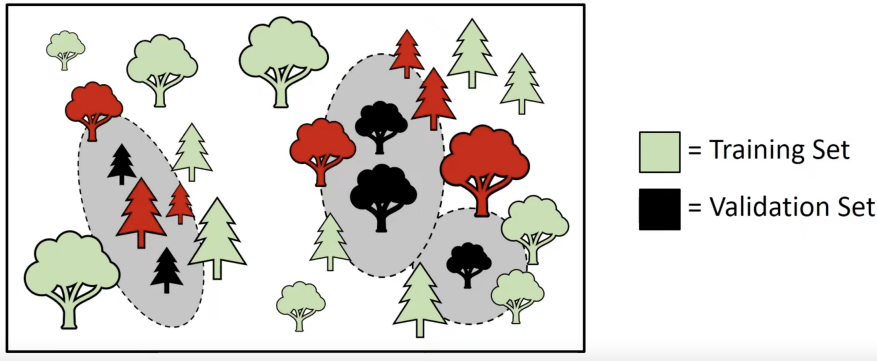
Spatial Prediction and Cross-Validation



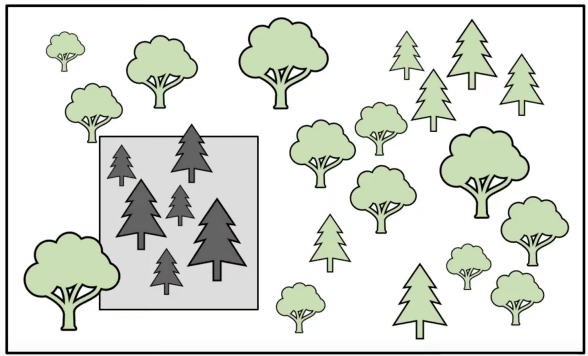
Spatial Prediction and Cross-Validation



Spatial Prediction and Cross-Validation



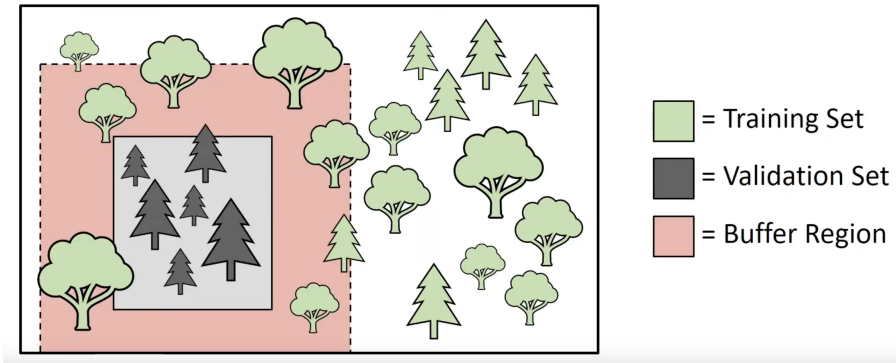
Spatial Prediction and Cross-Validation



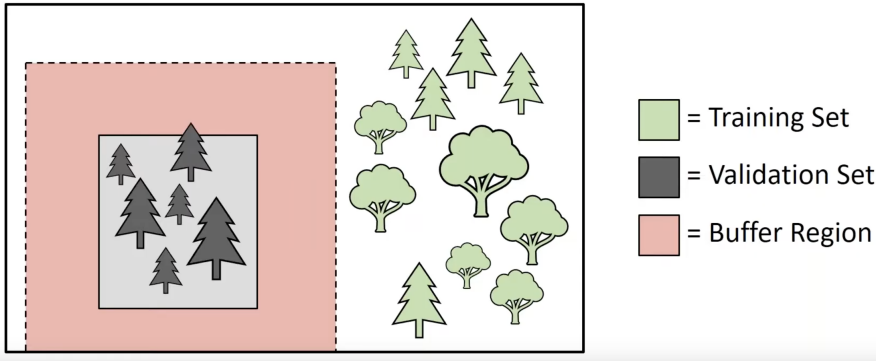
= Training Set

= Validation Set

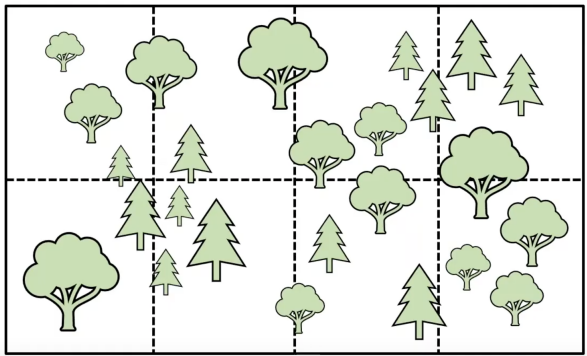
Spatial Prediction and Cross-Validation



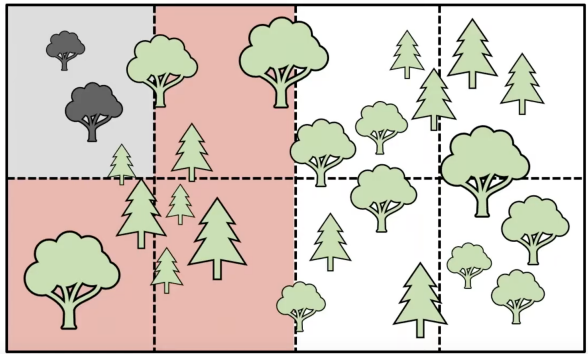
Spatial Prediction and Cross-Validation



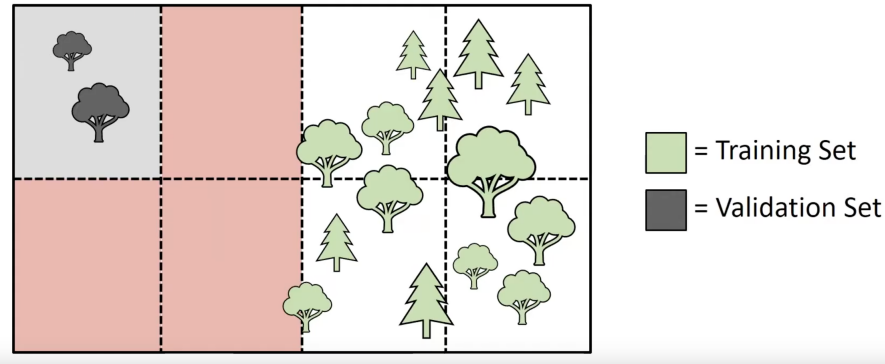
Spatial Prediction and Cross-Validation



Spatial Prediction and Cross-Validation



Spatial Prediction and Cross-Validation



Example: Spatial Cross-Validation

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/