jornadassiglibre
Geotech/spatial data science

Universitat de Girona
Servei de Sistemes d'Informació Geogràfica i Teledetecció

Técnicas de validación cruzada espacial en el paquete de R *CAST*

Carles Milà Garcia

Barcelona Institute for Global Health (ISGlobal)
University of Münster (WWU)

carles.mila@isglobal.org carles.mila@gmail.com https://github.com/carlesmila

ISGlobal Barcelona Institute for Global Health

WWU MÜNSTER

Sobre mi

jornadassiglibre
Geotech/spatial data science

Investigador predoctoral en ciencia de datos espacial

Investigación aplicada a la epidemiología ambiental

ISGlobal Barcelona Institute for Global Health

- Modelización espacio-temporal de exposiciones ambientales
- Métodos de deep learning para la estimación de riesgo de pobreza con imágenes de satélite

Investigación metodológica

WWU MÜNSTER

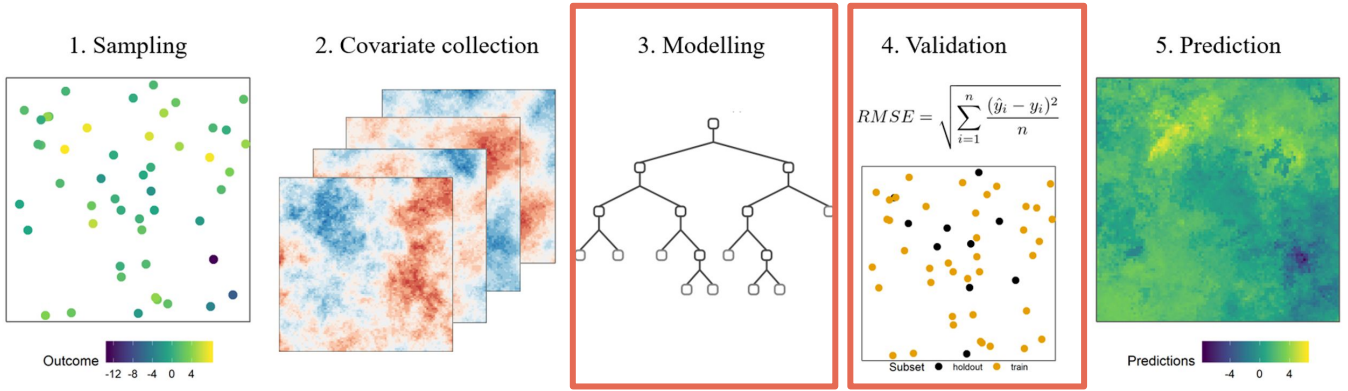
Remote Sensing & Spatial Modelling

- Cómo hacer que los modelos de machine learning sean espaciales
- Cómo estimar la calidad de las predicciones espaciales

Predicción espacial con datos medioambientales

jornadassiglibre
Geotech/spatial data science

Típico workflow:



- Meteorología
- Calidad aire
- Ecología
- Suelos
- Ingeniería forestal
- Ingeniería agrícola

Crear estas predicciones es relativamente fácil, pero ¿cuánto podemos confiar en ellas?

¿qué hiperparámetros deberíamos escoger?

Validación de modelos predictivos medioambientales

1. Muestras independientes para evaluación (*probability test sampling*)
   - Estadísticos sin sesgo (Milà et al. 2024, Wadoux et al. 2022)
   - Muestreo complejo y caro
   - Datos secundarios

2. Particiones train/validation/test
   - Número de muestras bajo
   - Necesitamos todos los datos para ajustar el modelo

3. Cross-validación

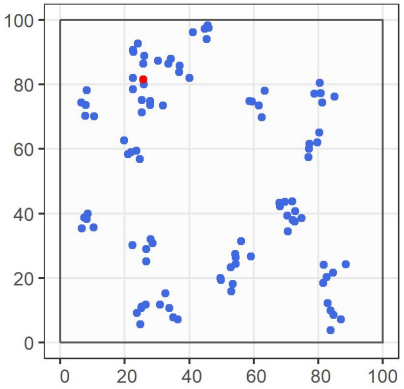
Estación de calidad del aire en Barcelona. Fuente: ASPB

Técnicas de cross-validación tradicionales

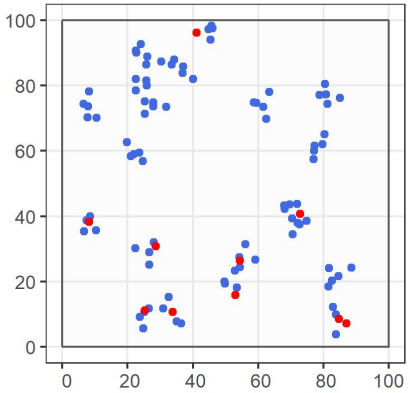
jornadassiglibre
Geotech/spatial data science

Métodos estándar en machine learning

Leave-one-out (LOO)



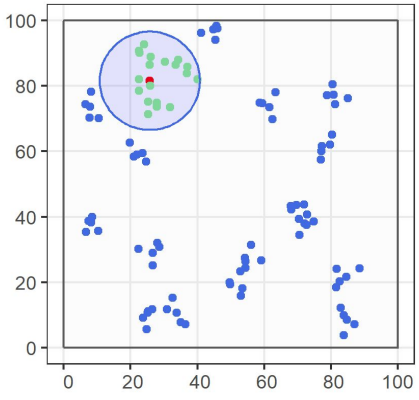
Random k-fold



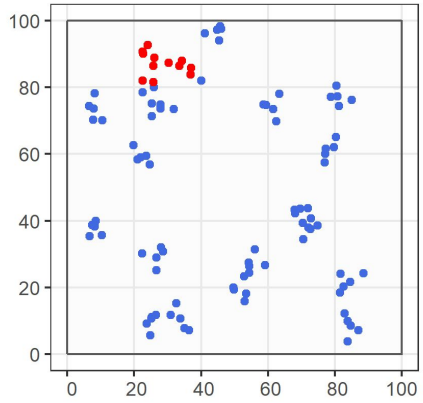
- **Asumen** independencia entre **train** y **test** (Roberts et al 2017)
- Buena estimación para muestras aleatorias, demasiado optimista para muestras en clúster (Milà et al 2022, Linnenbrink et al 2023, Wadoux et al 2022)

Métodos espaciales

bLOO



block CV

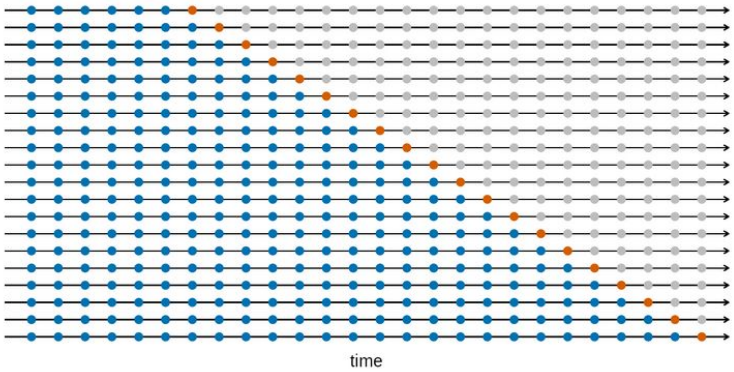


- **Imponen** independencia entre **train** y **test** (Roberts 2017)
- bLOO: Extensión de LOO con **buffer de exclusión**
- block CV: Extensión de k-fold por bloques espaciales
- ¿Tamaño del buffer/bloque?
- Evaluación demasiado pesimista (Milà et al 2022, Linnenbrink et al 2023, Wadoux et al 2022)

Cross-validación basada en el objetivo de la predicción

jornadassiglibre
Geotech/spatial data science

- Todos los métodos tradicionales tienen limitaciones. Investigación muy activa.
- **Idea clave:** La cross-validación debe reflejar las condiciones predictivas que se encontrarán al utilizar el modelo para un determinado objetivo.
- En series temporales (objetivo: predicción a t+1):

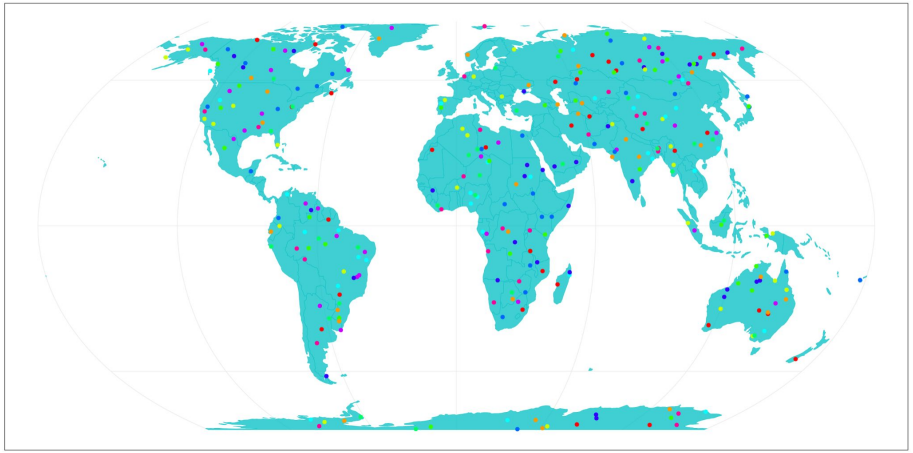


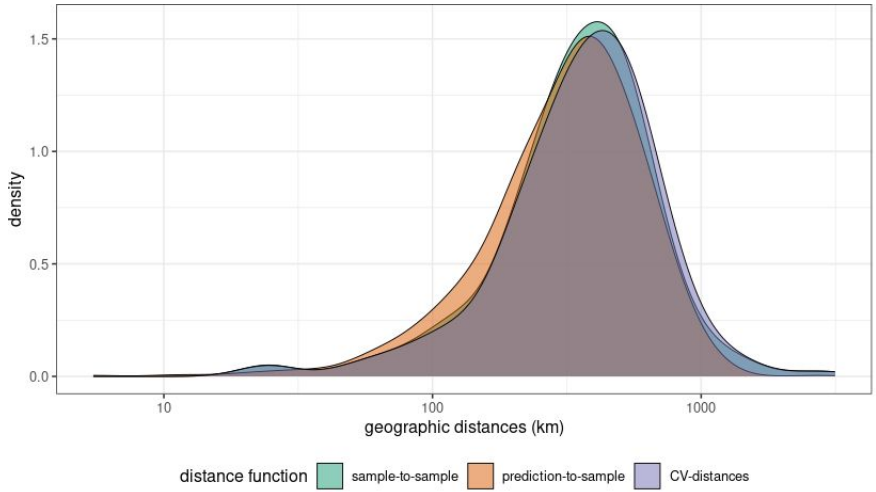
Nos fijamos en la distancia (temporal) entre la observación a predecir y nuestros datos y la reproducimos durante la cross-validación

Fuente: Hyndman and Athanasopoulos

- ¿Cómo aplicar estas ideas en espacio?

Cross-validación basada en el objetivo de la predicción





Fuente: https://hannameyer.github.io/CAST/articles/cast02-plotgeodist.html

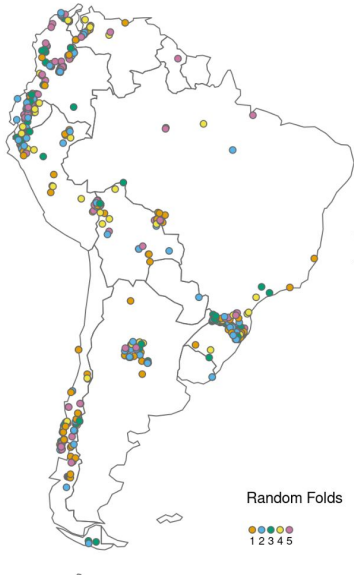
Objetivo y estrategia de validación: Predecir a escala global (malla regular) con muestras aleatorias, random 10-fold

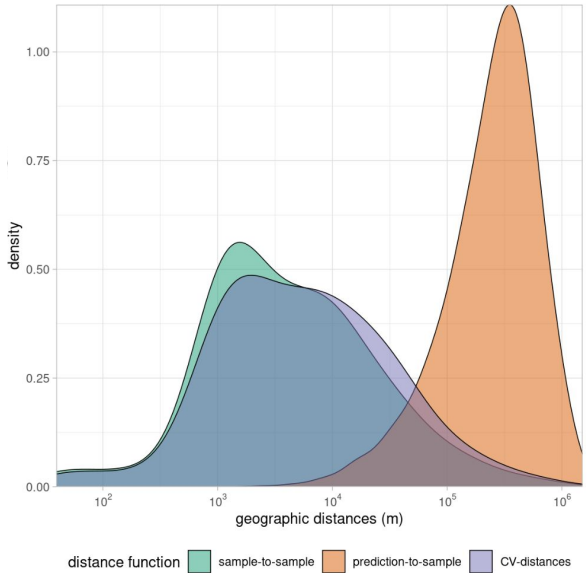
Condiciones predictivas: Distancias geográficas *nearest neighbour* entre puntos de predicción (toda la malla) y muestras.

Condiciones durante cross-validación: Distancias geográficas *nearest neighbour* entre puntos *test* y *train*.

Resultado: Para muestras aleatorias con random k-fold, la **distribución de distancias durante la validación cruzada** se aproxima a la **distribución de distancias durante la predicción.** Esto lleva a **evaluaciones correctas** (Linnenbrink et al 2023, Wadoux et al 2022).

Cross-validación basada en el objetivo de la predicción





Fuente: arXiv:2404.06978

Objetivo y estrategia de validación: Predecir a escala continental (malla regular) con muestras en cluster, random 5-fold

Condiciones predictivas: Distancias geográficas nearest neighbour entre puntos de predicción (toda la malla) y muestras.

Condiciones durante cross-validación: Distancias geográficas *nearest neighbour* entre puntos *test* y *train*.

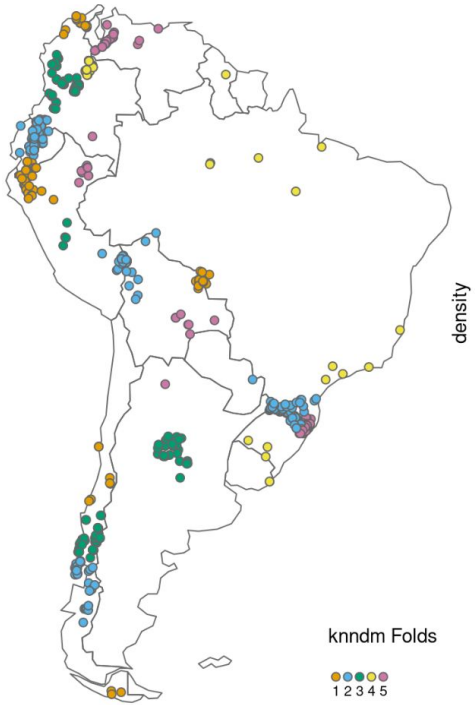
Resultado: Para muestras cluster con random k-fold, las **distancias durante la validación cruzada** son más cortas que **las distancias durante la predicción.** Esto lleva a **evaluaciones demasiado optimistas** (Linnenbrink et al 2023, Wadoux et al 2022).

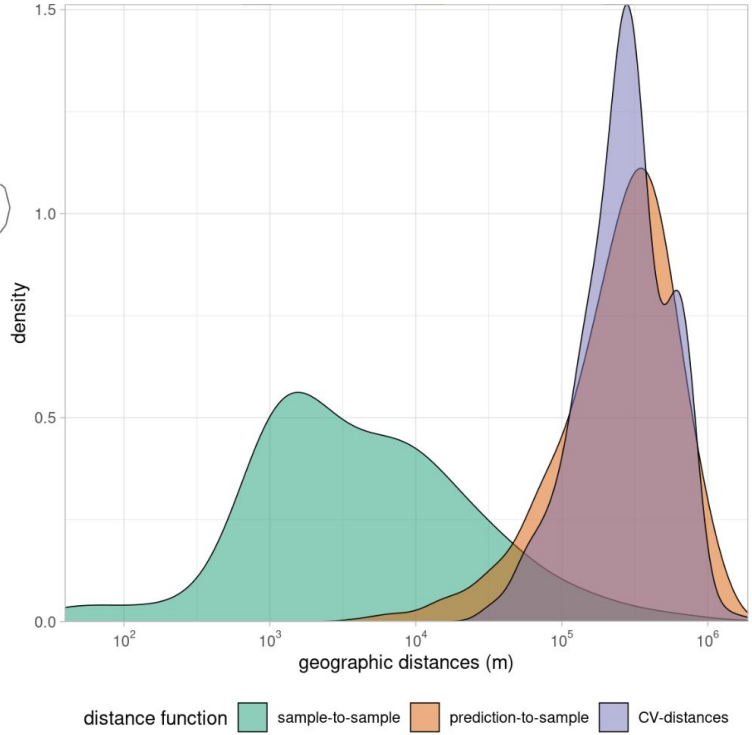
Método kNNDM para cross-validación espacial

jornadassiglibre
Geotech/spatial data science

kNNDM: k-fold Nearest Neighbour Distance Matching

Propone una configuración de validación cruzada espacial cuya **distribución de distancias durante la cross-validación** aproxima la **distribución de distancias durante la predicción**





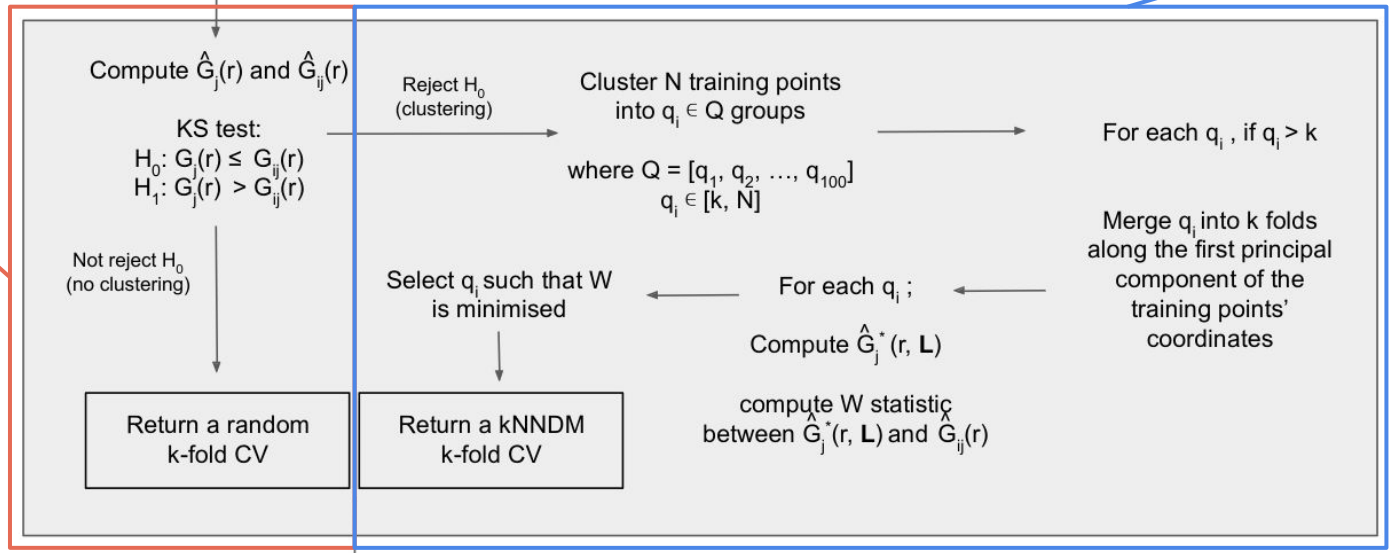
Fuente: Meyer et al 2024

Método kNNDM para cross-validación espacial

jornadassiglibre
Geotech/spatial data science

Si no se detecta clustering de las muestras, generaliza a random k-fold

Si se detecta clustering, se aplica algoritmo

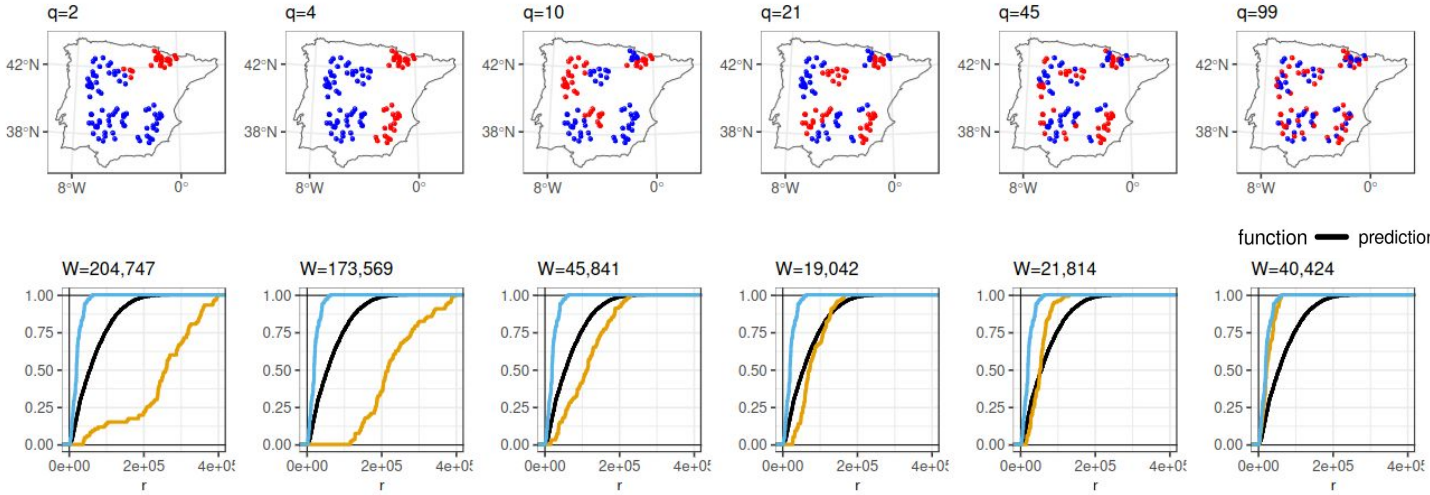


Simulaciones con evaluación del método:
https://doi.org/10.5194/egusphere-2023-1308

Método kNNDM para cross-validación espacial

jornadassiglibre
Geotech/spatial data science

Ejemplo: k=2

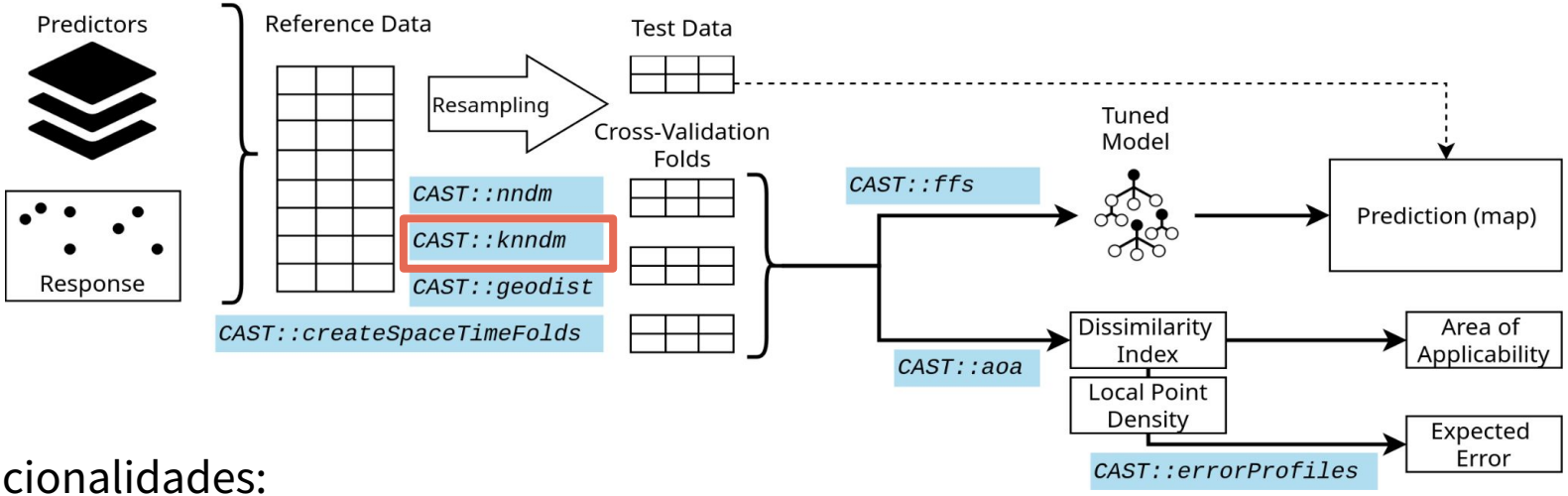


$$W = \int |\hat{G}_j^*(r, \mathbf{L}) - \hat{G}_{ij}(r)| \mathrm{d}r$$

Fuente: Linnenbrink et al 2023

1. Agrupamos observaciones utilizando algoritmos de clusterización con diferente número de clusters ***q*** que agrupamos en ***k*** folds.
2. Calculamos la función de distribución (ECDF) de las **distancias entre los puntos de predicción y muestra.**
3. Calculamos la ECDF de las **distancias entre los puntos de *test* y *train* para cada configuración.**
4. Seleccionamos la configuración que aproxime mejor la distribución de distancias de predicción con ***W***.

Presentación del paquete de R *CAST*



Fuente: Meyer et al 2024

Funcionalidades:

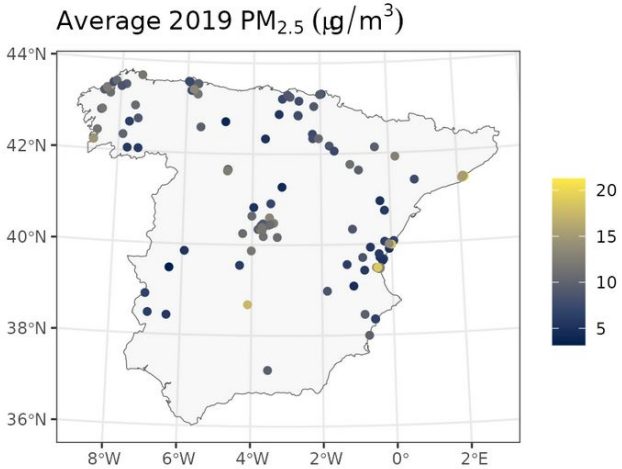
- **Validación cruzada espacial**
- Validación cruzada espacio-temporal
- Selección de variables espacial
- Análisis de extrapolación (*area of applicability*)
- Estimación de error a nivel de píxel

Website: https://hannameyer.github.io/CAST/index.html

Ejemplo

jornadassiglibre
Geotech/spatial data science

Variable respuesta



Predictores (simplificado)

- Densidad de población (Geostat)
- Densidad de carreteras (OSM)
- Superfícies impermeables (Copernicus)
- Luz nocturna (VIIRS)

Algoritmo: Random Forest

Validación: kNNDM 5-fold

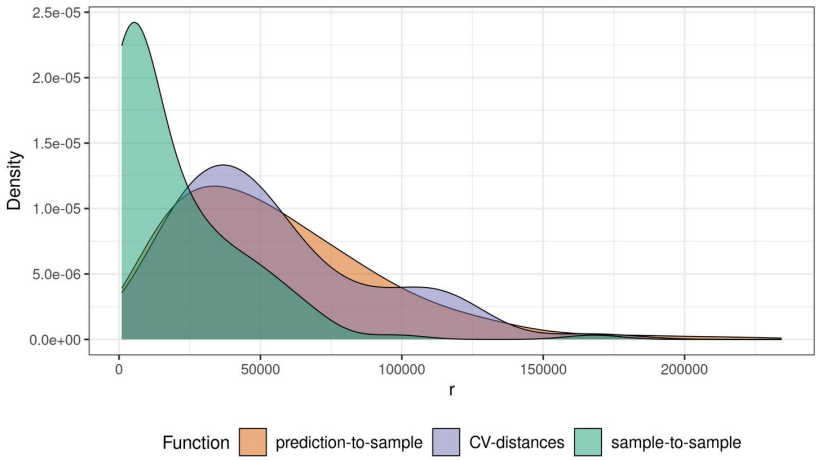
Fuente:
https://hannameyer.github.io/CAST/articles/cast03-CV.html

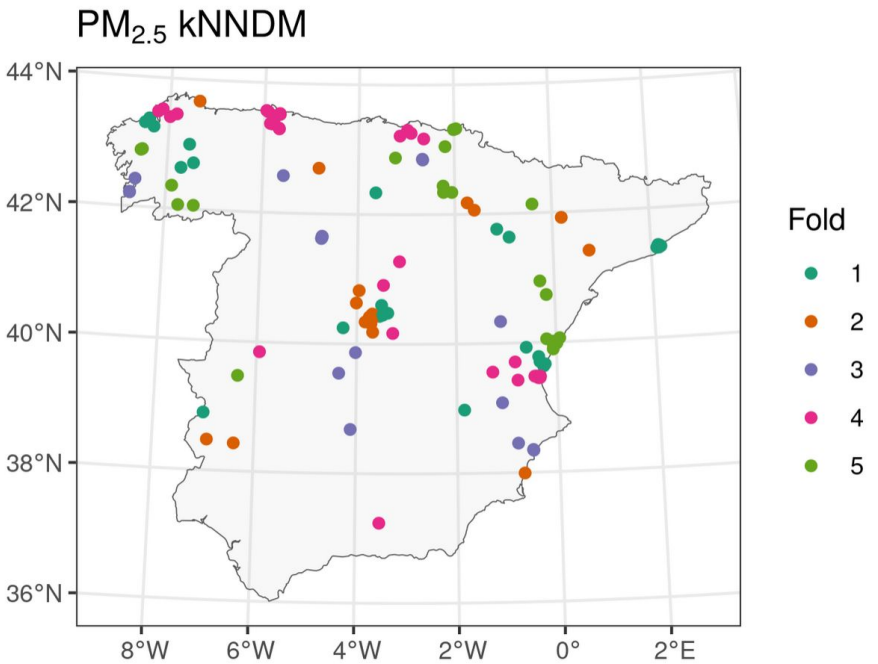
Ejemplo


```r
pm25_knndm <- knndm(pm25, k = 5, modeldomain = spain)
print(pm25_knndm)
```

```
## knndm object
## Space: geographical
## Clustering algorithm: hierarchical
## Intermediate clusters (q): 46
## W statistic: 4919.5574
## Number of folds: 5
## Observations in each fold:  29 22 18 32 23
```

```r
plot(pm25_knndm, type = "simple", stat = "density")
```





Ejemplo

jornadassiglibre
Geotech/spatial data science

```
# kNNDM 5-fold CV
pm25_knndm_ctrl <- trainControl(method="cv",
                                index=pm25_knndm$indx_train,
                                savePredictions=TRUE)
pm25_knndm_mod <- train(pm25_df[c("popdens", "primaryroads", "ntl", "imd")],
                        pm25_df[,"PM25"],
                        method="rf", importance=FALSE,
                        trControl=pm25_knndm_ctrl, ntree=100, tuneLength=1)
pm25_knndm_res <- global_validation(pm25_knndm_mod)
pm25_knndm_res <- t(as.data.frame(pm25_knndm_res))
kable(pm25_knndm_res, digits = 2, row.names = FALSE)
```

RMSE	Rsquared	MAE
3.26	0.19	2.47

Recursos y bibliografía

jornadassiglibre
Geotech/spatial data science

Métodos NNDM:

Milà, C., Mateu, J., Pebesma, E., & Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods in Ecology and Evolution*, *13*(6), 1304-1316.

Linnenbrink, J., Milà, C., Ludwig, M., & Meyer, H. (2023). kNNDM: k-fold Nearest Neighbour Distance Matching Cross-Validation for map accuracy estimation. *EGUsphere*, *2023*, 1-16.

Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, *13*(1), 2208.

Paquete de R *CAST*:

Meyer, H., Ludwig, M., Milà, C., Linnenbrink, J., & Schumacher, F. (2024). The CAST package for training and assessment of spatial prediction models in R. *arXiv preprint arXiv:2404.06978*.

https://hannameyer.github.io/CAST/index.html

Otra literatura sobre validación cruzada espacial:

Wadoux, A. M. C., Heuvelink, G. B., De Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913-929.

Milà, C., Ludwig, M., Pebesma, E., Tonne, C., & Meyer, H. (2024). Random forests with spatial proxies for environmental modelling: opportunities and pitfalls. EGUsphere, 2024, 1-30.

jornadassiglibre

Geotech/spatial data science

Muchas gracias por vuestra atención!

carles.mila@isglobal.org carles.mila@gmail.com https://github.com/carlesmila