Chapter 10: Spatial cross-validation for GeoAI

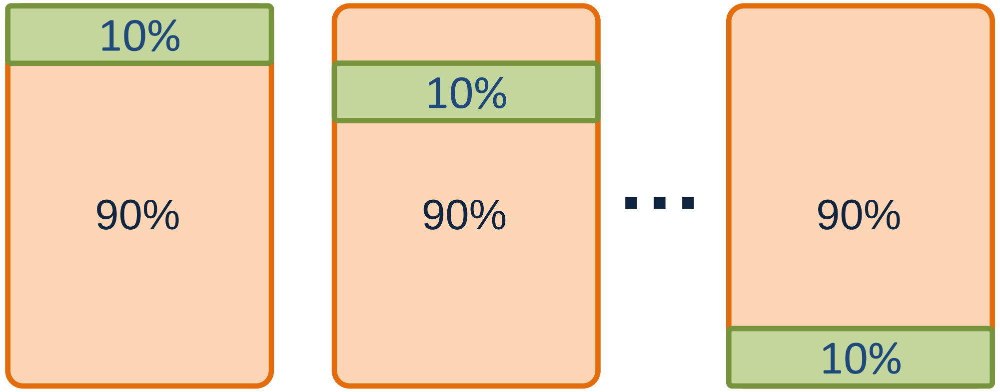
Kai Sun, Yingjie Hu, Gaurish Lakhanpal, and Ryan Zhenqi Zhou

GeoAI Lab, Department of Geography

University at Buffalo, State University of New York

Cross-validation and spatial cross-validation

- Cross-validation (CV) has been widely used in GeoAI research

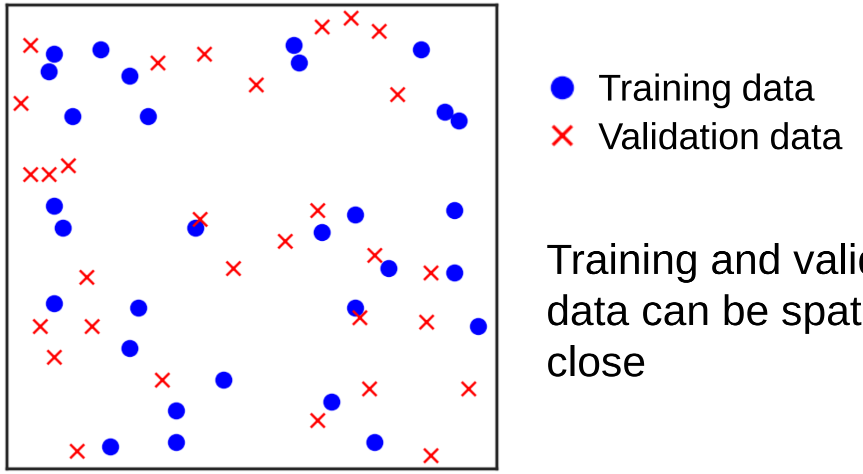


An example of 10-fold cross-validation

- While effective, CV could lead to an overestimate of model performance on geographic data

- Spatial CV is a spatially explicit CV approach that splits data spatially rather than randomly

- Different spatial CV methods exist across multiple disciplines

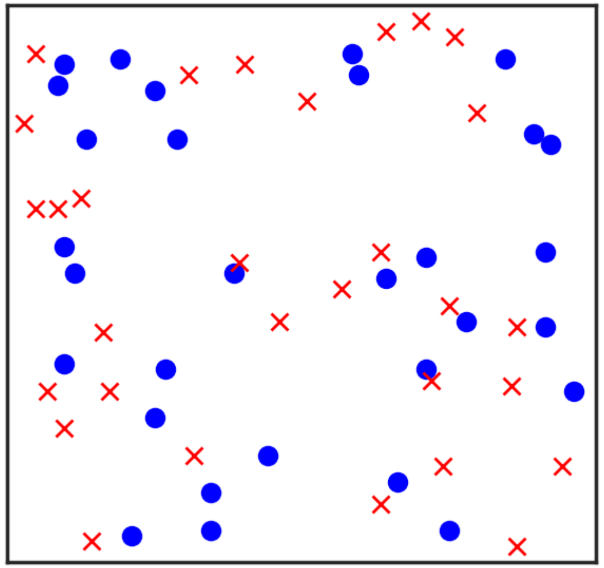


Training and validation data can be spatially close

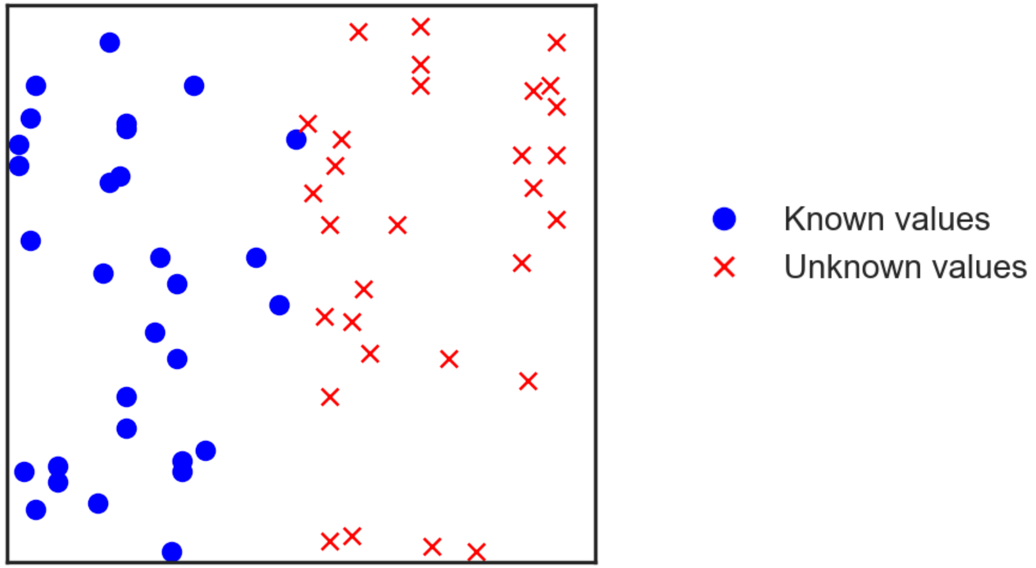
Cross-validation and spatial cross-validation

- Spatial CV does not always provide more suitable assessment of model performance than random CV

- Two common situations in GeoAI research: within-area prediction and between-area prediction (Roberts et al. 2017; Goodchild and Li 2021)



Within-area prediction (interpolation)



Between-area prediction (extrapolation)

Spatial CV methods

- Four main methods: Clustering-based spatial CV, grid-based spatial CV, geo-attribute-based spatial CV, and spatial leave-one-out CV

- Software packages for spatial CV developed by researchers

In R:

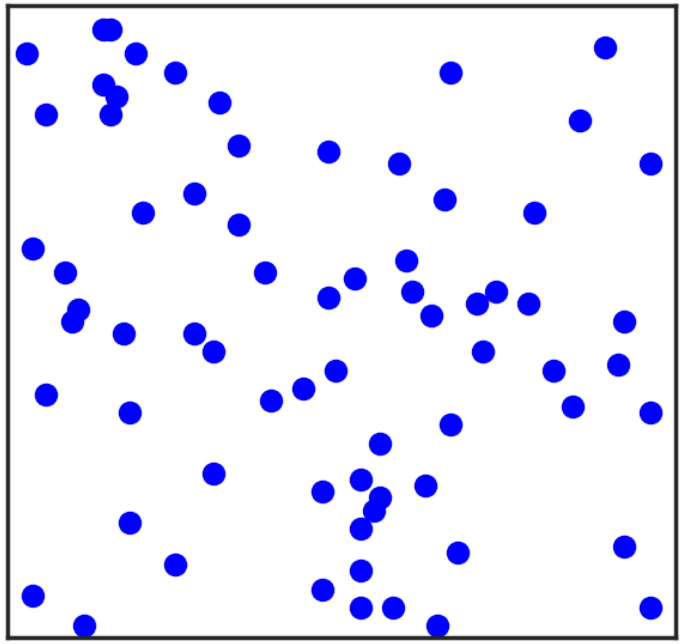
- ***sperrorest***: https://cran.r-project.org/web/packages/sperrorest/index.html (Brenning 2012)
- ***spatialsample***: https://cran.r-project.org/web/packages/spatialsample/index.html (Mahoney 2023)
- ***blockCV***: https://cran.r-project.org/web/packages/blockCV/index.html (Valavi et al. 2019)
- ***ENMeval***: https://cran.r-project.org/web/packages/ENMeval/index.html (Muscarella et al. 2014)
- ***Mlr3spatiotempcv***: https://cran.r-project.org/web/packages/mlr3spatiotempcv/index.html (Schratz et al. 2021)
- ***CAST***: https://cran.r-project.org/web/packages/CAST/index.html (Meyer et al. 2018)

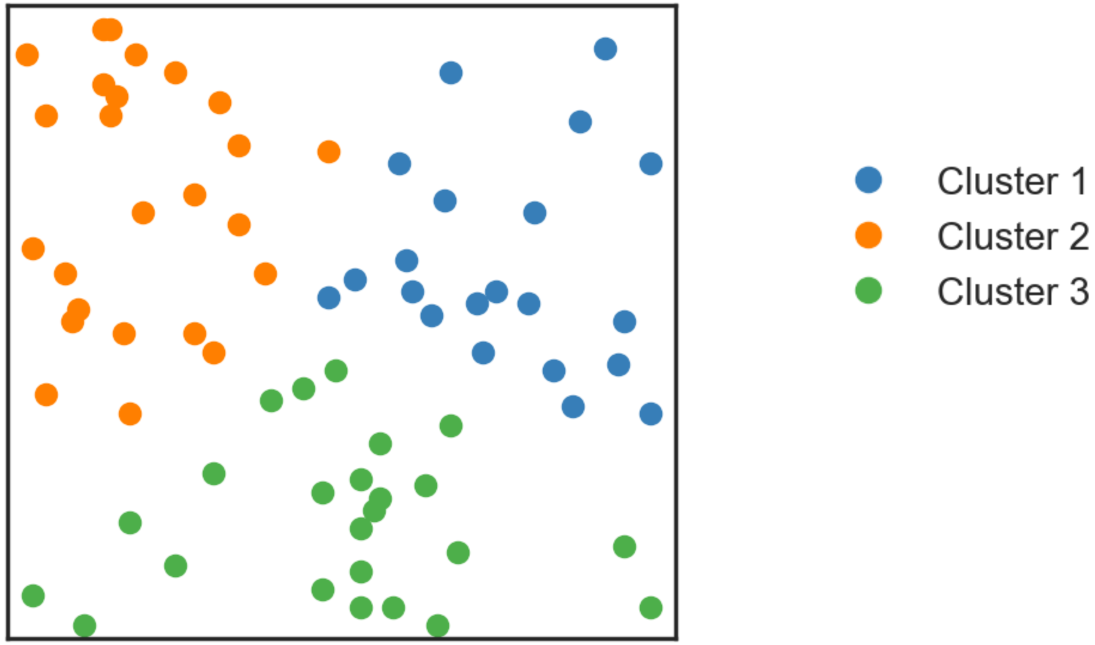
In Python:

- ***spacv***: https://github.com/SamComber/spacv (Comber 2020)
- ***MuseoToolbox***: https://museotoolbox.readthedocs.io/en/latest/ (Karasiak 2020)

Spatial CV methods

- Clustering-based spatial CV
- Using a clustering method (e.g., K-means) to spatially split data

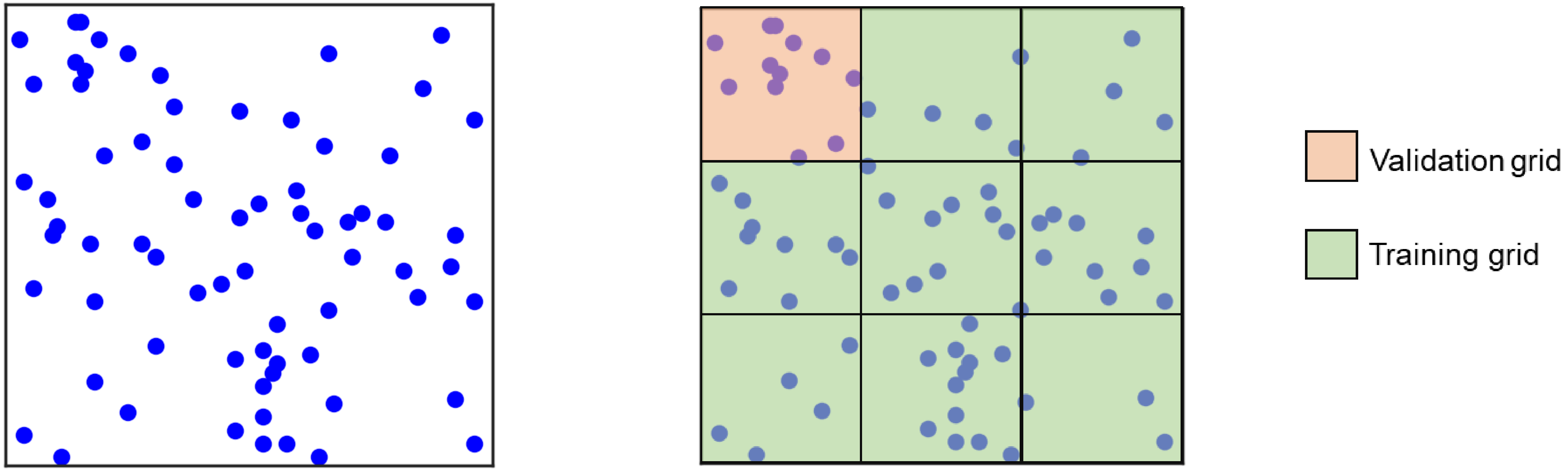




Implemented in the following packages: *spatialsample* (R), *sperrorest* (R), *Mlr3spatiotempcv* (R), *blockCV* (R), and *spacv* (Python)

Spatial CV methods

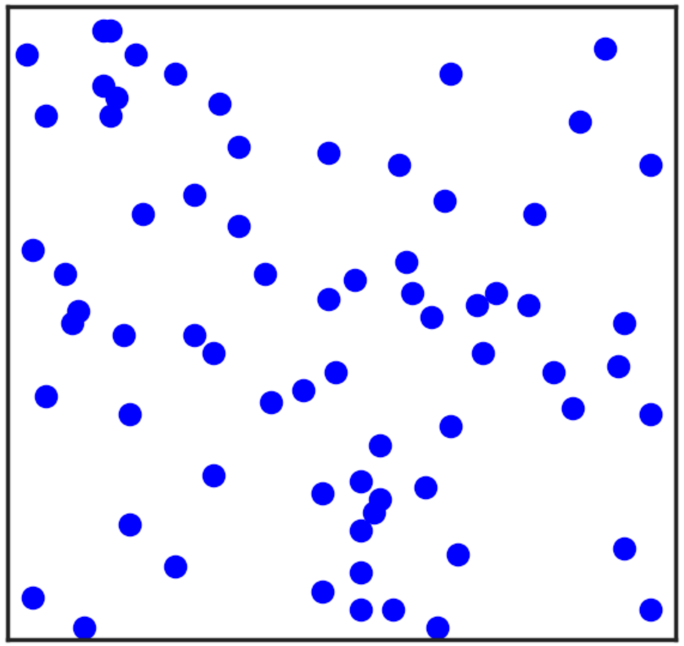
- Grid-based spatial CV
- Using a spatial grid to divide the study area into $n \times m$ grid cells

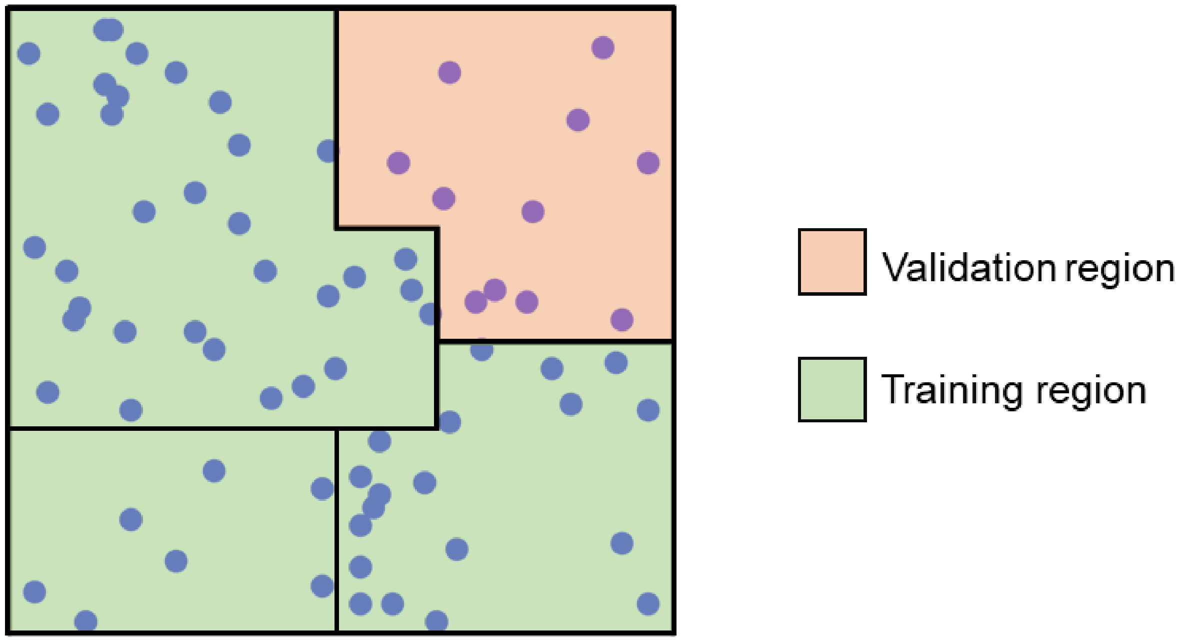


Implemented in the following packages: *ENMeval* (R), *spatialsample* (R), *sperrorest* (R), *Mlr3spatiotempcv* (R), *blockCV* (R), and *spacv* (Python).

Spatial CV methods

- Geo-attribute-based spatial CV
- Using a geo-attribute, such as county name or city district name, to spatially split data

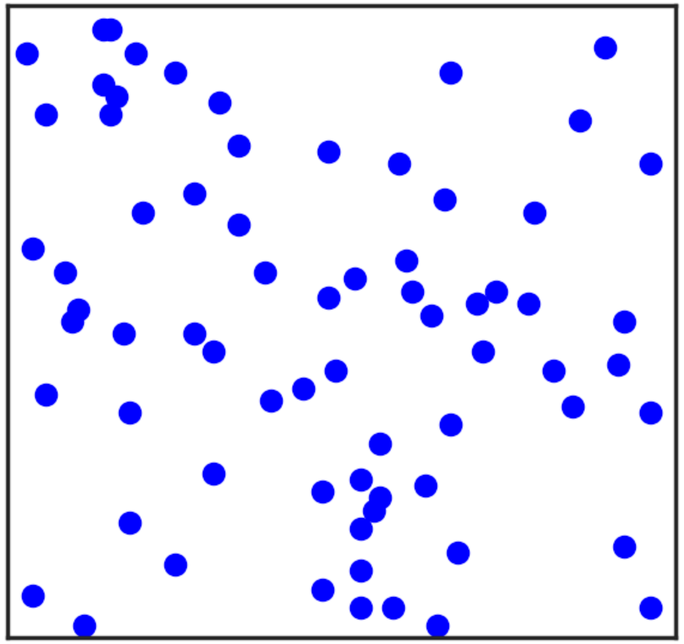


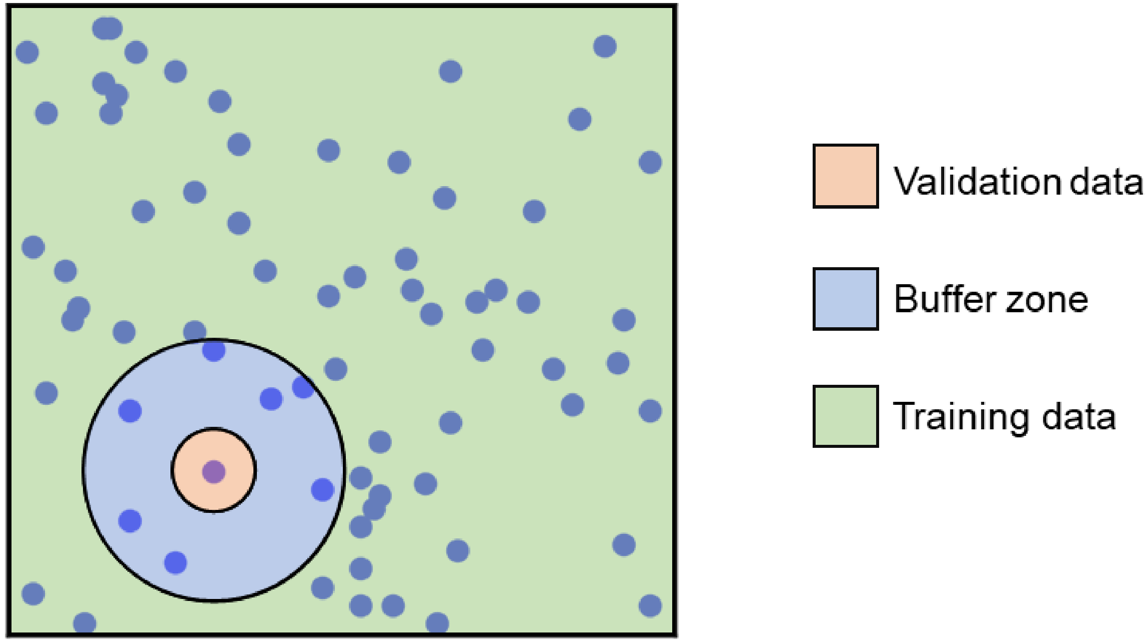


Implemented in the following packages: *Mlr3spatiotempcv* (R) and *spacv* (Python).

Spatial CV methods

- Spatial leave-one-out CV
- Using a buffer zone to spatially separate training and validation data



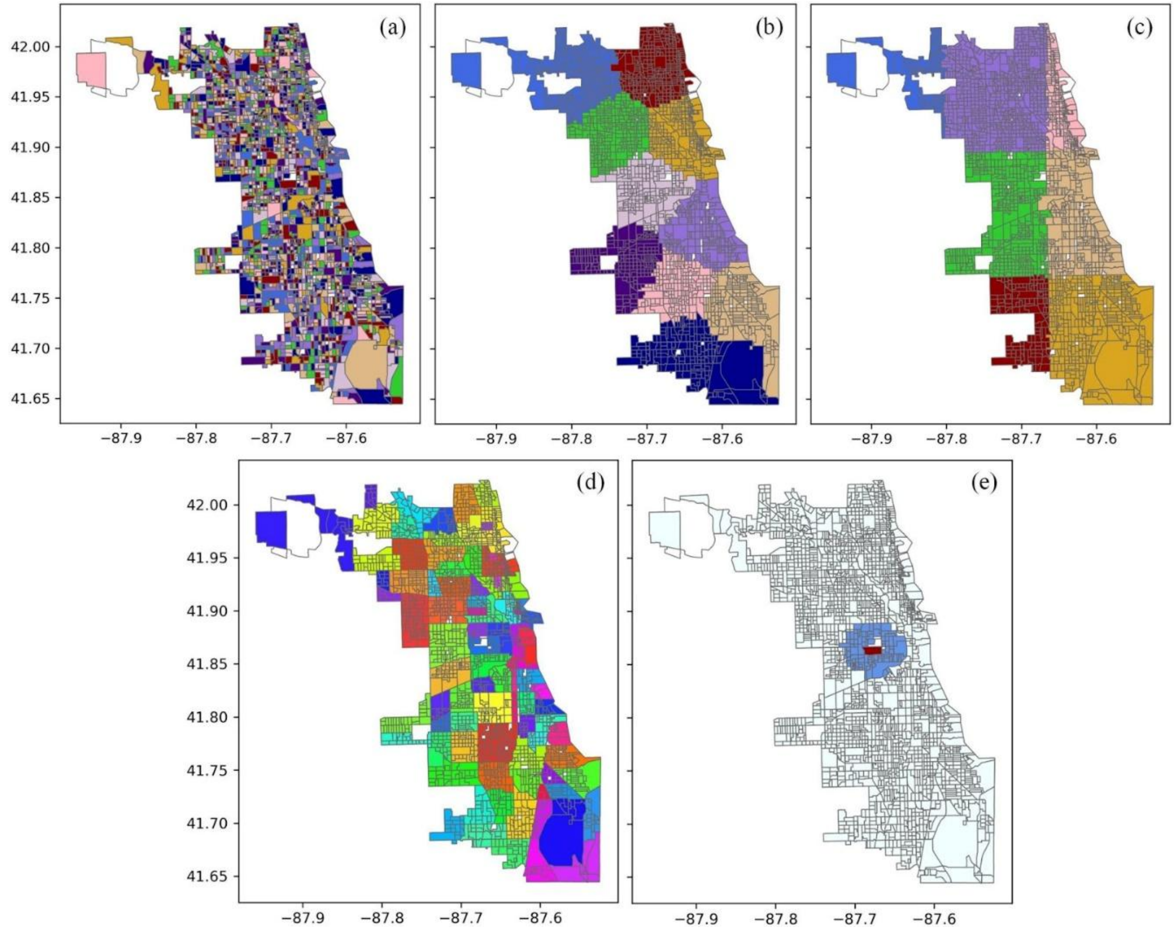


Implemented in the following packages: *CAST* (R), *Mlr3spatiotempcv* (R), *blockCV* (R), *sperrorest* (R), *spatialsample* (R), *spacv* (Python), and *museotoolbox* (Python).

Case study I

- Predicting neighborhood-level domestic violence rate in Chicago based on socioeconomic attributes using random forest

- Model is assessed under five CV:
  - Random CV
  - Clustering-based CV
  - Grid-based CV
  - Geo-attribute-based CV
  - Spatial leave-one-out CV

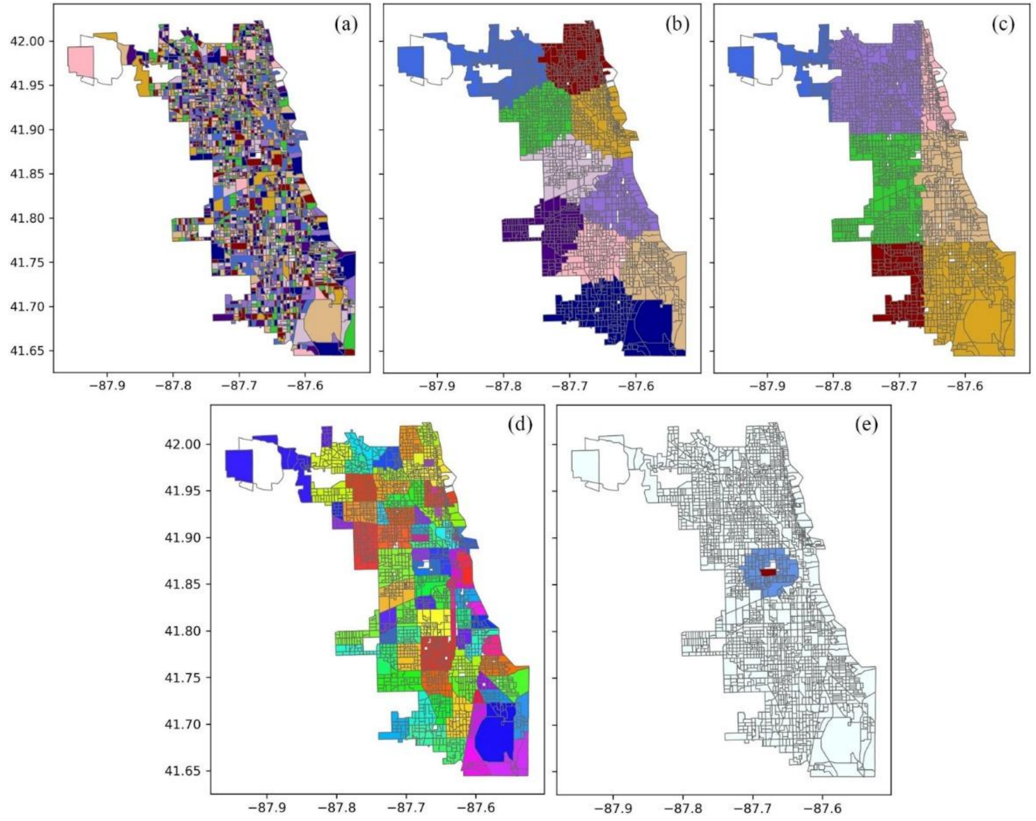


- The dataset is from a previous study (Chang et al. 2022)

Case study I

- Predicting neighborhood-level domestic violence rate in Chicago based on socioeconomic attributes using random forest

CV method	***R^2***	***RMSE***
Random CV	0.5952	8.9398
Clustering-based spatial CV	0.5443	9.4853
Grid-based spatial CV	0.5643	9.2752
Geo-attribute-based spatial CV	0.5667	9.2501
Spatial leave-one-out CV	0.5470	9.4575

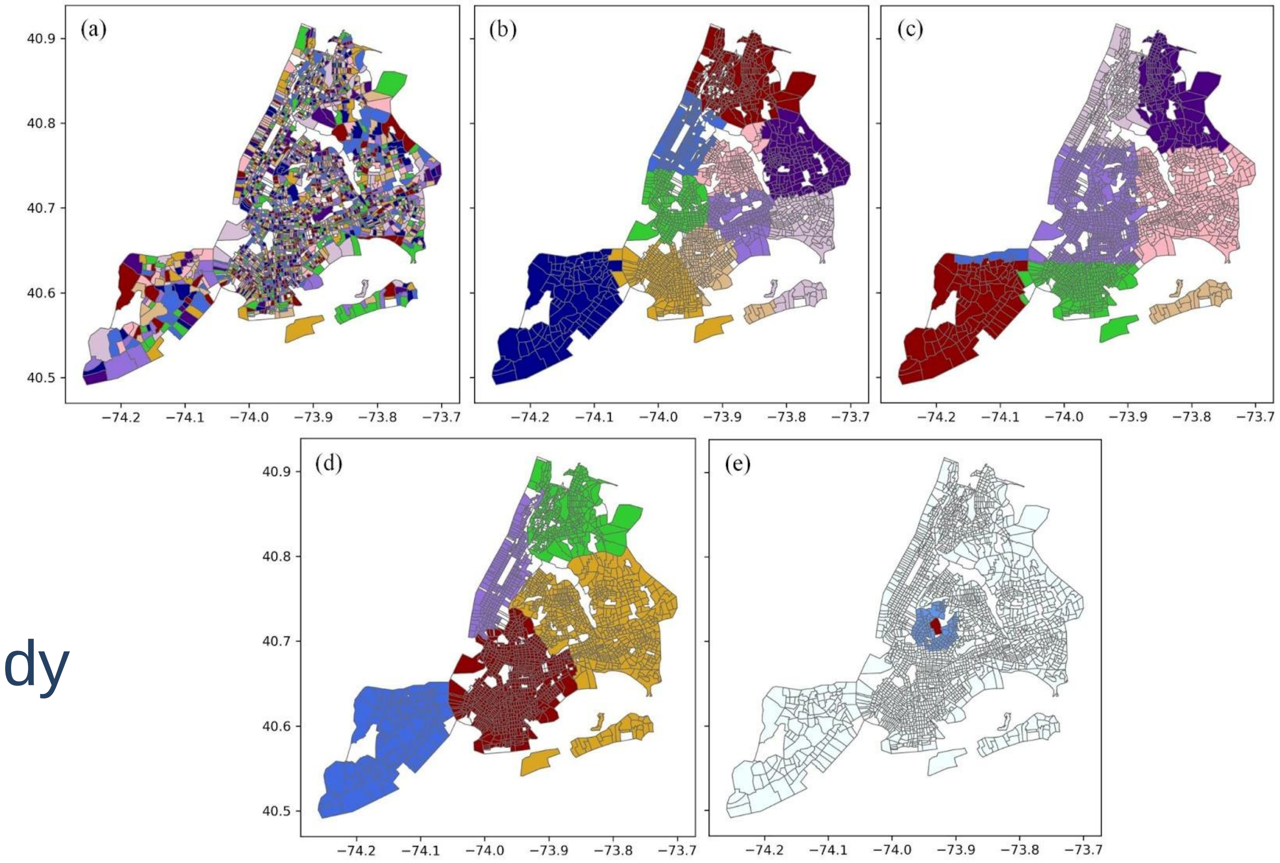


Case study II

- Predicting neighborhood-level obesity prevalence in New York City based on socioeconomic attributes using a fully-connected deep neural network

- Model is assessed under five CV:
  - Random CV
  - Clustering-based CV
  - Grid-based CV
  - Geo-attribute-based CV
  - Spatial leave-one-out CV

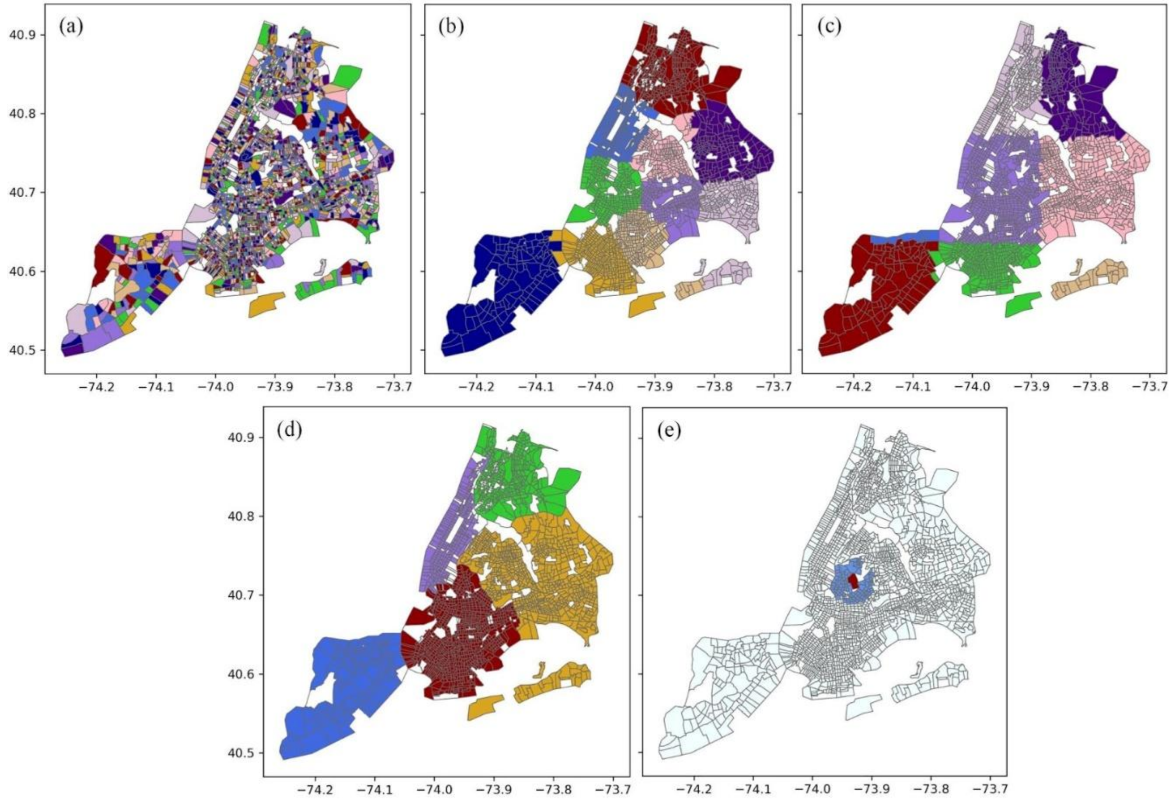
- The dataset is from a previous study (Zhou et al, 2022)



Case study II

- Predicting neighborhood-level obesity prevalence in New York City based on socioeconomic attributes using a fully-connected deep neural network

CV method	***R^2***	***RMSE***
Random CV	0.8692	2.1287
Clustering-based spatial CV	0.7244	3.0899
Grid-based spatial CV	0.7466	2.9624
Geo-attribute-based spatial CV	0.6613	3.4250
Spatial leave-one-out CV	0.8083	2.5766

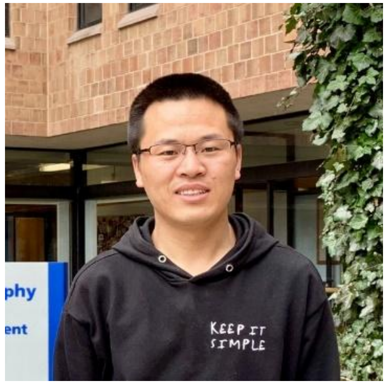


Conclusions

- Spatial CV presents a spatially explicit approach to assessing model performance by splitting data spatially rather than randomly
- Four main methods for spatial CV: Clustering-based spatial CV, grid-based spatial CV, geo-attribute-based spatial CV, and spatial leave-one-out CV
- Two case studies based on real-world data in two different U.S. cities and two different machine learning models
- Spatial CV is not always more suitable than random CV for GeoAI research; it depends on how the trained model will be used for making predictions

Acknowledgement

Student co-authors of this book chapter



Kai Sun

Ryan Zhou

Gaurish Lakhanpal

Code: https://github.com/geoai-lab/spatialCV

Thank you!

Yingjie Hu

University at Buffalo, SUNY
Email: yhu42@buffalo.edu
GeoAI@UB: https://geoai.geog.buffalo.edu