Geospatial Data Analysis and Machine Learning

Mark Asch - ML-PREP

2025

Geospatial-ML

Program

1. Introduction and Context.
2. Spatiotemporal resampling.
3. Network analysis.
4. Case-study: landslides in Ecuador

CONTEXT

1st law of Geography

The First Law of Geography, according to Waldo Tobler, is:

"everything is related to everything else, but near things are more related than distant things."

This first law is the foundation of the fundamental concepts of spatial dependence and spatial autocorrelation and is utilized specifically for the inverse distance weighting method for spatial interpolation and to support the regionalized variable theory for kriging. The first law of geography is the fundamental assumption used in all spatial analysis.

What is Geostatistics?

- Most properties of the environment, such as rainfall, plant nutrients in the soil, geological properties, pollutant distribution, epidemiology, are measured effectively at points between which there are large gaps.

- The environment is continuous, however, and environmental scientists typically want to know the values of those properties between the points, in the gaps; they want to predict in a spatial sense from their data, taking into account the locations of their observations.

- Geostatistics is based on the study of random fields, built upon stochastic processes that are themselves a generalization of random variables.

Geostatistics

Geostatistics is a branch of statistics for spatial or spatiotemporal datasets, in which the data consist of a finite sample of measured values, relating to an underlying spatially continuous phenomenon.

GEOSTATISTICS
and
MACHINE LEARNING

Geostatistics and Machine Learning

- Objective: Combine the predictive power of machine learning with the accuracy of geostatistics

  $\Rightarrow$ The capacity of geostatistical techniques, such as kriging or simulation, to incorporate secondary data is one of their main advantages.

  $\Rightarrow$ More accurate predictions and classifications are made possible by supervised (and unsupervised) machine learning methods such as support vector machines, neural networks, and random forests, which are excellent at capturing intricate non-linear correlations.

"A strong foundation for comprehending geographical correlations and variability is offered by geostatistics, while machine learning offers the adaptability needed to represent complex patterns and interactions in the data. When combined, they enable more sophisticated spatial models that can adjust to the

underlying intricacies of real-world occurrences, improving resource management and decision-making in a variety of environmental contexts."

- ML methods applicable are (see Basic ML training notes):

  ⇒ Supervised: RF (Xgboost), k-NN, SVM and CART.
  ⇒ Unsupervised: k-means, PCA, clustering (Dbscan).
  ⇒ Deep learning: CNNs.

- Conclusions:

  ⇒ A strong foundation for sophisticated spatial analysis is provided by combining geostatistics, in particular the Gaussian Variogram Model, with robust machine learning algorithms like Random Forest, k-nearest Neighbors (k-NN), Support Vector Machines (SVM), and Decision Trees.
  ⇒ We can greatly improve forecast accuracy and the interpretability of geographical data by quantifying spatial dependence using the variogram and incorporating these spatial insights into machine learning models.

⇒ We can solve intricate spatial problems more precisely thanks to the synergy between geostatistics and machine learning, which also enables us to see patterns and insights that could otherwise go missed.

⇒ The future of spatial data analysis will probably be greatly influenced by the combined use of geostatistics and machine learning, both of which are topics that are still developing.

SPATIAL MODEL EVALUATION

Spatial Model Evaluation

- When using machine learning methods and protocols, our objective is to choose a model that yields good enough predictive performance - see Machine Leaning Course.

- Predictive performance is measured by some resampling strategy to ensure a fair evaluation of the model - see Resampling Lecture.

- Cross-validation is the basis of the performance evaluations for

  $\Rightarrow$ hyperparameter tuning
  $\Rightarrow$ model precision.

- However, in geospatial applications, where the First Law of Geography applies, we need to take into account the spatial dependence of the measurements, otherwise there is a strong risk to obtain a biased, and often incorrect model.

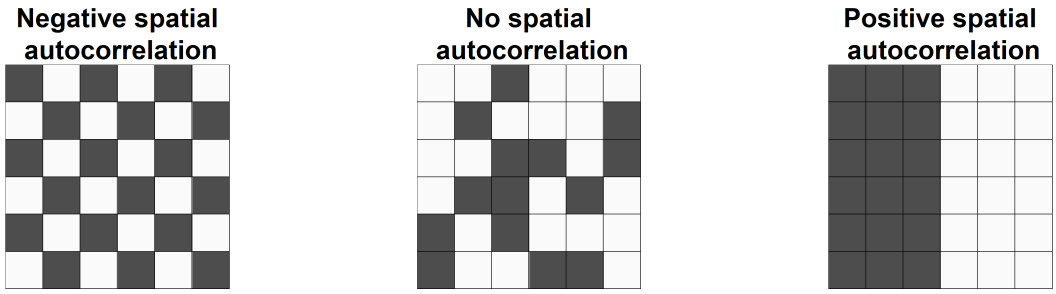
- To deal with this important issue we need:

  $\Rightarrow$ to quantify the spatial dependence (autocorrelation—see below)

  $\Rightarrow$ devise cross-validation techniques that take spatial dependence into account.

Spatial Autocorrelation

Definition 1. Spatial autocorrelation is the correlation of a variable with itself due to the spatial location of the observations.



- Spatial autocorrelation can be assessed using special indices that summarize the degree to which similar observations tend to occur near each other over the study area.

- Moran's I

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2},$$

  where

⇒ n is the number of regions,

⇒ Y_i is the observed value of the variable of interest in region i,

⇒ $\bar{Y}$ is the mean of all values,

⇒ w_{ij} are spatial weights that denote the spatial proximity between regions i and j, with $w_{ii} = 0$ and $i, j = 1, \ldots, n$.

⇒ The definition of the spatial weights depends on the variable of study and the specific setting.

- Moran's I values usually range from –1 to 1.

  ⇒ Moran's I values significantly above $E[I] = -1/(n-1)$ indicate positive spatial autocorrelation or clustering. This occurs when neighboring regions tend to have similar values.

  ⇒ Moran's I values significantly below $E[I]$ indicate negative spatial autocorrelation or dispersion. This happens when regions that are close to one another tend to have different values.

  ⇒ Finally, Moran's I values around $E[I]$ indicate randomness, that is, absence of spatial pattern.

- Other indices:

$\Rightarrow$ Local Moran's I
$\Rightarrow$ Geary's C
$\Rightarrow$ LISA - Local Indicator of Spatial Autocorrelation

- Testing for autocorrelation: one can use hypothesis testing (no spatial AC vs. spatial AC) and use a test statistic based on the assumption that I is normally distributed, which is valid for a large number of regions. See this reference, or Wikle's book [1].

CV for Geospatial Data

- Choosing the right cross-validation object is a crucial part of fitting a model properly.

  ⇒ There are many ways to split data into training and test sets in order to avoid model overfitting, to standardize the number of groups in test sets, etc.

  ⇒ For spatial, temporal and spatiotemporal data, we have to take great care that the resampling respects the underlying spatial and/or temporal distributions, and autocorrelations:

  → geospatial data are NOT independent, identically distributed (i.i.d.)

  → random sampling is NOT valid in this case!

  ⇒ Folds are defined using/respecting spatial boundaries and the underlying clusters.

- Large performance differences can be observed between the bias-reduced (spatial cross-validation) and overoptimistic (non-spatial, conventional cross-validation) cross-validation settings.

$\Rightarrow$ Hence the importance of accounting for the influence of spatial autocorrelation.

WARNING

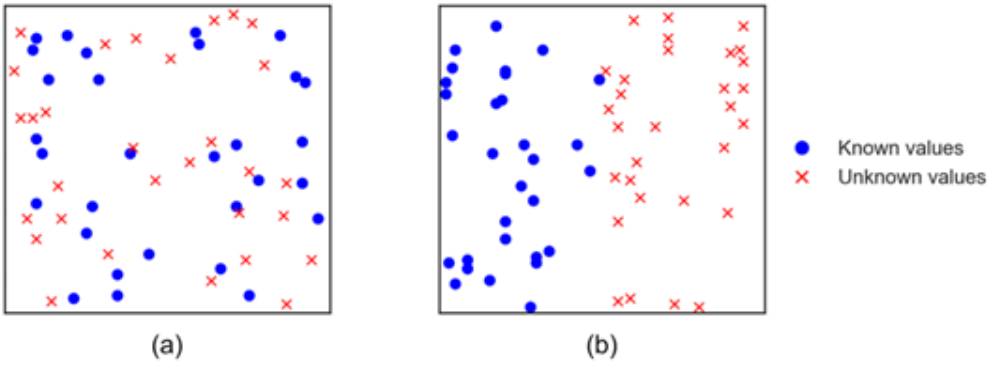
Overoptimistic performance estimates may lead to false actions in environmental decision-making based on biased model predictions.

Choice of CV Method

- Choosing the right cross-validation technique is crucial in building reliable machine learning models.
- The choice depends heavily on the specific characteristics of the dataset and the type of machine learning task at hand.
- Different techniques are designed to handle various challenges like imbalanced data, grouped data structures, spatiotemporal data.
- There is a very nice Example of the following methods on the scikit-learn website.

Type of CV	Usage	Description	When to Use
Standard K-Fold Cross-Validation	Both regression and classification	Splits the dataset into k equal-sized folds. Each fold is used once as a test set.	Best for balanced datasets to ensure comprehensive model evaluation.
Stratified K-Fold Cross-Validation	Primarily classification	Maintains the same proportion of class labels in each fold as the original dataset.	Classification tasks with imbalanced classes to maintain group proportions.
Leave-One-Out Cross-Validation (LOOCV)	Both regression and classification	Each data point is used once as a test set, with the rest as training.	Small datasets to maximize training data, though computationally intensive.
Group K-Fold Cross-Validation	Both regression and classification with groups	Ensures no group is in both training and test sets, which is useful when data points are not independent.	Datasets with logical groupings to test performance on independent groups.
Stratified Group K-Fold Cross-Validation	Primarily classification with grouped data	Combines stratification and group integrity, ensuring that groups are not split across folds.	Grouped and imbalanced datasets to maintain both class and group integrity.

Spatial CV - principles



- Basic distinction:

  ⇒ Randomized CV should be used for within-area prediction or interpolation (a).
  ⇒ Spatial CV should be used for between-area prediction or interpolation (b).

- 4 types of spatial-CV methods:

  ⇒ cluster-based
  ⇒ grid-based
  ⇒ geo-attribute based
  ⇒ spatial Leave-One-Out (LOO)

- Principle of the 4 methods:

  $\Rightarrow$ Common: perform a spatial split of training and validation data.

  $\Rightarrow$ Difference: how they split the data.

Spatial CV - in practice

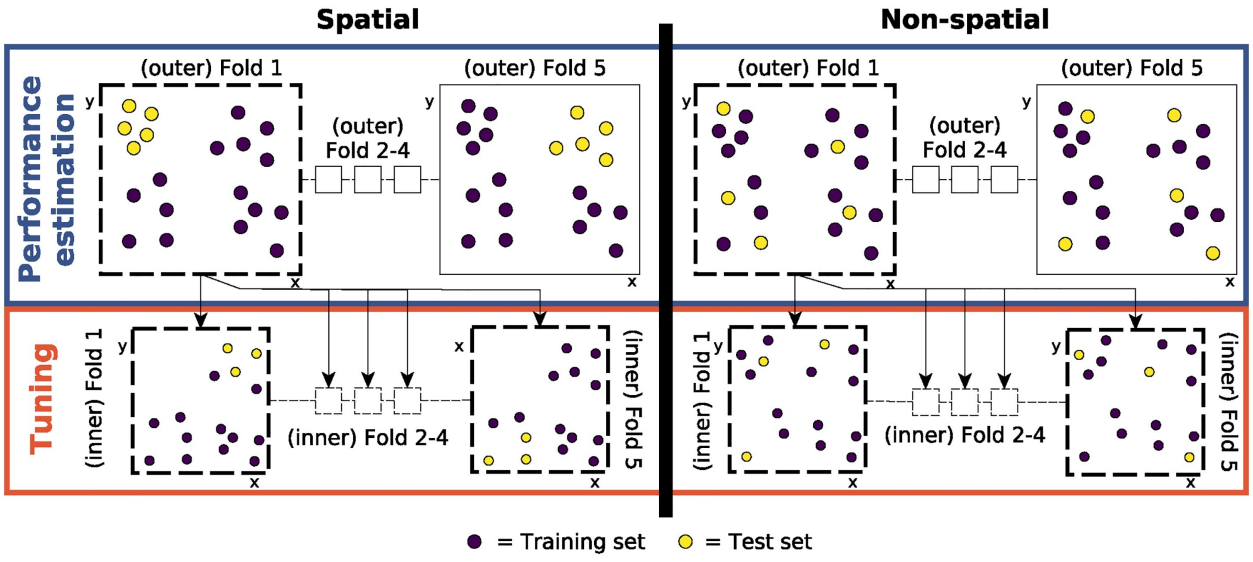
- To account for spatial autocorrelation, k-means clustering can be used before resampling:

  $\Rightarrow$ k-means partitions the dataset into spatially contiguous clusters

  $\Rightarrow$ perform standard CV on each cluster

  $\Rightarrow$ compute and report average performance over the clusters

- When tuning is also performed, a spatial nested k-fold CV approach should be used—see Advanced ML lecture.



Packages for Spatial CV

- The more sophisticated packages are available in R language, though there are some simpler ones available in python.

- They all execute some variations of scikit-learn's GroupKFold, or StratifiedGroupKFold - see this Example.

- Python packages for spatial CV:

  ⇒ museo
  ⇒ spacv
  ⇒ spatial-kfold

- R packages for spatial CV:

  ⇒ mlr3 - online book, spatial CV package - a general ML package with a spatial CV module.
  ⇒ CAST - caret framework.
  ⇒ spatialsample - tidymodels framework.
  ⇒ blockCV and on CRAN

GLM

- Generalized Linear Models are a flexible generalization of ordinary linear regression.

- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

- Generalized linear models cover diverse situations

  $\Rightarrow$ by allowing for response variables that have arbitrary distributions (rather than simply normal distributions),

  $\Rightarrow$ and for an arbitrary function of the response variable (the link function) to vary linearly with the predictors (rather than assuming that the response itself must vary linearly)

USE-CASES

Ecuador Landslide

- Full details are available on the Geocomputation with R site.
- Objective: implement spatial CV on a machine learning model.
- Uses a glm classifier.

Domestic Violence

- Adapted from this original.
- Does not require any specialized packages.
- Uses a random forest regressor.

Obesity Prevalence

- Adapted from this original.
- Does not require any specialized packages.
- Uses a feedforward neural network regressor.

Housing Values

- Adapted from this original.
- Requires the spatialkfold package.
- Simple linear regression.

References

1. C. Wikle, A. Zammit-Mangion, N. Cressie. *Spatio-Temporal Statistics with R*. CRC Press. 2019. https://spacetimewithr.org/

2. INSEE. *Handbook of Spatial Analysis: Theory and practical application with R*. 2018. https://www.insee.fr/en/information/3635545

There are a number of web-books available, mostly based on the R language (for the moment):

1. Geocomputation with R - see Chapter 12 for machine learning case-study on landslides. The python version is much more restricted.

2. MLR3 - this is a general applied ML modelling book - see Chapter 13 for spatial analysis.

3. Spatial sampling and resampling - based on MLR3.

4. Public Policy Analytics - essentially for policy deciders (social scientists, city planners, etc.) - treats the important subject of risk.

Some very nice python courses are available online:

1. Geographic Data Science by D. Arribas-Bel. - a basic course on spatial data analysis, using python, with social-sciences and urban analysis use-cases.

2. Two good courses of the University of Helsinki:
   (a) GeoPython - a very basic introduction to python programming for geodata.
   (b) Automating GIS - a more complete course on GIS processing in python.