**Performance evaluation for forecasting modeling with spatiotemporal structures in data**

Yujun Zhou and Kathy Baylis

**Abstract**

The growth in available geospatial data, along with the rise of machine learning methods, have let themselves to numerous spatial-temporal forecasting applications to solve real-world problems such as deforestation, pollution, and food security. Choosing the right performance evaluation matters for generating accurate and trustworthy out-of-sample predictions. However, with spatial-temporal dependencies between observations in both the training and testing data, the independence assumption of the testing set is violated. As a result, model performance evaluated using cross-validation (CV), and out-of-sample (OOS) can be over-optimistic. In this study, we show the changes in CV and OOS performance when we adjust for different types of spatiotemporal correlations in both simulated data and real-world panel data. We also show how the model selection is affected by the performance evaluation process to prefer overfitting models. Lastly, we propose and compare solutions such as blocking and clustering to improve performance evaluation procedures in both simulated and real-world data with spatiotemporal structures.

**JEL classifications: C33, C52, C53**

**Keywords:**  Machine-learning, Spatial autocorrelation, Spatial cross-validation

**Performance evaluation for forecasting modeling with spatiotemporal structures in data**

1. **Introduction**

The growth in available geospatial data, along with the rise of machine learning methods, have let themselves to numerous spatial-temporal forecasting applications. From predicting food security to predicting deforestation to forecasting weather or pollution, prediction accuracy is vital for the usability of these models. While the spatial and temporal correlations in panel data are frequently considered in econometrics and addressed through methods such as autocorrelation correction or spatial clustering, the interactions between spatial and temporal clustering and standard machine learning methods are less well understood. Baltagi et al. (2012, 2014) demonstrate the bias of estimators when they ignore the various spatial error structure in forecasting using simulated data in a panel setting and suggest unbiased estimators for each case. In the practice of machine learning, though, without knowing the true data generating process, we need to properly define the testing data for the model to be evaluated prior to the modeling phase. Choosing the right performance evaluation matters for generating accurate and trustworthy predictions, especially when the spatiotemporal correlations would affect the model training and model evaluation.

Two primary evaluation methods are cross-validation and Out-of-sample evaluation. Cross-validation (CV) is a resampling-based technique for the estimation of a model’s predictive performance. K-fold CV, for example, divides the data into K subsets of approximately the same size and then having each subset used successively as the test set. Out-of-sample (OOS) evaluation is to estimate the performance of the model in “unseen” data, i.e., data not used for in training the model. The standard way of doing this is called “hold-out validation,” namely splitting the data into training for the model, but retaining a portion of the data untouched, held out for evaluation. This is similar to out of sample forecasting in econometrics. However, with spatial-temporal dependencies between observations in both training and testing data, model performance evaluated using CV and OOS can be problematic. For example, if temporal autocorrelation were 0.9, a simple model that predicted the out-of-sample next period to be the same as the current period would have 90% accuracy, but one might be very wary about its predictive ability in a different setting. One could envision a similar problem for spatial prediction in a setting with a very high spatial correlation. In both cases, the procedures random resampling of the training data or random splitting into the testing data are no longer random. In other words, the observations in the testing set are no longer independent from the training set with the spatiotemporal correlations (Oliveira et al., 2018).

As is described in previous works (particularly Robert et al., 2017), the consequence of ignoring these correlations include the following:

1. Unreliable results: compared to the result tested on a truly independent set, CV and OOS results would appear to be better than they are, making predictions less reliable.

2. Overfitting: tests would falsely indicate more complicated models perform better when correlations are present. Imagine that the left-out data are perfectly correlated with the test data. Then a model that perfectly (over) fits the training data will perfectly predict the test data. Thus, cross-validation or OOS will not successfully limit overfitting.

3. Misinterpretation of variables: The correlations can be incorporated into the model process through some of the covariates that also have spatial-temporal variation and appear to be more predictive than they actually are, e.g., rainfall.

However, previous studies have not clearly shown the consequences of ignoring the spatiotemporal correlations under different spatial and temporal structures and how they differ in real-world applications in different fields. In this study, we ask three research questions. What are the preferred cross-validation (CV) and out-of-sample (OOS) performance evaluation methods for given spatiotemporal correlations in panel data? How do we adjust for spatiotemporal correlations to improve the out-of-sample performance of the models? Furthermore, how can we tackle spatial and temporal heterogeneity? This paper address this issue using Monte Carlo simulation to specify different correlation structures and varying amounts of correlations between observations. This allows us to see how significant bias can be across different cases. Bearing the same correlation structures in mind, we assess the results on real-world datasets in household surveys and deforestation. The performance is compared to a “truly” independent dataset created using stratified random sample of the original test set with the proposed correlations.

Previous studies have proposed several ways to address spatial and temporal autocorrelations. Brenning (2012) uses k-means clustering to partition to reduce the influence of spatial autocorrelation. Roberts et al. (2017) invent “block cross-validation” by introducing spatial blocks on contiguous geographic space to force the model to be tested on more distant data, with similar ideas for temporal blocks of grouping data that fall in the nearby time interval. Meyer et al. (2018) create “Leave-Location-and-Time-Out (LLTO) CV” to address spatiotemporal correlations. Schratz et al. (2019) focus on the hyperparameter tuning aspect concerning the underlying spatial correlations in the data. We compare the effectiveness of these methods in both the simulated dataset and the real-world panel data.

One closely related but under-addressed issue is how to appropriately capture heterogeneity in panel data. Such heterogeneity can come from seasonality (e.g., air pollution in winter), natural disasters (severe drought in one year), or geography (flooding in low altitude regions). All of these cases would make it hard to forecast on the test dataset, especially after we use methods like blocking to test the model on a dataset that are relatively less similar than the training set. In the simulated dataset, we depict the spatiotemporal heterogeneity by varying the variance in a spatially dependent error term. We show how different machine learning models perform on data with such and whether they can benefit from segmentation using clustering method. We then apply the same methodology on the real-world data to tackle spatial and temporal heterogeneity with

We see four contributions to this paper. First, we explore different spatial-temporal error structures in the panel dataset, how they affect the model performance, and model selection differently. Second, we showcase and measure the bias and apply the adjustment for spatial-temporal correlation in both environmental and development settings in real-world data. Third, we contribute to the econometrics literature on spatial panel forecasting with cross-validation strategies for evaluating model performance in a more robust way. Last, we provide modelling strategies for tackling spatiotemporal heterogeneity in the data.

1. **Methodology**

* 1. Monte Carlo simulation

Consider a Data Generating Process (DGP) in a panel setting with spatial-temporal correlations:

where is the dependent variable for region at time , is a vector of explanatory variables, are elements of the spatial weight matrixand is the spatial lag coefficient, and is the error component with different specifications of spatially autocorrelated residuals.

* + 1. Define different spatial-temporal error structures

e.g. allowing for spatial diffusion patterns, in the case of air pollution

is an individual-specific time-invariant effect, follow a spatial diffusion process (correlated with ) and a true random error .

* + 1. Demonstration of consequences for ignoring the spatial-temporal correlation
    2. Adjustment for spatial-temporal error structures (k-means clustering, blocking)
    3. Clustering methods for dealing with heterogeneity
  1. Empirical data application

Evaluate on two types of widely used panel datasets in development and environmental studies: household surveys and deforestation.

* + 1. Assess spatial-temporal correlations in the data
    2. CV and OOS performance comparison (on a correlated set and a relatively independent set)
    3. Model implications (model complexity, variable selection, the spatial distribution of prediction error)
    4. Adjustment for possible spatial-temporal correlations (k-means clustering, blocking)
    5. Clustering methods for dealing with heterogeneity

1. **Data**

Simulated data with known data generated process on the spatial-temporal correlation

LSMS (living standards monitoring survey) and Global Forest Change data (2000-2017)

**References**

Baltagi, Badi H., Georges Bresson, and Alain Pirotte. "Forecasting with spatial panel data." *Computational Statistics & Data Analysis* 56, no. 11 (2012): 3381-3397.

Baltagi, Badi H., Bernard Fingleton, and Alain Pirotte. "Estimating and forecasting with a dynamic spatial panel data model." *Oxford Bulletin of Economics and Statistics* 76, no. 1 (2014): 112-138.

Brenning, Alexander. "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest." In *2012 IEEE international geoscience and remote sensing symposium*, pp. 5372-5375. IEEE, 2012.

Meyer, Hanna, Christoph Reudenbach, Tomislav Hengl, Marwan Katurji, and Thomas Nauss. "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation." *Environmental Modelling & Software* 101 (2018): 1-9.

Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera‐Arroita, Severin Hauenstein, et al. "Cross‐validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography* 40, no. 8 (2017): 913-929.

Oliveira, Mariana, Luís Torgo, and Vítor Santos Costa. "Evaluation procedures for forecasting with spatio-temporal data." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 703-718. Springer, Cham, 2018.

Schratz, Patrick, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data." *arXiv preprint arXiv:1803.11266* (2018).