Evaluation procedures for forecasting with spatio-temporal data

Mariana Oliveira[1 . 2] · Luís Torgo[1 . 2 . 3] · Vítor Santos Costa[1 . 2]
(mariana.r.oliveira@inesctec.pt)

1 INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL
2 U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO
3 DALHOUSIE UNIVERSITY

Motivation

Assessing how a machine learning solution will perform on unseen data is crucial.

This involves choosing an evaluation procedure that can make the best use of available data to reliably estimate the chosen performance metrics.

When spatio-temporal dependencies are present in the data, the assumptions made by common procedures, such as cross-validation, are broken.

In this work, we investigate the predictive ability of multiple cross-validation and out-of-sample evaluation procedures for forecasting of geo-referenced time series.

Background

Out-of-sample (OOS) procedures divide the data into training and testing sets that respect the underlying order of the data (e.g., in time series the testing set is comprised of the more recent observations).

Prequential procedures use data to test in one step, and then add it to the training set on the next step. Monte Carlo procedures repeat a time-wise holdout at random time-points and average over several repetitions.

In **cross-validation** (CV) the data is split several times into training and test sets with each observation being part of the test set at least once.

Several variants of cross-validation have been proposed for time series. The main ideas are to keep consecutive observations in the same test set (block CV) and/or create a "buffer" of observations around the testing set that are not used for training (as in modified CV and hv-block CV). The same ideas can be applied to spatial data: keeping contiguous spatial points in each testing set (block CV) and/or adding a "buffer" around the observations in the test set.

Experimental setup

1. Split data into an in-set and an out-set (20% of the most recent observations);
2. Train a regression model (linear model, LM, or random forest, RF) on the in-set and test it on the out-set calculating NMAE: this is the "gold standard" error (*Gold*);
3. Apply evaluation procedures to the in-set: the result is the estimated error (*Est*);
4. Compare the error estimated by each procedure to the "gold standard".

Datasets

- 96 artificial datasets of different grid sizes (8x8 and 20x20) and time series length (150 and 300 time-points), generated by STARMA models;
- 17 variables from 7 real-world data sources including climate, air pollution and agronomic data of different sizes and levels of missing data.

Estimation procedures

We test holdout (H), Monte Carlo (MC), prequential (P) and cross-validation (CV).

Table 1. Cross-validation and prequential fold assignment methods.

		Time	Space		
Cross-validation	Standard	random	random	tRsR	• † ‡
	Time-sliced		all	tRsA	
	Spatial block	all	random block	tAsR	•
	Checkered spatial block		systematic	tAsS	
	Contiguous spatial block		contiguous	tAsC	•
Prequential evaluation	Time block	block	all	tBsA	†
	Spatio-temporal block		random block	tBsR	‡
	Spatio-temporal checkered block		systematic	tBsS	
	Spatio-temporal contiguous block		contiguous	tBsC	

† Time-buffered CV variation included
• Space-buffered CV variation included
‡ Space-time buffered CV variation included

Results

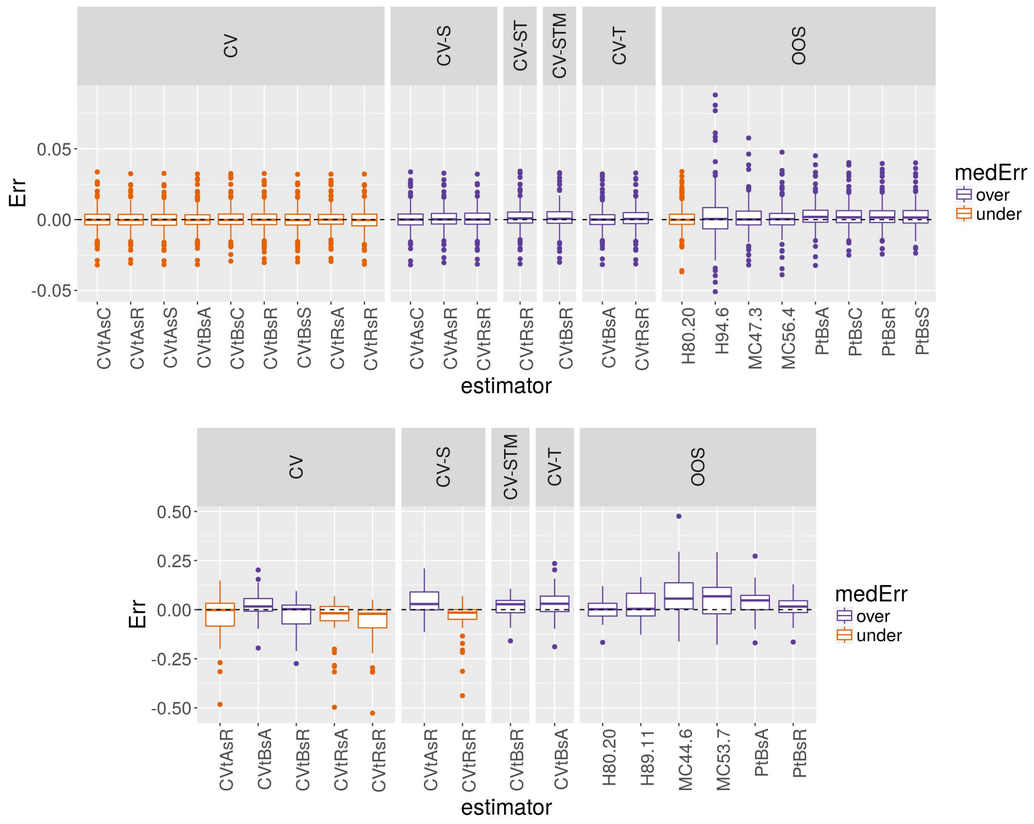

Figure 1. Box plots of estimation errors (*Err=Est-Gold*) incurred by CV and OOS procedures on artificial (top) and real-world (bottom) data.

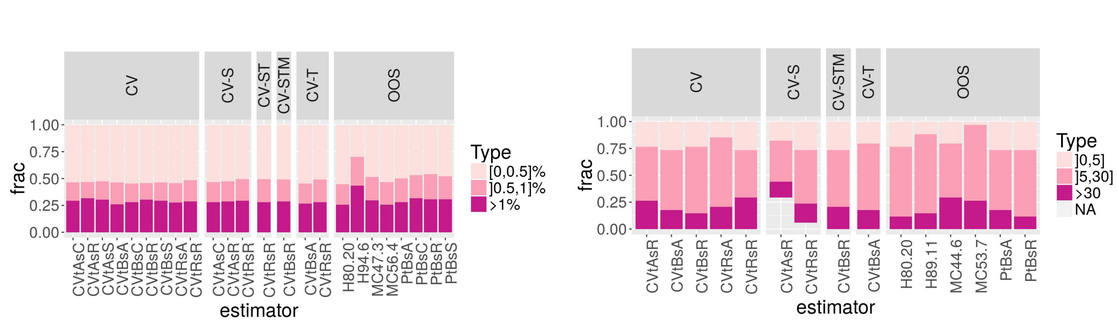

Figure 2. Bar plots of relative estimation errors (*RelErr=|Est-Gold|/Gold*) incurred by CV and OOS procedures on artificial (left) and real-world (right) data.

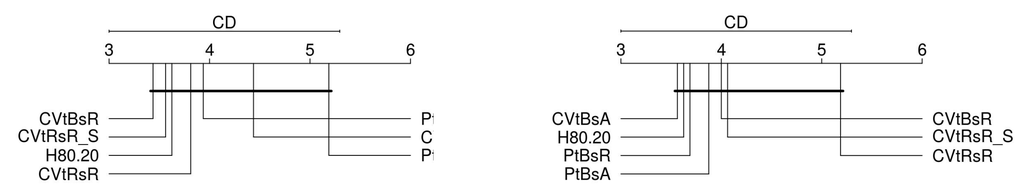

Figure 3. Critical difference diagram according to Friedman-Nemenyi test (at 5% confidence level) given the absolute error incurred by some of the better performing estimation procedures (*AbsErr=|Est-Gold|*) when using a linear model (left) and random forest (right).

Conclusions

- Most often error estimates are reasonably accurate;
- Standard CV underfits and exhibits outliers of severe error underestimation;
- Though the best error estimator is not always the same, most top performers block the data in time;
- In real-world datasets, spatio-temporal block and time block CV approximate the error better than other methods and avoid being overly optimistic;
- OOS procedures did not do as well, but they did avoid underestimation of the error in almost all real-world cases.
- Though there is some bias in the experimental design, the results seem to point to the temporal dimension being more important to respect during evaluation.

Acknowledgments. This work is partially funded by the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/2013. Mariana Oliveira is supported by a FCT/MAPi PhD research grant (PD/BD/128166/2016). Vítor Santos Costa is supported by the project POCI-01-0145-FEDER-016844.

FCT Fundação para a Ciência e a Tecnologia