Environmental Modelling & Software 101 (2018) 1–9

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation

Hanna Meyer [a, *], Christoph Reudenbach [a], Tomislav Hengl [b], Marwan Katurji [c], Thomas Nauss [a]

[a] *Faculty of Geography, Philipps-University Marburg, Deutschhausstr. 10, 35037 Marburg, Germany*
[b] *ISRIC — World Soil Information, P.O. Box 363, 6700 AJ Wageningen, The Netherlands*
[c] *Center for Atmospheric Research, University of Canterbury, Private Bag 4800, Christchurch 8020, New Zealand*

ARTICLE INFO

Article history:
Received 16 October 2017
Accepted 2 December 2017

Keywords:
Cross-validation
Feature selection
Over-fitting
Random forest
Spatio-temporal
Target-oriented validation

ABSTRACT

Importance of target-oriented validation strategies for spatio-temporal prediction models is illustrated using two case studies: (1) modelling of air temperature (T_{air}) in Antarctica, and (2) modelling of volumetric water content (VW) for the R.J. Cook Agronomy Farm, USA. Performance of a random *k*-fold cross-validation (CV) was compared to three target-oriented strategies: Leave-Location-Out (LLO), Leave-Time-Out (LTO), and Leave-Location-and-Time-Out (LLTO) CV. Results indicate that considerable differences between random *k*-fold ($R^2 = 0.9$ for T_{air} and 0.92 for VW) and target-oriented CV (LLO $R^2 = 0.24$ for T_{air} and 0.49 for VW) exist, highlighting the need for target-oriented validation to avoid an overoptimistic view on models. Differences between random *k*-fold and target-oriented CV indicate spatial over-fitting caused by misleading variables. To decrease over-fitting, a forward feature selection in conjunction with target-oriented CV is proposed. It decreased over-fitting and simultaneously improved target-oriented performances (LLO CV $R^2 = 0.47$ for T_{air} and 0.55 for VW).

1. Introduction

Machine learning algorithms are well established in environmental sciences (Lary et al., 2016; Kanevski et al., 2009) and find application in a variety of fields as for example mapping of land cover (Ludwig et al., 2016; Gislason et al., 2006), vegetation characteristics (Lehnert et al., 2015; Verrelst et al., 2012) and soil properties (Gasch et al., 2015; Lieβ et al., 2016) as well as in geomorphological (Messenzehl et al., 2017; Micheletti et al., 2014) or climatological (Kühnlein et al., 2014; Hong et al., 2004; Meyer et al., 2016a; Appelhans et al., 2015) studies. Most of the applications focus on static spatial predictions and are not aiming at estimating a certain variable simultaneously in space and time. However, though machine learning algorithms are still rarely applied in spatio-temporal models, the number of applications is increasing (Gokaraju et al., 2011; Gasch et al., 2015; Appelhans et al., 2015; Meyer et al., 2016b; Ho et al., 2014; Jing et al., 2016; Ke et al., 2016; Lary et al., 2014).

Machine learning algorithms in space-time applications learn from spatio-temporal observations to predict a certain variable for unknown locations and for an unknown point in time (within a defined model domain) allowing a monitoring of the environmental variable. The term *"prediction"*, in this context, should not to be confused with *"forecasting"* as most of the models are not aiming at predicting into the future but rather focus on predicting in past or present times as well as in space. In contrast to model-based geostatistics (Diggle and Ribeiro, 2007) as for example (co-) kriging, where one needs sufficiently distributed information on the variable at question for each interpolation time-step, spatio-temporal prediction models link a set of independent variables to the response (i.e. the variable in question) and only use those independent variables for the subsequent spatio-temporal prediction application. A typical example of spatio-temporal prediction models in environmental science might be the estimation of soil properties as done by Gasch et al. (2015). In this example, soil properties (volumetric water content, soil temperature and bulk electrical conductivity) are predicted in space and time on the basis of a machine learning model which is developed from a

[*] Corresponding author.
E-mail address: hanna.meyer@geo.uni-marburg.de (H. Meyer).

https://doi.org/10.1016/j.envsoft.2017.12.001
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

variety of spatial, temporal and spatio-temporal predictor variables as well as *"ground truth"* observations taken from data loggers.

Studies by Gasch et al. (2015) and Meyer et al. (2016a) have shown that the estimated performance of such models highly depends on the validation strategy: in both cases high differences between the performance estimated by a random test subset of the total dataset and the performance estimated by a Leave-Location-Out (LLO) Cross-Validation (CV) have been reported. LLO CV means that models are repeatedly trained by leaving the data from one location or a group of locations (i.e. climate stations, data loggers) out and using the respective held back data for model validation. The differences between a random subset validation (lower error estimates) and LLO CV (higher error estimates) strongly suggest spatial over-fitting as the models can very well predict on subsets of the time series of the locations used for training, but fail in the prediction of unknown locations. The prediction on unknown locations, however, is in most cases the major task of such models. The LLO CV error must therefore be considered as the decisive performance indicator of spatial as well as spatio-temporal models. Similarly, spatio-temporal models have a risk of temporal over-fitting which needs to be assessed by Leave-Time-Out (LTO) CV (Gudmundsson and Seneviratne, 2015). However, it is these *"target-oriented"* validation strategies that focus on the model performance in the context of unknown space or unknown time steps that are not yet fully prevailed in literature. This is especially a problem as case studies ignoring the spatio-temoral dependence in the data have to be considered too optimistic (Roberts et al., 2017). Even though LLO and LTO CV are used in some studies on spatial and spatio-temporal models (Ho et al., 2014; Gudmundsson and Seneviratne, 2015; Ruβ and Brenning, 2010; Meyer et al., 2017b; Brenning et al., 2012; Micheletti et al., 2014), random *k*-fold CV, where the dataset is randomly partitioned into folds, is still considered common practice (Ke et al., 2016; Messenzehl et al., 2017; Lieβ et al., 2016; Ludwig et al., 2016).

How to address spatial or spatio-temporal over-fitting in view to improved model selections? Over-fitting in machine learning models (when applied to spatial data) most likely happens due to poor representation of spatio-temporal sampling in predictor variable spaces. Hence, carefully selecting and interpreting predictor variables is a logical remedy for improving performance of spatial models. Many spatio-temporal prediction studies use auxiliary predictor variables which describe the properties of the location (e.g. elevation, slope, soil type, spatial coordinates). These variables vary in space but not in time which means that each station has a unique combination of static variables. We hypothesize hence that:

1. These temporally static variables are prone to over-fitting. Combinations of unique properties for each location are quasi comparable to a unique ID of the locations which is then used as predictor. Using such variables, the model is able to fit general characteristics of the individual time series.
2. Variables that lead to over-fitting can be automatically identified and removed using a feature selection method that accounts for the target-oriented performance.
3. Excluding misleading variables from the models does not only decrease over-fitting but also leads to improved target-oriented model performances.

Feature selection is an intuitive solution to reduce the number of variables to the most important ones. However, the commonly used method for feature selection, Recursive Feature Elimination (RFE) (see e.g. Brungard et al., 2015; Meyer et al., 2017a, b; Ghosh and Joshi, 2014; Stevens et al., 2013; in the field of environmental mapping), relies on variable importance scores which are calculated using solely the training subset (Kuhn and Johnson, 2013). If a variable leads to considerable over-fitting, it has a high importance in the models. Therefore, this variable will be selected as important variable in the RFE process and is not removed regardless of a resulting high LLO CV error. Alternative approaches for detecting the over-fitting variables are hence required.

We consider two published case studies to demonstrate the effect of different validation strategies, the risk of spatial or spatio-temporal over-fitting as well as the potential of feature selection algorithms to minimize the degree of over-fitting. To estimate the degree of over-fitting, we compare the results of a random *k*-fold CV with the results of the target-oriented validation strategies LLO, LTO and Leave-Location-and-Time-out (LLTO) CV. We then compare the RFE method with a newly proposed forward feature selection (FFS) method that works in conjunction with target-oriented performance to identify and remove variables that lead to over-fitting. As machine learning algorithm, the well-known Random Forest algorithm (Breiman, 2001) was applied as it appeals to a large community of users. We implement all steps of data analysis and modelling in the R environment for statistical programming (R Core Team, 2016). Most of the analysis is based on the caret package (Kuhn, 2016) that implements a wrapper to the Random Forest algorithm being used and provides functionality for data splitting and CV. All newly produced R functions and modelling steps are fully documented in https://github.com/environmentalinformatics-marburg/CAST.

2. Case studies and description of the datasets

2.1. Case study I: modelling air temperature in Antarctica

The first case study follows the approach of Meyer et al. (2016a) to spatio-temporally predict T_{air} in Antarctica based on LST data from the Moderate Resolution Imaging Spectroradiometer (MODIS) and auxiliary predictor variables. The dataset as it was used in the present study consists of 30666 hourly air temperature measurements from 32 weather stations distributed over Antarctica for the year 2013. The T_{air} values range from −78.40°C to 5.76°C with an average of −27.64°C and a standard deviation of 17.26°C.

Beside of MODIS based LST as a spatio-temporal predictor variable, several auxiliary spatial predictor variables were used that basically describe the terrain. In addition, a number of predictor variables that remain spatially constant but vary in time were used as temporal predictor variables. See Table 1 for the full list of predictors used in this study and Meyer et al. (2016a) for further information on the dataset.

2.2. Case study II: modelling volumetric water content of the "Cookfarm", USA

The second case study bases on the dataset applied in Gasch et al. (2015) to predict soil properties in 3D+time and can be freely accessed from the GSIF package in R. The research site of this case study is the R.J. Cook Agronomy Farm which is a 37 ha sized long-term agroecosystem research site in the Palouse region in the USA and operated by the Washington State University. The final dataset as prepared for this study consists of daily VW measurements from the years 2011–2013 taken by 5TE sensors (Decagon Devices, Inc., Pullman, Washington) initially installed in five depth