SPATIAL CROSS-VALIDATION AND BOOTSTRAP FOR THE ASSESSMENT OF PREDICTION RULES IN REMOTE SENSING: THE R PACKAGE SPERROREST

Alexander Brenning

University of Waterloo
Department of Geography and Environmental Management
200 University Avenue West
Waterloo, Ontario N2L 3G1, Canada

ABSTRACT

Novel computational and statistical prediction methods such as the support vector machine are becoming increasingly popular in remote-sensing applications and need to be compared to more traditional approaches like maximum-likelihood classification. However, the accuracy assessment of such predictive models in a spatial context needs to account for the presence of spatial autocorrelation in geospatial data by using spatial cross-validation and bootstrap strategies instead of their now more widely used non-spatial equivalent. These spatial resampling-based estimation procedures were therefore implemented in a new package 'sperrorest' for the open-source statistical data analysis software R. This package is introduced using the example of the detection of rock-glacier flow structures from IKONOS-derived Gabor texture features and terrain attribute data.

***Index Terms*—** Spatial cross-validation, spatial bootstrap, classification accuracy, land cover classification, Gabor filters, rock glaciers

1. INTRODUCTION

Novel prediction methods developed in the fields of computational statistics and machine learning, such as random forests and the support vector machine, have the potential to significantly improve land cover classification from remote sensing data, and provide a solution to high-dimensional prediction problems in hyperspectral remote sensing [1, 2]. However, while these methods are becoming increasingly popular in remote sensing, little attention has been paid to the challenges arising from the presence of spatial autocorrelation in geospatial data, which may result in overfitting these highly flexible methods to the training data or underestimating spatial prediction errors in resampling-based accuracy assessment [3, 4].

This contribution presents an implementation of resampling-based spatial estimation methods (section 3). It furthermore applies spatial cross-validation in the context of remotely-sensed land cover classification using a combination of texture filters and terrain attributes and four statistical and machine-learning classifiers of different complexity (section 4).

2. SPATIAL CROSS-VALIDATION AND BOOTSTRAP

One of the main implications of spatial autocorrelation for accuracy assessment is that the overfitting of classifiers to training (or calibration) samples will remain undetected if the test (or validation) samples are not independent of the training set. This phenomenon, which is also known in other contexts where twinning occurs (e.g., [5]), is particularly important in a spatial and temporal context (e.g., [3, 4, 6, 7, 8, 9]). Different spatial resampling-based error estimation techniques have therefore been proposed for estimation and testing in a spatial context, but their common denominator is to perform resampling at a spatially aggregated level.

Implementations of spatial cross-validation and bootstrap are currently not yet available in statistical or remote sensing software are currently not yet available but are required in order to promote the use of such bias-reduced methods. Open-source software is particularly suited for this.

3. IMPLEMENTATION IN R

Spatial and non-spatial cross-validation and boostrap methods were implemented in this contribution as an extension ('sperrorest' package) to the statistical data analysis software R [10]. R is an open-source software and programming language that is increasingly gaining popularity in spatial analysis and prediction due to its increasing integration with GIS software and data formats [11, 12] and the availability of cutting-edge statistical and machine-learning methods (compare, e.g., [13]). The 'sperrorest' package will be made

This research has been funded through an NSERC Discovery Grant – Individual awarded to A. Brenning. The collaboration of Shilei Long and Paul Fieguth, University of Waterloo, on the case study is gratefully acknowledged.

available online via the Comprehensive R Archive Network (CRAN).

At the core of the 'sperrorest' package are several functions for spatial and non-spatial partitioning and resampling and a main function '`sperrorest`' that performs the actual error estimation based on prediction models implemented in other R packages. This main function builds upon the general functionality and design of the non-spatial estimation function '`errorest`' in package 'ipred' [14]. Furthermore available are numerous functions for summarizing and plotting resampling schemes and for calculating and summarizing a variety of error measures.

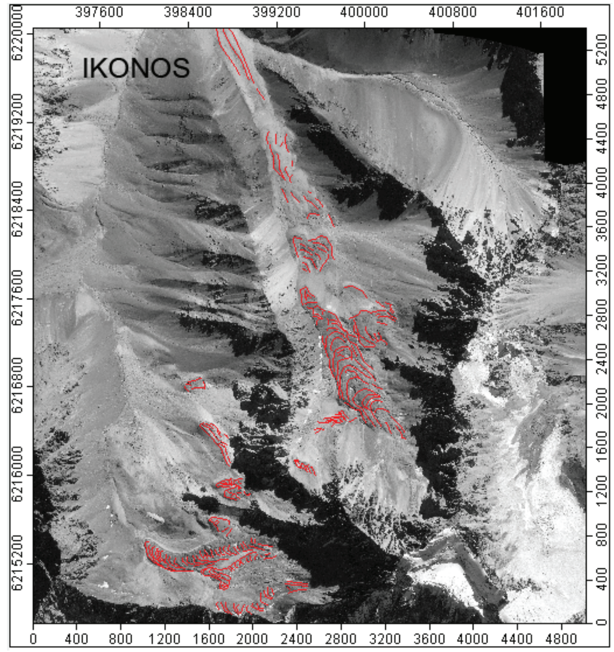
The 'sperrorest' implementation is extensible as user-defined resampling and partitioning functions can be used to accommodate specific sampling designs or dependence structures, including non-spatial ones. In addition to typical distance-based spatial partitioning, it is possible to perform cross-validation at the level of aggregated objects. An important application of this feature is the assessment of classifiers for the remotely-sensed classification of crop type as in this situation the agricultural field is the appropriate observational unit, not an individual pixel in the remote sensing imagery [15].

4. EXAMPLE: ROCK GLACIER MAPPING

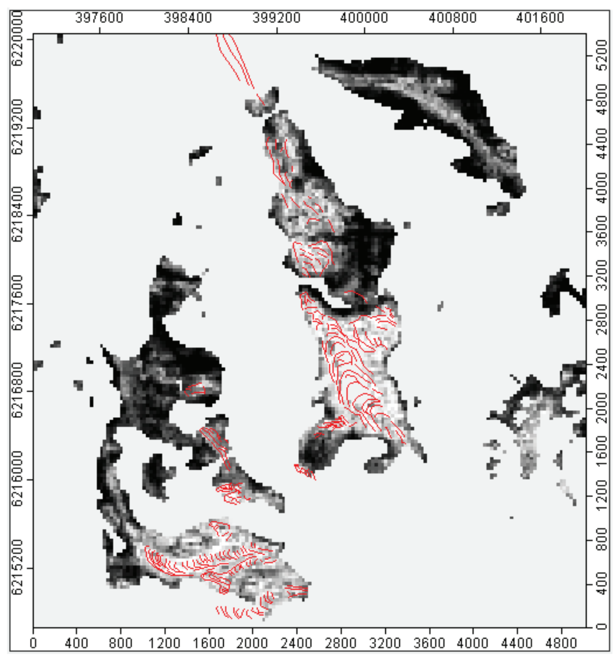
Automated remotely-sensed mapping of rock glaciers, a landform resulting from the creep of ice-rich mountain permafrost, is a challenging task since rock glaciers do not exhibit a distinct spectral signal [16]. The utility of a combination of Gabor texture filters and terrain attributes has therefore been explored in order to specifically detect flow structures developed on the surface of most rock glaciers based on their characteristics zebra-stripe pattern, which is visible in very-high-resolution imagery such as IKONOS panchromatic data (for details, see [17]).

In this case study, a filterbank of 40 Gabor texture filters was derived from two IKONOS panchromatic images from the Andes of Central Chile (Cerro Catedral and Laguna Negra areas), and a set of 6 terrain attributes was calculated as additional predictors. Terrain attributes as well as masks for exposed bedrock and snow-covered areas were furthermore used to narrow in the actual area of interest, i.e. rock glaciers versus other debris surfaces.

Four supervised classification methods (generalized linear model, GLM; generalized additive model, GAM [18]; support vector machine, SVM [2]; Bundling [19]) were utilized, and predictive performances were estimated using spatial cross-validation estimates of the area under the ROC (receiver operating characteristics) curve (AUROC) were calculated for performance assessment. Spatial cross-validation was implemented as a 100-repeated 5-fold cross-validation in which spatial partitions were derived by k-means clustering ($k = 5$).



(a) IKONOS Panchromatic



(b) Bundling prediction

Fig. 1. IKONOS panchromatic image with digitized rock glacier flow structures, and model predictions generated by the Bundling classifier; only a portion of the study area is shown. Bright gray tones in the prediction map represent high predicted probabilities of presence of flow structures. Masked areas correspond to unsuitable topography, bedrock, and uncovered glaciers/snow.

Using the 'sperrorest' package in R, a slightly simplified version of the present analysis using Bagging instead of Bundling can be implemented using the following call for 5-fold cross-validation using 100-repeated 5-means clustering for spatial partitioning:

```
sperrorest(
   formula = class ~ ., data = d,
   model.fun = bagging,
   pred.fun = predfun,
   smp.fun = partition.kmeans,
   smp.args = list(kfold = 5,
      repetition = 1:100),
   train.fun = train.strat.uniform,
   train.param = list(nstrat=500),
   test.fun = learn.strat.uniform,
   test.param = list(nstrat=Inf))
```

where `predfun` is a wrapper for the standard prediction methods of bagging:

```
function(object, newdata)
   predict(object, newdata,
      type = "prob")[,2]
```

The above `sperrorest` call utilizes 500 training samples from each of the response classes (presence and absence of rock glacier flow structures; `nstrat=500`). Test samples are also balanced (function `train.strat.uniform`) but use as many samples as possible from the test partition. Stratified sampling is desirable due to the small portion of the area with presence of flow structures.

A standard set of error measures for 'soft' two-class classification problems in which the classifier predicts a probability or index of class membership; this set includes the AUROC. User-defined error functions can, however, also be provided using the optional `err.fun` argument.

In this study the GAM performed best in one area, and Bundling in the other (Table 1). However, the performance of the GAM showed greater variation in median AUROC between the two study areas. Bundling predictions of the likelihood of the presence of rock glacier flow structures are shown in Figure 1.

	Median AUROC	
Classifier	**CAT area**	**LAG area**
GLM	0.756	0.685
GAM	**0.818**	0.732
SVM	0.778	0.766
Bundling	0.771	**0.795**

Table 1. Spatial cross-validation estimates of the AUROC of four classifiers using a combination of terrain attributes and texture filters in the CAT and LAG areas, respectively.

5. OUTLOOK

The 'sperrorest' implementation of spatial cross-validation and bootstrap methods aims at increasing the access to and popularity of these estimation techniques in spatial prediction problems. Yet these techniques are not only required to obtain unbiased estimates of spatial prediction performances, they also provide novel insights into the spatial structure of these prediction problems. Permutation-based spatial variable importance measures that are developed within the framework of spatial resampling-based estimation [20, 17] are just one example of the new opportunities arising from spatial resampling methods.

6. REFERENCES

[1] P.O. Gislason, J.A. Benediktsson, and J.R. Sveinsson, "Random Forests for land cover classification," *Pattern Recognition Letters*, vol. 27, pp. 294–300, 2006, 3rd Pattern Recognition in Remote Sensing Workshop, Kingston Univ, Kingston Upon Thames, England, August 27, 2004.

[2] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[3] A. Brenning, "Spatial prediction models for landslide hazards: review, comparison and evaluation," *Natural Hazards Earth System Sci.*, vol. 5, pp. 853–862, 2005.

[4] G. Ruß and A. Brenning, "Data mining in precision agriculture: Management of spatial information," in *Computational Intelligence for Knowledge-Based Systems Design*, E. Hüllermeier, R. Kruse, and F. Hoffmann, Eds., Berlin, Heidelberg, 2010, vol. 6178 of *LNAI*, pp. 350–359, Springer.

[5] A. Brenning and B. Lausen, "Estimating error rates in the classification of paired organs," *Statistics in Medicine*, vol. 27, pp. 4515–4531, 2008.

[6] P. Bühlmann, "Bootstraps for time series," *Statistical Science*, vol. 17, pp. 52–72, 2002.

[7] A. C. Davison, D. V. Hinkley, and G. A. Young, "Recent developments in bootstrap methodology," *Statistical Science*, vol. 18, pp. 141–157, 2003.

[8] D. N. Politis, "The impact of bootstrap methods on time series analysis," *Statistical Science*, vol. 18, pp. 219–230, 2003.

[9] M. Taper, "Bootstrapping dependent data in ecology," in *ASA Proceedings of the Joint Statistical Meetings*, 2004, pp. 2994–2999.

[10] E. C. Grunsky, “R: a data analysis and statistical programming environment — an emerging tool for the geosciences,” *Computers & Geosciences*, vol. 28, pp. 1219–1222, 2002, doi:10.1016/S0098-3004(02)00034-1.

[11] R. S. Bivand, “Using the R statistical data analysis language on GRASS 5.0 GIS data base files,” *Computers & Geosciences*, vol. 26, pp. 1043–1052, 2000.

[12] A. Brenning, “Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models,” in *SAGA Seconds Out*, Jürgen Böhner, Thomas Blaschke, and Luca Montanarella, Eds., vol. 19 of *Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie*, pp. 23–32. Institut für Geographie, Universität Hamburg, Hamburg, Germany, 2008.

[13] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*, Springer, New York, 4th edition, 2002.

[14] A. Peters and T. Hothorn, *ipred: Improved Predictors*, 2011, R package version 0.8-11.

[15] A. Brenning, K. Kaden, and S. Itzerott, “Comparing classifiers for crop identification based on multitemporal Landsat TM/ETM data,” in *Proceedings, Second Workshop of the EARSeL Special Interest Group on Remote Sensing of Land Use and Land Cover, 28–30 September 2006*, Bonn, Germany, 2006, pp. 64–71.

[16] A. Brenning, “Benchmarking classifiers to optimally integrate terrain analysis and multispectral sensing in automatic rock glacier detection,” *Remote Sensing of Environment*, vol. 113, pp. 239–247, 2009.

[17] A. Brenning, S. Long, and P. Fieguth, “Detecting rock glacier flow structures using gabor filters and IKONOS imagery,” *Remote Sensing of Environment*, vol. submitted, submitted.

[18] T. Hastie. and R. Tibshirani, *Generalized additive models*, Chapman and Hall, London, 1990.

[19] T. Hothorn and B. Lausen, “Bundling classifiers by bagging trees,” *Computational Statistics & Data Analysis*, vol. 49, pp. 1068–1078, 2005.

[20] G. Ruß and A. Brenning, “Spatial variable importance assessment for yield prediction in precision agriculture,” in *Advances in Intelligent Data Analysis IX, Proceedings*, Cohen, P.R. and Adams, N.M. and Berthold, M.R., Ed., 2010, vol. 6065 of *Lecture Notes in Computer Science*, pp. 184–195, 9th International Symposium on Intelligent Data Analysis, Tucson, AZ, May 19-21, 2010.