www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

GEODESY AND CARTOGRAPHY
Vol. 66, No 2, 2017, pp. 171-209

DOI: 10.1515/geocart-2017-0012

Original research paper

Thorough statistical comparison of machine learning regression models and their ensembles for sub-pixel imperviousness and imperviousness change mapping

Wojciech Drzewiecki

AGH University
Faculty of Mining Surveying and Environmental Engineering
Department of Geoinformation, Photogrammetry and Remote Sensing of Environment
al. Mickiewicza 30, 30–059 Kraków, Poland
e-mail: drzewiec@agh.edu.pl

https://orcid.org/0000-0002-9266-0000

Received: 09 August 2017 / Accepted: 2 October 2017

Abstract: We evaluated the performance of nine machine learning regression algorithms and their ensembles for sub-pixel estimation of impervious areas coverages from Landsat imagery. The accuracy of imperviousness mapping in individual time points was assessed based on RMSE, MAE and R^2. These measures were also used for the assessment of imperviousness change intensity estimations. The applicability for detection of relevant changes in impervious areas coverages at sub-pixel level was evaluated using overall accuracy, F-measure and ROC Area Under Curve. The results proved that Cubist algorithm may be advised for Landsat-based mapping of imperviousness for single dates. Stochastic gradient boosting of regression trees (GBM) may be also considered for this purpose. However, Random Forest algorithm is endorsed for both imperviousness change detection and mapping of its intensity.

In all applications the heterogeneous model ensembles performed at least as well as the best individual models or better. They may be recommended for improving the quality of sub-pixel imperviousness and imperviousness change mapping. The study revealed also limitations of the investigated methodology for detection of subtle changes of imperviousness inside the pixel. None of the tested approaches was able to reliably classify changed and non-changed pixels if the relevant change threshold was set as one or three percent. Also for five percent change threshold most of algorithms did not ensure that the accuracy of change map is higher than the accuracy of random classifier. For the threshold of relevant change set as ten percent all approaches performed satisfactory.

Keywords: impervious areas**,** sub-pixel classification, machine learning, model ensembles, Landsat

1. Introduction

Detection and quantification of changes in land use and land cover (LULC) is one of the most common applications of remote sensing. Despite many approaches developed over the years, this is still one of the very actual research areas (Hussain et al., 2015; Tewkesbury et al., 2015).

The majority of change detection techniques are based on processing of remote sensing images at the per-pixel level (Hussain et al., 2015; Tewkesbury et al., 2015). They are able to detect the conversion of LULC form, ie. the change of one land cover type into another. However, in case of remote sensing images, especially these with medium or coarse resolution, pure pixels containing only one land cover type are quite rare. As a result, the considered changes have usually the form of modification (Turner and Meyer, 1994). In such cases LULC category assigned to the mapping unit (pixel) does not change, but the proportions of land cover fractions inside do. For example, despite the substantial increase in impervious areas coverage (eg. from 10 to 30 percent), the pixel may be still classified as "discontinuous built-up".

Although other kinds of fractional coverages (eg. tree canopy) are also determined based on remote sensing images, the mapping of impervious surface areas (ISA) is probably the most frequent application of sub-pixel classification techniques. This is because the accurate information about ISA coverage and monitoring of its change is necessary for different kinds of environmental studies (Dams et al., 2013; Shahtahmassebi et al., 2014). The spectral mixture analysis-based methods are preferred for ISA mapping in urbanised areas (Ridd, 1995; Lu et al., 2014a). In areas dominated by non-urban types of land cover, the regression-based approaches are considered as more appropriate (Lu et al., 2014b; Heremans and Van Orshoven, 2015).

The methodology of sub-pixel imperviousness change detection was proposed by Yang et al. (2003). The approach belongs to layer arithmetic change detection techniques (Tewkesbury, 2015). To evaluate the change of imperviousness over time, the results of regression-based ISA assessments for particular points in time are subtracted one from another. The final map presents not only the location of ISA changes but also their intensity.

One can easily find studies done according to this methodology using a wide spectrum of ISA mapping techniques. Their authors usually evaluate carefully the accuracy of ISA maps made for assessed points in time. However, it is hard to find any studies where the accuracy of change map was determined as well. Such situation is not restricted to ISA mapping or other sub-pixel assessments. As reported by Olofsson et al. (2014), land change studies based on classification of remote sensing data "routinely fail to assess the accuracy of the final change maps".

It is commonly assumed that by maximizing the accuracy of individual ISA assessments one can obtain the best assessment of ISA change. Such assumption is true for post-classification change detection techniques (Hussain, 2013). However, in case of sub-pixel assessment of fractional coverages (eg. imperviousness) when the change is obtained as their difference, the change assessment error depends not only

on the errors of individual time point evaluations, but also on the correlation of these errors (Morgan and Herion, 1990; Kircher, 2001; Drzewiecki 2016b). As a result, the approaches most accurate for mapping ISA in individual points in time may give worse change assessment than less accurate ones.

Such effect was reported by Drzewiecki (2016a) and Drzewiecki (2016b). In the latter paper nine non-linear regression models were compared for sub-pixel impervious surface area mapping from Landsat images in three study areas. The imperviousness was evaluated for two points in time and the change in ISA coverage assessed as well. Root Mean Squared Error (RMSE) and Mean Absolut Error (MAE) were used as performance measures. For individual points in time the best results were obtained using the Cubist algorithm. But for ISA change evaluation the Cubist algorithm was outperformed by others. The best results of change assessment were achieved using Random Forest algorithm which also gave the most correlated errors of individual time point evaluations.

Drzewiecki (2016b) has also shown that when machine learning methods are used for ISA mapping, it is possible to improve the accuracy of sub-pixel imperviousness change assessment using ensembles of heterogeneous non-linear regression models. The best models trained using individual tested algorithms were successfully ensembled using backward selection schema approach (Coelho and Von Zuben, 2006) into models which outperformed best individual models in ISA change evaluation (gave lower RMSE and MAE values).

In Drzewiecki (2016a) and Drzewiecki (2016b) both cross-validation and independent validation datasets were used to evaluate the results. However, with regard to these studies one may consider several issues.

Firstly, in both studies the paired *t*-tests with Bonferroni correction were used for comparison of model performances for both cross-validation and independent dataset validation results. But, in case of repeated cross-validation *t*-test based assessment may lead to wrong conclusions due to high probability of Type I error, i.e. the rejection of the null hypothesis incorrectly (Bouckaert, 2003; Bouckaert and Frank, 2004). This is because the independence assumption necessary for *t*-test is violated as some data are re-used in different cross-validation realisations and the train and test sets overlap (Bouckaert and Frank, 2004; Santafe et al., 2015). As a result we may consider particular algorithm as significantly worse, although in fact no significant difference exists. In Drzewiecki (2016a; 2016b) the danger of this kind was reduced to some extent by using very low threshold for *p*-values to reject the null hypothesis ($p < 0.001$) and very conservative Bonferroni approach to correction of family-wise error in multiple comparisons. Nevertheless, the more appropriate method might be used and some corrected approach to calculate *t*-test statistic adopted (Japkowicz and Shah, 2011; Santafe et al., 2015).

On the other hand, the Bonferroni correction used in both studies may result in the low power of the test (Santafe et al., 2015). In consequence, the null hypothesis might not be rejected although there actually was the difference in the performance of considered algorithms (Type II error). This might cause the lack of statistically

significant differences of performance measures reported in Drzewiecki (2016b) for validation datasets. Again, more powerful method to adjust the significance level values for multiple comparisons may be chosen from the approaches proposed in literature (Santafe et al., 2015).

Thirdly, Drzewiecki (2016b) compared for imperviousness change assessment performances of model ensembles and individual algorithms. But, to create the change map we can also use the best (in terms of RMSE or MAE) estimates for individual time steps, which may be obtained using different algorithms. Thus, in order to fully evaluate the merits of using model ensembles we should also compare their performance to such alternatives.

Finally, the research of Drzewiecki (2016a) and Drzewiecki (2016b) focused on the accuracy of ISA change intensity evaluation. However, in some applications one may be interested how accurate the occurrence of change may be detected rather than what the accuracy of change intensity estimation is. The quality of sub-pixel ISA change map from change detection point of view may be assessed in the same way as quality of change maps obtained at pixel level, i.e. based on the numbers of pixels correctly and incorrectly identified as changed or unchanged.

Because of the issues outlined above we decided to refine and extend the previous research. Therefore, this paper present the results of the study aimed at:

1) thorough statistical re-evaluation of the sub-pixel imperviousness and imperviousness change intensity assessment results reported in Drzewiecki (2016b) using new approaches as suggested in machine learning literature (Santafe et al., 2015);
2) comparison of selected machine learning algorithms in the context of sub-pixel imperviousness change detection accuracy, and
3) answering a question if by ensembling of heterogeneous non-linear regression models one can find the approach with higher ability of ISA change detection and/or more accurate ISA change intensity evaluation than using individual algorithms or the best models for individual time points assessments.

2. Methods

2.1. Study areas and datasets

The research was intended as a continuation and extension of the study presented in Drzewiecki (2016b). The same image datasets, comprised of Landsat images of three watersheds (Raba, Dunajec and Soła) located in South Poland, were used. All three regions are rural areas with large forest cover. Details about current land use and land cover in studied watersheds may be found in Wężyk et al. (2016). Urbanized areas covers from ca. 5 (Dunajec) to ca. 8 percent (Raba) of the area. In Soła and Dunajec watersheds the most (ca. 80%) of built-up areas were classified as dense development. In Raba watershed the sparse development prevails.

For each watershed two image datasets acquired in middle 1990s and late 2000s were available. Aerial orthophotomaps were used as the reference. All the details about preparation of calibration and validation datasets (including calculation of prediction variables and splitting of datasets), machine learning algorithms used, tuning the models (using 10-times repeated 5-fold cross-validation procedure) and creating model ensembles are provided in Drzewiecki (2016b). Within the research presented in this paper the data had not been reprocessed, but the results of ISA change evaluations reported in Drzewiecki (2016b) were subjected to further analysis (Figure 1).

2.2. Detection of relevant changes

The results of automating change detection using remote sensing images may be used in different applications. Depending on application the potential user may define a set of criteria that determine what kind of change is considered as relevant and as not relevant (Klaric, 2014). In case of imperviousness the users may be interested in detection of ISA changes of varying intensity. As a result, the relevant changes may be defined as increase (or decrease) of ISA greater than the user-defined threshold value.

In Drzewiecki (2016b) nine machine learning (ML) regression algorithms were tested: Cubist (Quinlan, 1993), Random Forest (RF) (Breiman, 2001), stochastic gradient boosting of regression trees (GBM) (Friedman, 2002), k-nearest neighbors (kNN), random k-nearest neighbors (rkNN) (Li et al., 2011), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991), averaged neural networks (avNN) (Ripley, 1996), support vector machines (Smola and Schölkopf, 2004) with polynomial (SVMp) and radial (SVMr) kernels. For every study area, each of them was used to predict imperviousness for both mid 1990s and late 2000s. The change maps were also obtained as difference maps. ISA and ISA change assessment was also done for model ensembles. In present study, to detect relevant changes of ISA, the imperviousness change maps from Drzewiecki (2016b) and additional change maps created by subtracting the most accurate (according to RMSE and MAE) individual time points predictions were thresholded. Four different threshold values were used: 1%, 3%, 5% and 10%. These values correspond to changes of impervious surface area within a pixel of 9, 27, 45 and 90 squared meters, respectively. As a result pixels when relevant changes occurred were coded in binary form (1/0 – change/no change) and compared to the reference (i.e. thresholded ISA change evaluated based on high resolution aerial orthos). The characteristics of calibration and validation datasets taking into account the numbers of changes considered as relevant are presented in Table 1.

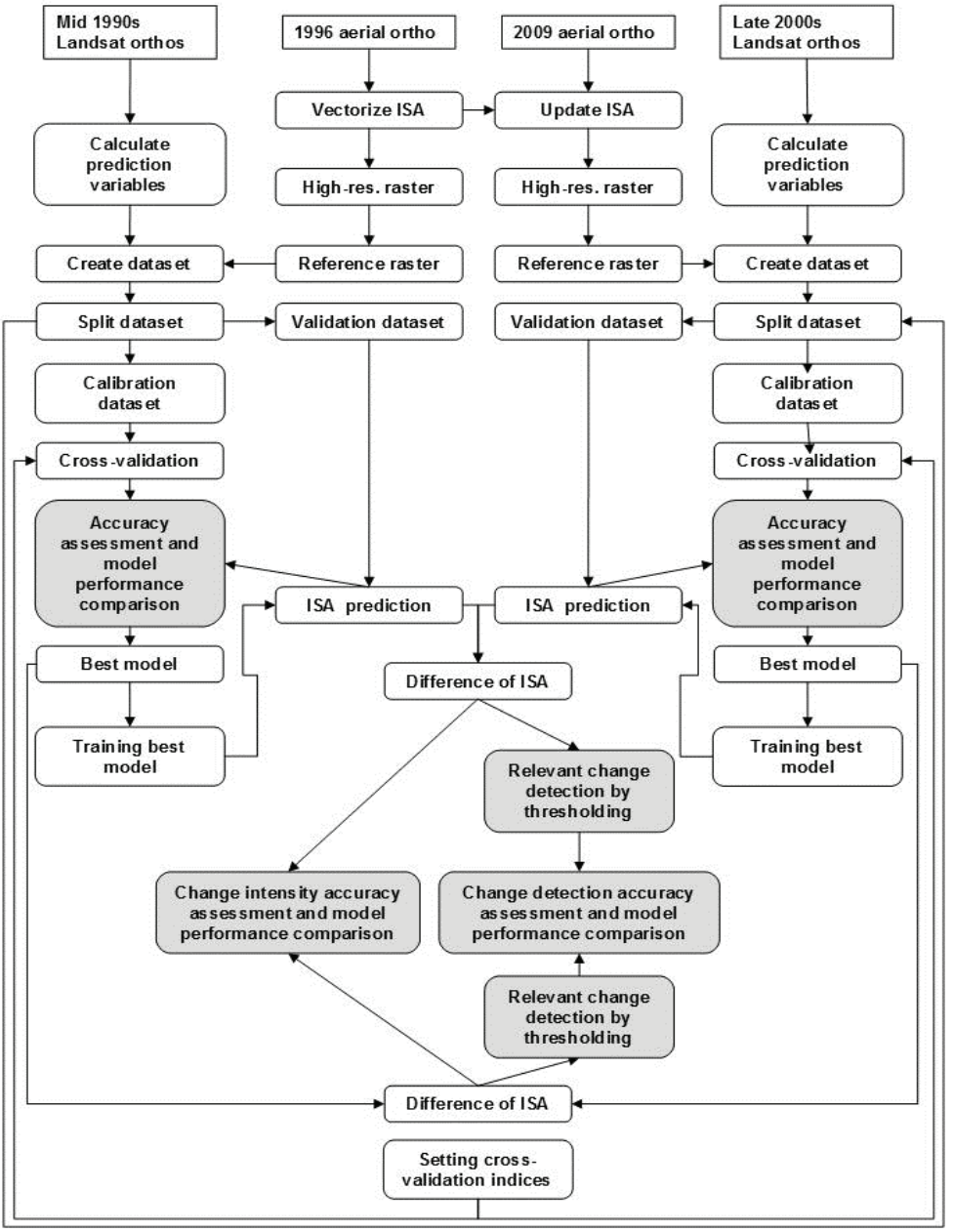

Fig. 1. Framework of ISA relevant change detection accuracy assessment
(steps shown in gray done within this study – remaining ones taken from Drzewiecki 2016b)

Table 1. Datasets characteristics

Dataset	Time period	No of predictors	calibration dataset					validation dataset				
			No of pixels	Relevant changes				No of pixels	Relevant changes			
				1%	3%	5%	10%		1%	3%	5%	10%
Raba	mid 1990s	132	2310	579	429	329	188	578	145	102	87	48
	late 2000s	66										
Dunajec	mid 1990s	66	1382	266	222	206	176	346	62	53	45	40
	late 2000s	165										
Sola	mid 1990s	66	1507	329	273	231	165	376	68	61	54	39
	late 2000s	99										

2.3. Change detection accuracy assessment

When comparing the relevant changes detected based on imperviousness estimations to the reference, four cases are possible (Klaric, 2014):

- true positive (TP) results: relevant change is present in ground truth data and it is detected by the model;
- false positive (FP) results: model predicts the relevant change where there is no change in reference or the change is present in reference but it is not relevant;
- false negative (FN) results: relevant change is present in ground truth data, but it is not detected by the model;
- true negative (TN) results: there is no change in reference or the change is present in reference but it is not relevant and the model predicts no relevant change as well.

These values are often presented as a confusion matrix and several common metrics of model performance may be calculated based on them (Fawcett, 2006).

The most popular one in remote sensing applications is the accuracy:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

What accuracy level is acceptable depends on the application. However, the random classifier has the overall accuracy of 0.5.

In case of imbalanced datasets accuracy alone is not sufficient. If change is rare, we can easily achieve very high accuracy by predicting all cases as no change. But such model is completely useless.

Cohen's kappa statistic (Cohen, 1960; Landis and Koch, 1977), often used in remote sensing applications for the assessment of classification results, may be seen as better performance measure in case of imbalanced class problems. It is defined as:

$$\kappa = \frac{Accuracy - p_e}{1 - p_e}$$

where p_e is the expected agreement defined as the sum of the products of reference likelihood and result likelihood for each class. In case of change detection confusion matrix, it can be calculated as (Santafe et al., 2015):

$$p_e = \frac{(TP + FN) \cdot (TP + FP)}{(TP + FP + FN + TN)^2} + \frac{(TN + FP) \cdot (TN + FN)}{(TP + FP + FN + TN)^2}$$

The usefull classifier should have kappa values higher then zero.

Although Cohen's kappa is more appropriate then overall accuracy for application as the performance measure in case of imbalanced datasets, in change detection studies recall, precision and F-measure are used as the standard performance measures (Klaric, 2014; Wieland et. al., 2016):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$F1 = \frac{2}{1/\text{Precision} + 1/\text{Recall}} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision+Recall}} = \frac{2 * \text{TP}}{2 * \text{TP+FP+FN}}$$

Especially F-measure is commonly-used in evaluation as it takes into account both precision and recall (Klaric, 2014). In case of imbalanced datasets precision and F-measure are able to show differences in model performances that are not revealed when using accuracy (Saito and Rehmsmeier, 2015). The useful prediction model should have the precision value higher than precision of random classifier (PRC) given by the formula (Saito and Rehmsmeier, 2015):

$$\text{PRC} = \frac{P}{\text{P+N}} = \frac{\text{TP+FN}}{\text{TP+FN+FP+TN}}$$

where P stands for the number of positive (change) and N for the number of negative (no change) instances in dataset (reference).

Other approach often utilized in evaluation of change detection model performances is based on Receiver Operating Characteristic (ROC) graphs (Fawcett, 2006; Wieland et. al., 2016; Aleksandrowicz et al., 2016). ROC graph is a two-dimensional graph of the true positive rate (recall) plotted against the false positive rate (fpr), calculated as:

$$\text{fpr} = \frac{\text{FP}}{\text{FP+TN}}$$

ROC graphs are most often used for so called scoring classifiers. Such classifiers yield a value (probability or score) representing a degree to which classified instance belongs to a class. Their results may be thresholded to produce discrete (binary) classifiers. Plotting recall against fpr for various thresholds gives so called ROC curve (Fawcett, 2006).

The ROC graph may also be used to compare the performance of discrete classifiers. In this case the classifier performance is represented by one point in the ROC space. Two measures may be used for comparison of the discrete classifier performances (Fawcett, 2006; Powers, 2011): the distance to the optimum point (fpr = 0 and recall = 1) and the ROC Area Under Curve (ROC_{AUC}) which in case of discrete classifier is defined as (Powers, 2011):

$$ROC_{AUC} = \frac{\text{recall} - \text{fpr} + 1}{2}$$

The latter is commonly used for comparison of classifiers (Amancio et al., 2014). The ROC_{AUC} values are between 0 and 1, but the useful prediction model should have the value over 0.5 (the value of random guessing).

As Amancio et al. (2014) note, it is not possible to fully compare the performance of classifiers with a single metric. In this study we used the overall accuracy, F-measure and the ROC Area Under Curve (ROC_{AUC}).

2.4. Comparison of machine learning models

The question if one (usually newly proposed) machine learning algorithm performs better than a competitor or better than the state of art algorithms is very frequent. To answer such a question appropriate statistical tests can be used to evaluate the results obtained with particular models (Trawiński et al., 2012; Santafe et al., 2015). Different scenarios may be considered depending on the number of algorithms in comparison and the number of available datasets (Japkowicz and Shah, 2011; Santafe et al., 2015): two or many algorithms may be compared on one dataset or in several datasets.

When two algorithms are compared using a single dataset the dataset is usually resampled using cross-validation or bootstrap. For each data split the selected score (eg. the error of estimation) is calculated for both algorithms. Differences of the scores are then statistically tested to find if the two algorithms of interest differ in performance with respect to the score or not (Santafe et al., 2015). Usually parametric tests are used assuming the normal distribution of score differences. However, if the number of estimations (resamples) is not large enough this assumption should be verified with appropriate test. If the distribution is not normal non-parametric alternatives to *t*-test should be used such as Wilcoxon signed-rank test (Wilcoxon, 1945) or sign test.

Application of standard paired *t*-test for comparison of two algorithms on one resampled dataset is generally considered as unsafe due to underestimation of the statistic variance caused by the violation of the assumption of the score values independence (Santafe et al., 2015). One can find in literature several proposals of modified or corrected statistical test for different resampling methods (Nadeau and Bengio, 2003; Bouckaert, 2004; Bouckaert and Frank, 2004; Japkowicz and Shah, 2011; Santafe et al., 2015). Corrected resampled *t*-test (Nadeau and Bengio, 2003) and corrected *t*-test for repeated cross-validation (Bouckaert and Frank, 2004) are the most often recommended (Santafe et al., 2015). The corrected statistic for r-times repeated k-fold cross-validation is calculated as (Bouckaert and Frank, 2004):

$$t = \frac{\frac{1}{k \cdot r}\sum_{i=1}^{k}\sum_{j=1}^{r} x_{ij}}{\sqrt{(\frac{1}{k \cdot r} + \frac{n_2}{n_1})\hat{\sigma}^2}}$$

where

$$\hat{\sigma}^2 = \frac{1}{k \cdot r - 1}\sum_{\mathrm{i}=1}^{k}\sum_{\mathrm{j}=1}^{r}(x_{\mathrm{ij}} - m)^2$$

$$\mathrm{m} = \frac{1}{k \cdot r}\sum_{\mathrm{i}=1}^{k}\sum_{\mathrm{j}=1}^{r} x_{\mathrm{ij}}$$

and x_{ij} is observed difference of algorithm scores for cross-validation fold i and run j, n_1 is the number of instances used for training, and n_2 the number of instances used for testing.

The other possible solution is to average the results of repeated cross-validation over runs (Bouckaert, 2004; Japkowicz and Shah, 2011). In this approach the results of a k-fold cross-validation are sorted in every single run from the lowest to the highest one. Then averaging is done on each fold over all runs. As a result, the estimate for the minimum value is calculated from the minimum values in all folds, the one but lowest from the one but lowest results in all folds, etc.. According to Bouckaert (2004) such sorted runs sampling scheme should be combined with *t*-test to achieve the best results.

The independence between the score values of two algorithms may be obtained using independent validation dataset. Unfortunately, such approach may give a pessimistically biased estimation of the error. Because of that, the approaches based on dataset resampling are generally preferred (Santafe et al., 2015).

Nevertheless, when the accuracy of two classifiers is being compared using the same independent validation dataset McNemar's test is usually applied (Santafe et al., 2015). The other possibility is to use the approach based on confidence intervals

(Foody, 2009). Confidence intervals can be calculated not only for the accuracies but also for some other measures. The method of confidence intervals estimation for ROC_{AUC} was proposed for example by Hanley and McNeil (1982).

The significance tests cannot be applied directly to compare the performance of two classifiers on validation dataset using composite metrics like F-measure, since they do not have any probabilistic interpretation (Joshi, 2002). However, recall and precision have. To compare the performance of classifiers for rare classes Joshi (2002) suggested comparison of improvements of recall and precision using *p*-test proposed by Yang and Liu (1999) and direct comparison of F-measures only when *p*-test outcomes on recall and precision are not in agreement.

Comparing several machine learning algorithms using only one dataset creates additional problems. Both, parametric statistical tests for multiple comparisons such as ANOVA (Fisher, 1937) and non-parametric omnibus tests such as Friedman test (Friedman, 1940) should not be used with one resampled dataset due to the lack of independence among the obtained score values. The only possibility is to compare algorithms by pairs and apply multiple comparison corrections to control family-wise error (Santafe et al., 2015). Several procedures for such corrections have been proposed in literature. The simplest one, but also the least powerful is the Bonferroni-Dunn procedure (Dunn, 1961). In the case of 1:N comparisons (i.e. when the one algorithm is compared to the others) the Finner method (Finner, 1993) is suggested as a good choice (Trawiński et al., 2012; Santafe et al., 2015)

When several datasets are available the other approaches to algorithm comparison are possible. If two algorithms are considered Wilcoxon signed-rank test or sign test are usually used. The non-parametric tests are preferred as the scores obtained on different datasets are hardly commensurable (Santafe et al., 2015). In case of multiple algorithm comparisons an omnibus test is done first to compare all algorithms together and decide if they have the same performance. If not, the algorithms are compared by pairs using post-hoc tests. Of course, we should apply the corrections to control family-wise error as well. Again, non-parametric omnibus tests are prefered (Demsar, 2006; Trawiński et al., 2012; Santafe et al., 2015). If the number of algorithms in comparison is higher than five the Friedman test with Iman and Davenport extension (Iman and Davenport, 1980) is recommended. If number of the models is lower the Friedman aligned ranks (Daniel, 1990; Garcia et al., 2010) or the Quade tests are considered as more useful (Santafe et al., 2015).

In presented study the performance of machine learning models was evaluated for the assessment of ISA in individual time points and assessment of ISA change. The latter was done both for prediction of change intensity and detection of relevant changes. Evaluation was done based on cross-validation results obtained for calibration datasets and using independent validation datasets. Two scenarios for algorithm comparison were used. In all comparisons the significance level of $\alpha = 0.05$ was applied to decide about the rejection of null hypotheses.

2.4.1 Scenario 1: comparison of multiple algorithms on individual datasets

In the first scenario, multiple machine learning models were compared on individual datasets. The 1:N approach was used, i.e. the algorithm with the best score (RMSE,MAE or R^2 in case of ISA and ISA change intensity or accuracy, F-measure and ROC_{AUC} for ISA relevant change detection) was compared to the others. As a result for each dataset and each performance measure the best algorithm and the algorithms with no statistically significant difference in performance were found.

For calibration datasets the differences in scores of cross-validation realisations were compared directly. For all performance measures the Shapiro-Wilk normality test were used to check the distribution of differences. For RMSE and MAE no violations for normality assumption were find and the paired corrected *t*-test for repeated cross-validation (Bouckert and Frank, 2004) was used. In case of R^2 the differences were not normally distributed. Because of that the sorted runs scheme (Bouckaert, 2004) was applied. As there was no evidence to reject the hypothesis about normal distribution of obtained averaged differences, the *t*-test was used for their evaluation. For correction of family-wise error in multiple comparisons we used the procedure proposed by Finner (1993).

In case of ISA change detection capabilities evaluation, the performance of individual methods was compared to the performance of random classifier based on accuracy, precision and ROC_{AUC} scores. This check procedure was applied for every of 50 realisations of cross-validation in every dataset. Only algorithms which passed this test were analysed further. The procedure applied may be considered as very conservative. On the other hand, such approach may be seen as very safe as any underperformance eliminated the algorithm from further considerations.

In case of independent validation datesets paired *t*-test were used to compare squared errors and absolute errors for every pixel. Instead of R^2, we compared the differences in correlation coefficients. Two approaches were used for this purpose (Diedenhofen and Musch, 2015): the test for the difference of two dependent correlations as proposed in Steiger (1980) and the method of Zou (2007) based on confidence intervals. McNemar test was used to compare achieved accuracies. In case of ROC_{AUC} measure the approaches were assessed as significantly different only when their confidence intervals calculated according Hanley and McNeil (1982) did not overlap. The performance of best model chosen according to F-measure was compared to others using the approach proposed by Joshi (2002).

2.4.2 Scenario 2: comparison of multiple algorithms on multiple datasets

In the second scenario multiple algorithms were compared on multiple datasets. This scenario was possible to implement only for single time-step evaluations of ISA, as two image datasets were available in every of three study areas (six datasets altogether). For change assessment the number of datasets (three) was too low. In this

scenario we used the Friedman test and then the best algorithm was compared to the remaining ones using post-hoc tests. Again, the family-wise error was controlled with the Finner procedure. Like in the first scenario two comparisons were done – one based on cross-validation and the second for the scores obtained with independent validation dataset.

3. Results

3.1. Comparison of individual machine learning algorithms performances for ISA mapping

3.1.1 Single time points

This subsection presents the results obtained using individual machine learning approaches for single time-point sub-pixel ISA mapping in researched study areas. Table 2 presents average values of RMSE and MAE and Table 3 average values of R^2 obtained in cross-validation procedure on calibration dataset. The RMSE and MAE values obtained with tested algorithms for single time points ISA assessment on independent validation datasets are presented in Table 4 and R^2 values in Table 5. The best results are bolded.

The best performed algorithms were compared to the others and obtained *p*-values adjusted for multiple comparisons are presented in Table 6–9. Bolded entries indicate the cases when algorithms performance do not differ significantly from the best one.

Table 2. Model performances (RMSE and MAE) on calibration datasets – ISA evaluation for single time steps (Drzewiecki, 2016b)

	Raba datasets				Dunajec datasets				Sola datasets			
Method	Mid 1990s		Late 2000s		Mid 1990s		Late 2000s		Mid 1990s		Late 2000s	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
avNN	0.1149	0.0680	0.1154	0.0675	0.1623	0.1098	0.1414	0.0961	0.1338	0.0822	0.1360	0.0861
RF	0.1099	0.0653	0.1116	0.0650	**0.1590**	0.1089	0.1390	0.0956	0.1257	0.0787	0.1210	0.0756
Cubist	0.1074	0.0608	**0.1115**	**0.0619**	0.1609	0.1070	**0.1312**	**0.0876**	0.1213	**0.0720**	**0.1178**	**0.0703**
GBM	**0.1073**	0.0617	0.1119	0.0640	0.1594	0.1068	0.1324	0.0893	**0.1205**	0.0728	0.1196	0.0723
kNN	0.1080	**0.0599**	0.1166	0.0656	0.1668	0.1063	0.1434	0.0932	0.1289	0.0740	0.1276	0.0755
rkNN	0.1074	0.0620	0.1165	0.0664	0.1716	0.1131	0.1464	0.0979	0.1293	0.0778	0.1288	0.0799
SVMp	0.1267	0.0734	0.1132	0.0643	0.1839	0.1160	0.1462	0.0978	0.1279	0.0776	0.1254	0.0768
SVMr	0.1085	0.0613	0.1153	0.0647	0.1698	**0.1062**	0.1350	0.0927	0.1255	0.0762	0.1262	0.0767
MARS	0.1191	0.0731	0.1179	0.0680	0.1805	0.1245	0.1616	0.1113	0.1346	0.0836	0.1310	0.0799

Table 3. Model performances (R^2) on calibration datasets – ISA evaluation for single time steps

	Raba datasets		Dunajec datasets		Sola datasets	
Method	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s
	R^2	R^2	R^2	R^2	R^2	R^2
avNN	0.6871	0.7495	0.6917	0.8060	0.7946	0.8127
RF	0.7151	0.7670	**0.7050**	0.8144	0.8194	0.8519
Cubist	**0.7270**	**0.7670**	0.6969	**0.8321**	0.8312	**0.8589**
GBM	0.7266	0.7649	0.7027	0.8302	**0.8332**	0.8545
kNN	0.7235	0.7447	0.6747	0.8010	0.8090	0.8345
rkNN	0.7256	0.7463	0.6549	0.7928	0.8084	0.8323
SVMp	0.6459	0.7592	0.6082	0.7926	0.8118	0.8401
SVMr	0.7213	0.7504	0.6650	0.8236	0.8188	0.8378
MARS	0.6644	0.7391	0.6188	0.7470	0.7922	0.8251

Table 4. Model performances (RMSE, MAE) on validation datasets – ISA evaluation for single time steps (Drzewiecki, 2016b)

	Raba datasets				Dunajec datasets				Sola datasets			
Method	Mid 1990s		Late 2000s		Mid 1990s		Late 2000s		Mid 1990s		Late 2000s	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
avNN	0.1218	0.0736	0.1131	0.0679	**0.1496**	0.1041	0.1396	0.0954	0.1156	0.0702	0.1386	0.0818
RF	0.1128	0.0678	0.1120	0.0647	0.1500	0.1051	0.1368	0.0951	0.1062	0.0645	0.1113	0.0676
Cubist	**0.1086**	**0.0624**	**0.1084**	**0.0611**	0.1538	0.1008	0.1441	0.0948	0.1074	0.0608	0.1158	0.0670
GBM	0.1117	0.0639	0.1116	0.0639	0.1509	0.1039	0.1335	**0.0895**	**0.1002**	**0.0592**	0.1171	0.0692
kNN	0.1117	0.0636	0.1182	0.0680	0.1670	0.1046	0.1478	0.0961	0.1203	0.0641	0.1104	**0.0633**
rkNN	0.1098	0.0640	0.1167	0.0684	0.1638	0.1101	0.1452	0.0977	0.1139	0.0656	0.1113	0.0670
SVMp	0.1334	0.0783	0.1091	0.0625	0.1543	0.0987	0.1384	0.0941	0.1206	0.0689	**0.1104**	0.0648
SVMr	0.1154	0.0665	0.1120	0.0654	0.1533	**0.0962**	**0.1332**	0.0917	0.1061	0.0617	0.1154	0.0690
MARS	0.1248	0.0768	0.1110	0.0629	0.1662	0.1176	0.1534	0.1104	0.1199	0.0736	0.1245	0.0745

Table 5. Model performances (R^2) on validation datasets – ISA evaluation for single time steps

	Raba datasets		Dunajec datasets		Sola datasets	
Method	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s
avNN	0.6824	0.79	0.7260	0.8001	0.8540	0.8016
RF	0.7304	0.7924	**0.7270**	0.8091	0.8761	0.8737
Cubist	**0.7477**	**0.8069**	0.7083	0.7881	0.8696	0.8619
GBM	0.7344	0.7941	0.7210	0.8169	**0.8870**	0.8585
kNN	0.7334	0.7689	0.6642	0.7781	0.8342	0.8746
rkNN	0.7434	0.7768	0.6747	0.7853	0.8556	0.8741
SVMp	0.6512	0.8060	0.7072	0.8045	0.8342	**0.8748**
SVMr	0.7164	0.7930	0.7095	**0.8177**	0.8723	0.8627
MARS	0.6714	0.7960	0.6615	0.7581	0.8366	0.8404

Table 6. Model performances on calibration datasets for individual time points ISA assessments (*p*-values for H0 = performance of the algorithm does not differ from the best model) – RMSE and MAE

	Raba datasets				Dunajec datasets				Sola datasets			
Method	Mid 1990s		Late 2000s		Mid 1990s		Late 2000s		Mid 1990s		Late 2000s	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
avNN	0.004	<0.001	**0.123**	<0.001	**0.450**	**0.360**	0.003	<0.001	<0.001	<0.001	<0.001	<0.001
RF	**0.094**	<0.001	**0.988**	<0.001	**BEST**	**0.360**	0.003	<0.001	0.005	<0.001	**0.179**	0.001
Cubist	**0.962**	**0.246**	**BEST**	**BEST**	**0.457**	**0.833**	**BEST**	**BEST**	**0.675**	**BEST**	**BEST**	**BEST**
GBM	**BEST**	**0.083**	**0.860**	0.010	**0.831**	**0.833**	**0.699**	**0.243**	**BEST**	**0.500**	**0.418**	**0.134**
kNN	**0.814**	**BEST**	0.017	<0.001	**0.056**	**0.974**	<0.001	0.001	0.037	**0.255**	0.010	0.004
rkNN	**0.814**	<0.001	0.009	<0.001	0.001	0.025	<0.001	<0.001	0.008	<0.001	0.003	<0.001
SVMp	<0.001	<0.001	**0.273**	0.004	<0.001	<0.001	<0.001	<0.001	0.005	<0.001	0.026	0.004
SVMr	**0.758**	**0.185**	**0.123**	0.007	0.006	**BEST**	**0.320**	**0.015**	0.025	0.007	0.015	0.001
MARS	<0.001	<0.001	0.022	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.005	<0.001

Table 7. Model performances on calibration datasets for individual time points ISA assessments (*p*-values for H0 = performance of the algorithm does not differ from the best model) – R^2

	Raba datasets		Dunajec datasets		Sola datasets	
Method	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s
avNN	0.021	**0.137**	**0.350**	0.027	0.034	0.029
RF	**0.205**	**1.000**	**BEST**	0.027	**0.053**	**0.238**
Cubist	**BEST**	**BEST**	**0.350**	**BEST**	**0.664**	**BEST**
GBM	**0.940**	**0.768**	**0.681**	**0.682**	**BEST**	**0.391**
kNN	**0.716**	**0.104**	**0.086**	0.018	**0.071**	**0.067**
rkNN	**0.842**	**0.104**	0.030	0.018	**0.053**	**0.067**
SVMp	0.015	**0.173**	0.008	0.018	0.045	**0.067**
SVMr	**0.716**	**0.137**	0.046	**0.301**	**0.057**	**0.067**
MARS	0.015	**0.104**	0.008	0.008	0.011	**0.067**

The results obtained from repeated cross-validation show that Cubist algorithm outperformed the others. If RMSE and R^2 are considered, Cubist and GBM algorithms are the best or do not differ significantly from the best ones. The same is true for the Cubist in case of MAE. The result of GBM is significantly worse for one dataset only (Raba late 2000s).

Table 8. Model performances on validation datasets for individual time points ISA assessments (*p*-values for H0 = performance of the algorithm does not differ from the best model) – squared and absolute errors

	Raba datasets				Dunajec datasets				Sola datasets			
Method	Mid 1990s		Late 2000s		Mid 1990s		Late 2000s		Mid 1990s		Late 2000s	
	squared error	absolute error	squared error	absolute error	squared error	absolute error	squared error	absolute error	squared error	absolute error	squared error	absolute error
avNN	0.003	<0.001	**0.135**	<0.001	**BEST**	**0.068**	**0.287**	**0.141**	**0.111**	0.002	0.015	<0.001
RF	**0.178**	**0.141**	**0.392**	0.009	**0.912**	0.049	**0.426**	0.009	**0.192**	0.024	**0.759**	**0.324**
Cubist	**BEST**	**BEST**	**BEST**	**BEST**	**0.696**	**0.238**	0.024	**0.131**	**0.192**	**0.850**	**0.795**	**0.324**
GBM	**0.353**	**0.365**	**0.392**	0.042	**0.800**	**0.067**	**0.939**	**BEST**	**BEST**	**BEST**	**0.759**	**0.155**
kNN	**0.353**	**0.447**	**0.076**	0.001	**0.060**	**0.078**	0.024	**0.131**	**0.111**	**0.397**	**0.759**	**BEST**
rkNN	**0.543**	**0.274**	**0.076**	<0.001	**0.063**	0.002	0.025	0.032	**0.111**	**0.075**	**0.795**	**0.091**
SVMp	<0.001	<0.001	**0.756**	**0.339**	**0.696**	**0.319**	**0.351**	**0.214**	**0.192**	0.042	**BEST**	**0.347**
SVMr	**0.307**	**0.062**	**0.316**	0.007	**0.696**	**BEST**	**BEST**	**0.464**	**0.192**	**0.397**	**0.759**	**0.131**
MARS	<0.001	<0.001	**0.552**	**0.339**	**0.060**	<0.001	0.024	<0.001	**0.192**	<0.001	**0.195**	0.039

Table 9. Evaluation of model performances on validation datasets for individual time points ISA assessments (*p*-values for H0 = performance of the algorithm does not differ from the best model) – correlation coefficients

	Raba datasets		Dunajec datasets		Sola datasets	
Method	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s
avNN	<0.001	0.012	**0.940**	**0.200**	<0.001	<0.001
RF	0.025*	0.005	**BEST**	**0.432**	0.021	**0.954**
Cubist	**BEST**	**BEST**	**0.208**	0.010	<0.001	**0.300**
GBM	**0.083**	0.012	**0.578**	**0.940**	**BEST**	**0.195**
kNN	**0.113**	<0.001	0.001	0.003	<0.001	**0.970**
rkNN	**0.400**	<0.001	0.001	0.011	<0.001	**0.960**
SVMp	<0.001	**0.890**	**0.272**	**0.245**	<0.001	**BEST**
SVMr	0.004	0.044	**0.306**	**BEST**	0.023	**0.260**
MARS	<0.001	**0.108**	0.001	<0.001	<0.001	0.004

** – does not differ significantly from the best model according to the method of Zou (2007)*

In case of squared errors obtained for control pixels of independent validation datasets Random Forest, GBM and SVMr algorithms are the best models or do not differ significantly from the best ones at $\alpha = 0.05$. Cubist, kNN and rkNN models gave results with no difference from the best one with level of significance not worse than $\alpha = 0.01$. In case of absolute errors only the Cubist algorithm was the best or without significant difference from the best one at $\alpha = 0.05$. GBM algorithm gave the results not different from the best scores at $\alpha = 0.01$. When differences in correlation coefficients are considered we cannot find any model to be constantly the best one or without significant difference from the best one at $\alpha = 0.05$. However, GBM model seems to be the best performer, as it is the only one which for every dataset gave the results not different from the best scores at $\alpha = 0.01$.When all three measures are considered together GBM and Cubist algorithms may be pointed out as the best performers on validation dataset as they were in cross-validation approach as well.

Table 10 presents average ranks obtained by each algorithm in the Friedman tests done for RMSE, MAE and R^2 scores calculated for different datasets in cross-validation and independent dataset validation approaches. In both cases the Cubist algorithm has the highest position for Mean Absolute Error. It is also the best when RMSE and R^2 results from cross-validation are considered. In case of validation datasets the GBM model has the highest average rank for RMSE and R^2.

Table 10. Model performance for individual time points ISA assessments – average ranks of Friedman tests

Method	Ranking (cross-validation)			Ranking (validation datasets)		
	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	6.17	7.33	6.50	5.83	7.00	6.33
RF	3.67	5.33	3.08	3.67	5.50	3.5
Cubist	**1.75**	**1.67**	**1.58**	3.83	**2.42**	3.83
GBM	2.08	2.50	2.00	**3.42**	3.33	**3.33**
kNN	5.42	3.17	5.83	6.50	4.67	6.42
rkNN	5.58	6.75	6.33	5.42	6.42	5.17
SVMp	7.17	6.33	6.50	5.25	4.17	5.08
SVMr	4.67	3.33	4.67	3.92	3.83	4.00
MARS	8.5	8.58	8.5	7.17	7.67	7.33

The best algorithms were compared to others in 1:N post-hoc analysis. Adjusted *p*-values obtained through the application of Finner procedure are presented in Table 11. Bolded entries indicate the cases when algorithms do not differ significantly from the control one. According to Friedman tests done for the results from cross-validation, Cubist, GBM and Supprot Vector Machines with radial kernel algorithms may be considered as the best ones for single time-point ISA assessment. For validation datasets the differences between scores of individual algorithms were much lower. In case of RMSE and R^2 no algoritm may be considered as significantly different from the best one (*p*-value computed by Friedman test with Iman and Davenport correction are 0.149 and 0.101, respectively).

Table 11. Adjusted p-values for 1xN comparisons of algorithms for Friedman post-hoc tests

Method	Adjusted p-value (cross-validation)			Adjusted p-value (validation datasets)		
	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	0.013851	0.001353	0.007473	**0.302581**	0.014902	**0.189491**
RF	**0.25319**	0.032432	0.381032	**0.874367**	**0.099717**	**0.916051**
Cubist	**control algorithm**	**control algorithm**	**control algorithm**	**0.844045**	**control algorithm**	**0.796631**
GBM	**0.833029**	**0.598161**	**0.792147**	**control algorithm**	**0.562083**	**control algorithm**
kNN	0.032432	**0.381032**	0.011478	**0.189491**	**0.235824**	**0.189491**
rkNN	0.030431	0.003475	0.007473	**0.36941**	0.030143	**0.431863**
SVMp	0.00245	0.006316	0.007473	**0.36941**	**0.340756**	**0.431863**
SVMr	**0.085827**	**0.368789**	**0.067635**	**0.844045**	**0.410523**	**0.774982**
MARS	0.000157	0.000097	0.000097	**0.133174**	0.007169	**0.087732**

3.1.2 Estimation of ISA change intensity

This subsection presents the results obtained using individual machine learning approaches for sub-pixel mapping of ISA change intensities. Change intensity means the difference of ISA values estimated for two points in time. Table 12 presents average values of performance measures obtained in cross-validation procedure on calibration dataset. The values obtained with tested algorithms for ISA change intensity assessment on independent validation datasets are presented in Table 13. The best performed algorithms were compared to the others and obtained *p*-values adjusted for multiple comparisons using Finner procedure are presented in Table 14 and Table 15. Bolded entries indicate the cases when algorithms do not differ significantly from the best one.

Table 12. ISA models performance (average valuess) on calibration datasets – change asseement

Method	Raba dataset			Dunajec datasets			Sola datasets		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	0.1053	0.0626	0.2662	0.1467	0.1043	0.5474	0.1290	0.8340	0.2876
RF	**0.0902**	0.0520	**0.3129**	**0.1286**	**0.0889**	**0.6104**	**0.1077**	**0.0667**	0.3594
Cubist	0.0953	0.0531	0.2852	0.1389	0.0954	0.5881	0.1124	0.0695	**0.3630**
GBM	0.0947	0.0540	0.2970	0.1346	0.0923	0.5959	0.1096	0.0677	0.3628
kNN	0.0978	0.0543	0.2515	0.1516	0.1004	0.5369	0.1257	0.0760	0.2670
rkNN	0.0917	**0.0505**	0.2795	0.1497	0.1015	0.5333	0.1163	0.0704	0.2894
SVMp	0.1095	0.0640	0.2522	0.1538	0.1032	0.5088	0.1107	0.0685	0.3552
SVMr	0.0967	0.0575	0.3032	0.1474	0.1000	0.5410	0.1199	0.0746	0.2883
MARS	0.1081	0.0656	0.2288	0.1655	0.1158	0.4349	0.1290	0.0791	0.2905

Table 13. ISA models performance on validation datasets – change asseement

Method	Raba dataset			Dunajec datasets			Sola datasets		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	0.0998	0.0629	0.2810	0.1383	0.0987	0.5657	0.1396	0.0856	0.2365
RF	**0.0853**	0.0504	0.3136	0.1316	0.0889	0.5775	0.1060	0.0632	0.3885
Cubist	0.0875	0.0505	0.2958	0.1494	0.0954	0.5237	0.1176	0.0692	0.3604
GBM	0.0873	0.0505	0.2977	**0.1307**	**0.0873**	**0.5898**	0.1157	0.0688	0.3540
kNN	0.0938	0.0534	0.2435	0.1493	0.0952	0.5446	0.1227	0.0687	0.3379
rkNN	0.0863	**0.0486**	0.2993	0.1403	0.0929	0.5724	**0.1057**	**0.0618**	0.3960
SVMp	0.1012	0.0620	0.3324	0.1366	0.0920	0.5440	0.1097	0.0634	**0.4484**
SVMr	0.0854	0.0522	**0.3503**	0.1364	0.0921	0.5736	0.1139	0.0682	0.3764
MARS	0.0915	0.0594	0.3404	0.1542	0.1069	0.4443	0.1330	0.0809	0.2786

Table 14. ISA models performance on calibration datasets – change asseement
(*p*-values for H0 = performance of the algorithm does not differ from the best model)

Method	Raba dataset			Dunajec datasets			Sola datasets		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	<0.001	<0.001	**0.122**	<0.001	<0.001	0.043	<0.001	<0.001	**0.151**
RF	**BEST**	**0.159**	**BEST**	**BEST**	**BEST**	**BEST**	**BEST**	**BEST**	**0.888**
Cubist	0.004	0.019	**0.177**	0.015	0.012	**0.246**	**0.171**	**0.133**	**BEST**
GBM	0.004	0.009	**0.198**	0.015	**0.058**	**0.190**	**0.442**	**0.533**	**0.993**
kNN	0.002	<0.001	**0.122**	<0.001	<0.001	0.025	<0.001	<0.001	**0.070**
rkNN	**0.474**	**BEST**	**0.177**	<0.001	<0.001	0.025	0.023	**0.105**	**0.151**
SVMp	<0.001	<0.001	**0.122**	<0.001	<0.001	0.017	**0.446**	**0.442**	**0.844**
SVMr	<0.001	<0.001	**0.484**	<0.001	<0.001	0.036	0.005	<0.001	**0.151**
MARS	<0.001	<0.001	**0.057**	<0.001	<0.001	0.016	<0.001	<0.001	**0.151**

Table 15. ISA models performance on validation datasets – change asseement
(*p*-values for H0 = performance of the algorithm does not differ from the best model)

Method	Raba dataset			Dunajec datasets			Sola datasets		
	squared error	absolute error	correlation coefficient	squared error	absolute error	correlation coefficient	squared error	absolute error	correlation coefficient
avNN	0.002	<0.001	**0.051***	**0.340**	**0.057**	**0.535**	<0.001	<0.001	<0.001
RF	**BEST**	**0.403**	**0.169**	**0.805**	**0.549**	**0.535**	**0.785**	**0.942**	**0.067**
Cubist	**0.590**	**0.403**	**0.051***	**0.066**	**0.092**	0.006	**0.201**	**0.155**	0.012
GBM	**0.590**	**0.403**	**0.051***	**BEST**	**BEST**	**BEST**	**0.201**	**0.155**	0.007
kNN	**0.091**	<0.001	<0.001	**0.066**	**0.190**	**0.174**	0.023	**0.149**	0.007
rkNN	**0.819**	**BEST**	**0.051***	**0.282**	**0.280**	**0.535**	**BEST**	**BEST**	**0.110**
SVMp	0.003	<0.001	**0.588**	**0.414**	**0.333**	**0.174**	**0.519**	**0.591**	**BEST**
SVMr	**0.972**	**0.132**	**BEST**	**0.414**	**0.319**	**0.535**	**0.286**	**0.155**	0.012
MARS	**0.091**	<0.001	**0.710**	**0.012**	<0.001	<0.001	**0.104**	<0.001	<0.001

** – differs significantly from the best model according to the method of Zou (2007)*

When looking into cross-validation results from the ISA change intensity mapping point of view, the Random Forest outperforms other approaches. It gave the lowest RMS errors in every area, the highest R^2 for Raba and Dunajec catchments and the lowest MAE for Dunajec and Sola datasets. In case of Raba dateset, there is no significant difference between RF and the best model for MAE score. The same is true for R^2 in Sola area. In case of RMSE score, all other algorithms differ significantly in at least one area.

Random Forest and random k-nearest neighbors algorithms are also the only ones which are the best or not significantly different from the best ones for all performance measures in every study area when independent validation datasets are taken into consideration. If differences in correlation coefficients are evaluated based on their confidence intervals rknn algorithm performance for Raba dataset is significantly worse then performance of radial SVM (the best one). Nevertheless, the difference between correlation coefficients of radial SVM and Random Forest models is still not significant.

3.1.3 Estimation of ISA change detection capabilities

Analysis of accuracy scores from cross-validation revealed the fact that it is risky to use the differences of ISA maps for detection of subtle ISA changes. For Dunajec dataset the accuracies of ISA change maps created with 1% and 3% thresholds were not higher than the accuracy of random classifier (Table 16). In case of 5% threshold the same appeared in at least one cross-validation realisation for the most of algorithms. Only three of them (Random Forest, Cubist and kNN) passed the test, but with very low results. For the change maps based on 10% change threshold the accuracies were much higher.

Table 16. ISA models performance on calibration datasets – change detection (minimal values of overall accuracy measure). Scores below 0.5 are bolded

Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
avNN	**0.459**	0.571	0.652	0.749	**0.238**	**0.336**	**0.422**	0.606	**0.331**	**0.452**	0.538	0.679
RF	0.564	0.610	0.610	0.799	**0.246**	**0.391**	0.502	0.682	**0.414**	0.533	0.616	0.744
Cubist	0.591	0.608	0.682	0.794	**0.274**	**0.394**	0.502	0.664	**0.430**	0.515	0.565	0.711
GBM	0.543	0.623	0.681	0.786	**0.249**	**0.386**	**0.480**	0.663	**0.419**	0.542	0.621	0.748
kNN	0.615	0.634	0.671	0.784	**0.319**	**0.397**	0.502	0.663	**0.465**	0.526	0.585	0.708
rkNN	0.591	0.621	0.688	0.803	**0.278**	**0.359**	**0.477**	0.706	**0.455**	0.508	0.571	0.732
SVMp	**0.471**	0.594	0.658	0.766	**0.254**	**0.354**	**0.464**	0.645	**0.374**	**0.490**	0.588	0.745
SVMr	**0.462**	0.582	0.682	0.771	**0.242**	**0.351**	**0.457**	0.650	**0.342**	**0.495**	0.568	0.725
MARS	**0.445**	0.530	0.602	0.752	**0.170**	**0.285**	**0.377**	0.565	**0.349**	**0.482**	0.555	0.705

When precision scores are considered the results are a little better (Table 17). The similar pattern is visible in Table 18, where minimum values of ROC_{AUC} score are presented (useful classifiers should have value above 0.5).

Table 17. ISA models performance on calibration datasets – change detection (comparison of precision with PRC). Bolded entries indicate performances below PRC

	Number of cases with precision below PRC treshold											
Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
avNN	**40**	0	0	0	**49**	**5**	0	0	**48**	**6**	0	0
RF	0	0	0	0	**42**	**2**	0	0	**14**	0	0	0
Cubist	0	0	0	0	**42**	0	0	0	**3**	0	0	0
GBM	**5**	0	0	0	**42**	0	0	0	**12**	0	0	0
kNN	**2**	0	0	0	**33**	0	0	0	**4**	0	0	0
rkNN	0	0	0	0	**38**	0	0	0	**8**	0	0	0
SVMp	**4**	0	0	0	**40**	**1**	0	0	**26**	**1**	0	0
SVMr	**31**	0	0	0	**44**	**2**	0	0	**48**	0	0	0
MARS	**44**	**1**	0	0	**49**	**17**	**2**	0	**31**	0	0	0

Table 18. ISA models performance on calibration datasets – change detection (minimal values of ROC_{AUC} measure). Scores below 0.5 are bolded

Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
avNN	**0.391**	0.526	0.582	0.629	**0.369**	**0.444**	0.527	0.679	**0.363**	**0.475**	0.529	0.591
RF	0.502	0.568	0.605	0.610	**0.385**	**0.490**	0.582	0.721	**0.436**	0.537	0.573	0.648
Cubist	0.513	0.554	0.586	0.603	**0.397**	0.509	0.577	0.696	**0.475**	0.536	0.566	0.621
GBM	**0.465**	0.555	0.619	0.602	**0.399**	0.516	0.568	0.697	**0.459**	0.546	0.585	0.636
kNN	**0.475**	0.539	0.565	0.597	**0.424**	0.506	0.582	0.714	**0.466**	0.516	0.544	0.586
rkNN	0.505	0.549	0.587	0.607	**0.398**	0.504	0.566	0.706	**0.478**	0.530	0.548	0.624
SVMp	**0.398**	0.548	0.622	0.667	**0.350**	**0.465**	0.523	0.673	**0.408**	**0.492**	0.559	0.612
SVMr	**0.425**	0.548	0.598	0.655	**0.367**	**0.475**	0.556	0.679	**0.367**	**0.500**	0.531	0.618
MARS	**0.391**	**0.498**	0.555	0.608	**0.312**	**0.386**	**0.478**	0.583	**0.390**	0.510	0.541	0.578

Only these methods which performed better than random for particular threshold in all datasets were evaluated further. There were all algorithms for 10% threshold and just three approaches (Random Forest, Cubist and kNN) for 5% threshold. Average values of accuracy, F-measure and ROC_{AUC} obtained for those algorithms in cross-validation and for independent validation datasets are presented in Table 19 and Table 20, respectively. The highest scores are shown with bolded italics. The best performed algorithms were compared to the others and obtained *p*-values adjusted for multiple comparisons using Finner procedure. Bolded entries indicate the cases when algorithms do not differ significantly from the best one.

Table 19. ISA models performance on calibration datasets – change detection
(average values of performance measures)

Method	Raba datasets						Dunajec datasets						Sola datasets					
	5%			10%			5%			10%			5%			10%		
	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC
avNN				0.811	**0.315**	**0.695**				0.670	0.384	0.736				0.737	0.321	0.666/
RF	**0.717**	***0.360***	***0.659***	**0.834**	***0.349***	**0.708**	**0.556**	***0.352***	***0.667***	**0.736**	***0.447***	***0.788***	***0.663***	***0.356***	**0.645**	**0.795**	**0.393**	**0.718**
Cubist	***0.729***	**0.349**	**0.647**	0.825	**0.319**	**0.684**	**0.546**	**0.341**	**0.652**	0.707	**0.415**	**0.761**	**0.654**	**0.356**	***0.646***	**0.773**	**0.381**	**0.719**
GBM				0.824	**0.335**	**0.701**				0.723	**0.432**	**0.777**				**0.793**	**0.400**	***0.728***
kNN	**0.718**	**0.328**	**0.627**	0.823	**0.313**	**0.679**	***0.559***	**0.344**	**0.654**	0.708	**0.414**	**0.759**	**0.632**	**0.316**	0.602	0.754	**0.334**	**0.675**
rkNN				***0.844***	**0.346**	**0.697**				***0.757***	**0.409**	**0.757**				**0.777**	**0.368**	**0.700**
SVMp				0.798	**0.333**	***0.724***				0.698	**0.404**	**0.750**				***0.796***	***0.401***	**0.725**
SVMr				0.813	**0.331**	**0.709**				0.702	**0.410**	**0.757**				0.771	**0.361**	**0.697**
MARS				0.797	**0.313**	**0.698**				0.653	0.363	0.710				0.749	**0.344**	**0.690**

Acc – overall accuracy
F – F-measures
ROC – ROC_{AUC}

Table 20. ISA models performance on validation datasets – change detection
(average values of performance measures)

Method	Raba datasets						Dunajec datasets						Sola datasets					
	5%			10%			5%			10%			5%			10%		
	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC
avNN				0.791	**0.328**	**0.719**				0.658	**0.330**	**0.691**				0,710	0,315	0,685
RF	**0.713**	**0.378**	***0.663***	**0.843**	**0.381**	**0.725**	**0.571**	**0.330**	**0.677**	**0.730**	**0.395**	**0.747**	**0.649**	**0.378**	**0.693**	***0,814***	**0,455**	**0,791**
Cubist	**0.727**	**0.373**	**0.658**	**0.832**	**0.340**	**0.691**	**0.577**	**0.322**	**0.664**	**0.722**	**0.398**	**0.760**	**0.644**	**0.364**	**0.675**	0,750	**0,342**	0,699
GBM				***0.846***	***0.397***	**0.744**				***0.739***	***0.418***	**0.773**				0,774	**0,374**	**0,724**
kNN	***0.730***	***0.378***	**0.662**	**0.817**	**0.329**	**0.692**	***0.580***	***0.338***	***0.688***	**0.710**	**0.391**	**0.761**	***0.686***	***0.408***	***0.718***	**0,787**	**0,388**	**0,731**
rkNN				**0.836**	**0.345**	**0.692**				**0.733**	**0.415**	***0.777***				**0,787**	**0,397**	**0,742**
SVMp				0.782	**0.357**	***0.758***				**0.699**	**0.388**	**0.753**				***0,814***	***0,463***	***0,802***
SVMr				**0.830**	**0.388**	**0.747**				***0.739***	**0.397**	**0.748**				**0,806**	**0,436**	**0,775**
MARS				0.808	**0.363**	**0.748**				**0.687**	**0.345**	**0.700**				0,723	**0,341**	**0,715**

Acc – overall accuracy
F – F-measures
ROC – ROC_{AUC}

For change maps created based on 5% change threshold the three considered models performed very comparable. They gave no significant differences for validation dataset. However, in case of cross-validation the ROC_{AUC} score of kNN algorithm on Sola dataset differs significantly from the Cubist one.

For 10% change maps the image is different for different measures. Most algorithms show no significant difference from the best ones when F-measure and ROC_{AUC} scores are considered. Only avNN and MARS performed worse. But, in case of accuracy Random Forest and random kNN algorithms outperformed the others for both Raba and Dunajec datasets. These two algorithms proven their good performance on validation dataset as well.

3.2. Performance of model ensembles

3.2.1 Single time points

Table 21 presents the best model ensembles obtained for individual time point assessments together with average values of performance measures from cross-validation. Bolded values show cases when the performance of model ensemble is better than performance of any individual algorithm.

Table 21. Performance of model ensembles for individual time points ISA assessments (values averaged from cross-validation on calibration dataset)

Dataset	Mid 1990s				Late 2000s			
	Ensembled models	RMSE	MAE	R^2	Ensembled models	RMSE	MAE	R^2
Raba	GBM, CUB, kNN, SVMr, avNN	**0.1039**	**0.0597**	**0.7428**	CUB,RF,GBM, avNN, kNN, SVMp	**0.1094**	0.0626	**0.7759**
Dunajec	avNN, RF, GBM, kNN	**0.1544**	**0.1036**	**0.7208**	CUB, GBM, SVMr	**0.1279**	**0.0870**	**0.8419**
Sola	GBM, CUB, SVMr, kNN	**0.1185**	**0.0713**	**0.8387**	CUB, GBM, kNN, SVMp	**0.1167**	0.0710	**0.8618**

Performance of ensembled models was compared to the performances of individual algorithms. Table 22 and Table 23 present *p*-values obtained for the hypothesis about no difference in compared performances. For three datasets (Raba mid 1990s, Dunajec mid 1990s and Dunajec late 2000s) ensembled models outperformed significantly all individual algorithms when RMSE and R^2 are considered. In other cases no statistically significant differences from the best individual models were found.

Table 22. ISA models performance (RMSE, MAE) on calibration datasets (*p*-values for H0 = performance of the algorithm does not differ from the performance of ensembled model)

	Raba datasets				Dunajec datasets				Sola datasets			
Method	Mid 1990s		Late 2000s		Mid 1990s		Late 2000s		Mid 1990s		Late 2000s	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
avNN	<0.001	<0.001	0.001	<0.001	0.005	0.002	0.007	<0.001	<0.001	<0.001	<0.001	<0.001
RF	<0.001	<0.001	**0.081**	<0.001	0.007	<0.001	0.009	<0.001	<0.001	<0.001	0.002	<0.001
Cubist	0.001	**0.065**	0.010	**0.179**	0.001	0.021	<0.001	**0.558**	**0.078**	**0.357**	**0.445**	**0.500**
GBM	0.003	<0.001	0.030	0.014	0.008	0.021	0.013	0.009	**0.169**	**0.093**	0.043	**0.123**
kNN	0.001	**0.788**	<0.001	<0.001	<0.001	**0.133**	<0.001	<0.001	<0.001	0.047	<0.001	0.001
rkNN	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SVMp	<0.001	<0.001	0.001	0.023	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SVMr	0.009	0.043	0.003	0.031	<0.001	**0.178**	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
MARS	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 23. Model performances (R^2) on calibration datasets for individual time points ISA assessments (*p*-values for H0 = performance of the algorithm does not differ from the performance of ensembled model)

	Raba datasets		Dunajec datasets		Sola datasets	
Method	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s	Mid 1990s	Late 2000s
avNN	0.006	0.025	0.025	0.006	0.017	0.031
RF	0.007	**0.122**	0.035	0.006	0.018	0.031
Cubist	0.015	0.037	0.013	0.044	**0.094**	**0.395**
GBM	0.036	**0.058**	0.037	0.016	**0.170**	**0.075**
kNN	0.021	0.025	0.011	0.005	0.018	0.031
rkNN	0.015	0.025	0.004	0.005	0.017	0.031
SVMp	0.004	0.025	0.003	0.005	0.017	0.031
SVMr	0.037	0.028	0.011	0.012	0.017	0.031
MARS	0.004	0.025	0.003	0.005	0.013	0.031

Model ensembles were also used to predict imperviousness for validation datasets (Table 24). In this case no statistically significant differences from the best individual models were observed.

Table 24. Performance of model ensembles for individual time points ISA assessments on validation datasets

Dataset	Mid 1990s			Late 2000s		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Raba	**0.1081**	0.0630	**0.7513**	**0.1082**	0.0629	**0.8085**
Dunajec	**0.1471**	0.1007	**0.7354**	**0.1322**	**0.0880**	**0.8207**
Sola	0.1013	**0.0578**	0.8852	**0.1068**	0.0630	**0.8830**

Ensembled models were also included into comparison using the Friedman tests (Table 25). They are ranked first for all performance measures in cross-validation and for RMSE as well as R^2 calculated for independent validation. In case of MAE for validation datasets Cubist and GBM algorithms have higher average ranks. Comparison of the best algorithms to others in 1:N post-hoc analysis (Table 26) shows that only ensembled models and GBM do not differ significantly from the best ones in every case.

Table 25. Model (individual algorithms and ensembles) performances for individual time points ISA assessments – average ranks of Friedman tests

Method	Ranking (cross-validation)			Ranking (validation datasets)		
	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	7.17	8.33	7.5	6.83	7.83	7.33
RF	4.5	6.33	4.08	4.67	6.17	4.5
Cubist	2.58	2.33	2.58	4.83	**2.75**	4.83
GBM	2.91	3.50	3	4.25	4.00	4.17
kNN	6.25	4.17	6.83	7.5	5.33	7.42
rkNN	6.42	7.75	7.33	6.42	7.08	6.17
SVMp	8.17	7.33	7.5	6.25	4.50	6.08
SVMr	5.50	4.33	5.67	4.92	4.33	5
MARS	9.50	9.58	9.5	8.17	8.67	8.33
ens**emble**	**2.00**	**1.33**	**1**	**1.17**	4.33	**1.17**

Table 26. Adjusted *p*-values for 1xN comparisons of individual algorithms and model ensembles (Friedman post-hoc tests)

Method	Adjusted p-value (cross-validation)			Adjusted p-value (validation datasets)		
	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	0.009329	0.00028	0.000902	0.003559	0.016262	0.001572
RF	**0.191832**	0.007603	**0.09883**	**0.050767**	**0.110329**	**0.063368**
Cubist	**0.738597**	**0.567269**	**0.365047**	0.04751	**control algorithm**	0.045968
GBM	**0.643285**	**0.238572**	**0.279268**	**0.077748**	**0.474549**	**0.086119**
kNN	0.026915	**0.132974**	0.001523	0.001309	**0.236862**	0.001572
rkNN	0.025722	0.000725	0.000902	0.005997	0.039007	0.009495
SVMp	0.001884	0.001345	0.000902	0.006537	**0.435247**	0.009495
SVMr	**0.067109**	**0.126357**	0.011366	0.04751	**0.442324**	0.042163
MARS	0.00016	0.000021	0.00001	0.000559	0.006393	0.000372
ensemble	**control algorithm**	**control algorithm**	**control algorithm**	**control algorithm**	**0.442324**	**control algorithm**

3.2.2 Change intensity

Table 27 presents the average values of performance measures from cross-validation for model ensembles and assessments based on the best models according to RMSE or MAE scores. Again, bolded values show cases when the model outperformed the others.

Table 27. Performance of model ensembles on calibration datasets – change asseement

Dataset	Model ensembles			Best RMSE			Best MAE		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Raba	**0.0879**	**0.0498**	**0.3352**	0.0936	0.0537	0.2960	0.0962	0.0535	0.2802
Dunajec	**0.1275**	**0.0879**	**0.6274**	0.1362	0.0941	0.5955	0.1503	0.1036	0.5427
Sola	**0.1047**	**0.0646**	**0.3880**	0.1115	0.0706	0.3694	0.1124	0.0695	0.3630

Model ensembles performed better than any other approaches for RMSE in Raba dataset (Table 28). In other cases their performances were the best, but without significant differences to the performances of the best individual algorithms. Similar picture is visible for validation datasets (Table 29, Table 30).

Table 28. ISA models performance on calibration datasets – change asseement
(*p*-values for H0 = performance of the algorithm does not differ from the ensembled model)

Method	Raba dataset			Dunajec datasets			Sola datasets		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
avNN	<0.001	<0.001	0.0254	<0.001	<0.001	0.0115	<0.001	<0.001	0.0305
RF	0.038	0.001	**0.0707**	**0.686**	**0.521**	**0.2073**	**0.189**	**0.067**	**0.1586**
Cubist	<0.001	<0.001	0.0254	<0.001	<0.001	0.0151	<0.001	<0.001	**0.1177**
GBM	<0.001	<0.001	0.0264	0.002	0.008	0.0316	0.014	0.008	**0.1609**
kNN	<0.001	<0.001	0.0254	<0.001	<0.001	0.0087	<0.001	<0.001	0.0128
rkNN	0.037	**0.436**	0.0254	<0.001	<0.001	0.0087	<0.001	0.001	0.0305
SVMp	<0.001	<0.001	0.0254	<0.001	<0.001	0.0087	0.033	0.015	**0.1686**
SVMr	<0.001	<0.001	0.0302	<0.001	<0.001	0.0087	<0.001	<0.001	0.0305
MARS	<0.001	<0.001	0.0254	<0.001	<0.001	0.0087	<0.001	<0.001	0.0305
Best RMS	<0.001	<0.001	0.0254	<0.001	<0.001	0.0151	0.031	0.033	**0.5078**
Best MAE	<0.001	<0.001	0.0254	<0.001	<0.001	0.0087	<0.001	<0.001	**0.1177**

Table 29. Performance of model ensembles on validation datasets – change asseement

Dataset	Model ensembles			Best RMSE			Best MAE		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Raba	**0.0799**	**0.0470**	**0.3609**	0.0845	0.0486	0.3158	0.0921	0.0530	0.2650
Dunajec	**0.1272**	**0.0841**	**0.6106**	0.1329	0.0873	0.5767	0.1507	0.0958	0.5174
Sola	**0.1025**	**0.0614**	0.4436	0.1132	0.0678	0.3859	0.1176	0.0692	0.3604

Table 30. ISA models performance on validation datasets – change asseement (*p*-values for H0 = performance of the algorithm does not differ from the ensembled model)

Method	Raba dataset			Dunajec datasets			Sola datasets		
	squared error	absolute error	correlation coefficient	squared error	absolute error	correlation coefficient	squared error	absolute error	correlation coefficient
avNN	<0.001	<0.001	<0.001	**0.052**	0.001	**0.053***	<0.001	<0.001	<0.001
RF	0.019	0.017	0.004	**0.341**	**0.084**	0.040	**0.945**	**0.768**	0.007
Cubist	<0.001	0.009	<0.001	0.002	<0.001	<0.001	0.032	0.001	<0.001
GBM	0.014	0.017	<0.001	**0.341**	**0.197**	**0.100**	0.005	0.001	<0.001
kNN	<0.001	<0.001	<0.001	0.002	0.003	0.004	0.025	**0.065**	<0.001
rkNN	0.031	**0.301**	0.001	0.022	0.010	**0.053***	**0.713**	**0.717**	0.020
SVMp	<0.001	<0.001	**0.331**	**0.095**	0.047	0.005	**0.454**	**0.336**	**0.860**
SVMr	0.036	0.001	**0.570**	0.046	0.015	**0.053***	**0.066**	0.006	0.002
MARS	<0.001	<0.001	**0.419**	0.002	<0.001	<0.001	0.032	<0.001	<0.001
Best RMS	**0.052**	**0.258**	0.005	**0.233**	**0.261**	0.040	0.042	0.009	0.001
Best MAE	<0.001	<0.001	<0.001	0.002	0.001	<0.001	0.005	0.001	<0.001

** – differs significantly from the best model according to the method of Zou (2007)*

3.2.3 Estimation of ISA change detection capabilities

New models were also evaluated from ISA change detection point of view. First of all their performance was compared to the performance of random classifier (Table 31– 33). Like individual algorithms they do not fulfill the criterion of outperforming random classifiers accuracy for cross-validation runs in case of 1% and 3% thresholds change maps. For 5% threshold the accuracy of random guessing was beaten only by ensembled models. For change maps indicating at least 10% change all three models passed the tests.

Table 31. Ensembled models performance on calibration datasets – change detection (minimal values of overall accuracy measure). Scores below 0.5 are bolded

Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
ensemble	0.563	0.638	0.710	0.822	**0.250**	**0.383**	0.513	0.690	**0.425**	0.538	0.568	0.754
Best RMS	0.571	0.628	0.675	0.790	**0.246**	**0.362**	**0.486**	0.659	**0.404**	**0.477**	0.578	0.721
Best MAE	0.606	0.623	0.688	0.790	**0.206**	**0.340**	**0.460**	0.652	**0.430**	0.515	0.565	0.711

Table 32. Ensembled models performance on calibration datasets – change detection (comparison of precision with PRC). Bolded entries indicate performances below PRC

	Number of cases with precision below PRC treshold											
Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
ensemble	**3**	0	0	0	**44**	0	0	0	**8**	0	0	0
Best RMS	**1**	0	0	0	**45**	**1**	0	0	**17**	0	0	0
Best MAE	0	0	0	0	**48**	**2**	0	0	**3**	0	0	0

Table 33. Ensembled models performance on calibration datasets – change detection (minimal values of ROC_{AUC} measure). Scores below 0.5 are bolded

Method	Raba datasets				Dunajec datasets				Sola datasets			
	1%	3%	5%	10%	1%	3%	5%	10%	1%	3%	5%	10%
ensemble	**0.474**	0.562	0.621	0.635	**0.364**	0.518	0.578	0.722	**0.466**	0.557	0.583	0.638
Best RMS	**0.488**	0.568	0.587	0.624	**0.359**	**0.482**	0.552	0.700	**0.446**	0.514	0.570	0.641
Best MAE	0.514	0.533	0.557	0.617	**0.339**	**0.465**	0.557	0.680	**0.475**	0.536	0.566	0.621

Table 34 presents the average values of accuracy, F-measure and ROC_{AUC} from cross-validation. For 5% change maps model ensembles gave the highest values of considered measures (what is indicated by bolded italics). They also gave the highest values of accuracy and F-measure for 10% threshold change maps in Raba and Sola catchments. However, the differences to the best individual models in all these cases were not significant. Model ensembles do not differ significantly from the best individual algorithms in other cases neither. No significant differences to the best models were also observed for validation using independent datasets (Table 35)

Table 34. Ensembled models performance on calibration datasets – change detection
(average values of performance measures)

Method	Raba datasets						Dunajec datasets						Sola datasets					
	5%			10%			5%			10%			5%			10%		
	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC
ensemble	***0.741***	***0.371***	***0.666***	***0.849***	***0.368***	**0.715**	***0.570***	***0.355***	***0.670***	**0.740**	**0.446**	**0.783**	***0.672***	***0.363***	***0.651***	***0.800***	***0.402***	**0.723**
Best RMS				0.828	0.332	**0.696**				0.711	0.416	**0.760**				0.776	**0.383**	**0.719**
Best MAE				0.819	0.307	0.676				0.685	0.403	**0.753**				0.773	**0.381**	**0.719**

Acc – overall accuracy
F – F-measures
ROC – ROC_{AUC}

Table 35. Ensembled models performance on validation datasets – change detection
(average values of performance measures)

Method	Raba datasets						Dunajec datasets						Sola datasets					
	5%			10%			5%			10%			5%			10%		
	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC	Acc	F	ROC
ensemble	***0.751***	***0.408***	***0.682***	***0.851***	**0.377**	**0.710**	**0.577**	**0.321**	**0.662**	***0.745***	**0.410**	**0.762**	**0.678**	**0.390**	**0.698**	**0.800**	**0.421**	**0.761**
Best RMS				**0.837**	**0.347**	**0.693**				**0.739**	***0.420***	***0.781***				0.769	**0.386**	**0.743**
Best MAE				0.818	**0.319**	**0.687**				**0.719**	**0.407**	**0.773**				0.750	**0.342**	0.699

Acc – overall accuracy
F – F-measures
ROC – ROC_{AUC}

4. Discussion

The research presented in this paper may be divided into several parts. First of all we compared the performances of nine non-linear regression algorithms for sub-pixel imperviousness mapping from Landsat imagery in individual time points. The comparison was done across six datasets using RMSE, MAE and R^2 as the performance measures. The Cubist algorithm seems to outperform the other methods. When Friedman test was applied to the results from cross-validation procedure, it had the highest averaged ranks for all measures. It had also the highest averaged rank for MAE when the results obtained for independent validation datasets were considered. For RMSE and R^2 Cubist had the third best result. However, based on post-hoc analyses we cannot reject the hypotheses that two other algorithms (i.e. stochastic gradient boosting and support vector machines with radial kernel) performed equally well. When we compared the algorithms by pairs for individual datasets using cross-validation on calibration datasets, the Cubist was also the only one constantly present in the group of the best models (i.e. its performance was the best or not significantly differed from the best one), irrespectively of the dataset and the measure used for performance evaluation.

This finding is consistent with the previously reported in Drzewiecki (2016b). Cubist gave also better results than other machine learning algorithms in comparisons done by Walton (2008) and Mohapatra and Wu (2010). It should be noted however, that in eighteen comparisons done in actual study there is only one case (MAE for late 2000 Raba dataset) when GBM algorithm performance differed from the best one at significance level of 0.05. GBM performed also a little bit better then Cubist on independent validation datasets.

The next step in our research was to assess the performance of selected algorithms for evaluation of sub-pixel imperviousness changes. When comparing individual algorithms for estimation of change intensity, Random Forest seems to be better than others. For individual time point assessments it gave accuracies significantly lower than the best algorithms. It is true especially for MAE and visible in both Friedman test and by pair algorithm comparisons for cross-validation results. Despite this fact, RF is the only algorithm which is the best or not significantly differ from the best when change of imperviousness is estimated. This is also in accordance with findings of Drzewiecki (2016b). No other comparison of machine learning approaches for estimation of sub-pixel imperviousness change intensity has been found in literature.

Random Forest seems also to perform better than other considered algorithms when we assess the ability for correct detection of pixels where relevant change in impervious area coverage occurred. All researched approaches showed limited ability to reliably detect sub-pixel changes of imperviousness. To accept the method as a potential tool for detection of changes we wanted it to beat the performance of random classifier for every considered measure (accuracy, F-measure and ROC_{AUC}) in every cross-validation realisation. This is a very conservative condition, but as the number of datasets is quite limited we wanted to be as sure as possible that

the approach can be used safely. The overall accuracy turned out to be the most limiting factor. None method gave accuracies above 50% for change maps based on 1% and 3% change threshold in Dunajec catchment as well as 1% threshold in Sola chatchment.

For change maps created using the threshold of 5% change of imperviousness in Dunajec study area only three methods (RF, Cubist and k-NN) gave the accuracies better than 50% in every cross-validation realisation. In this task the performances of RF and Cubist were comparable, while kNN algorithm gave significantly worse ROC_{AUC} values for Sola dataset. When relevant change was defined as an increase (or decrease) of imperviousness of at least 10%, Random Forest and random kNN algorithms may be seen as the best ones. Most of remaining approaches gave the performances which did not differ significantly for F-measure and ROC_{AUC}. But RF and random kNN outperformed them significantly in overall accuracies.

The essential aim of the presented research was to compare the performance of single machine learning algorithms with the performance of their heterogeneous ensembles. When Friedman test was applied the ensembled models obtained the highest average ranks for all measures in cross-validation as well as for RMSE and R^2 in validations with independent datasets. For MAEs of validation datasets the ensembles were ranked third after Cubist and GBM models. Post-hoc tests showed that only model ensembles and generalised boosted regression trees gave always the best results or results not significantly different from the best ones. One should note however, that the Cubist and radial kernel SVM algorithms gave significantly different results for RMSE and R^2 obtained for independent dataset validations with *p*-values just a little below 0.05 threshold. Nevertheless, by pair comparisons using individual datasets showed in cross-validation case significantly better performance of model ensembles for RMSE and R^2 measures for three of six datasets.

Model ensembles gave also the lowest errors of imperviousness change intensity evaluation for both cross-validation and validation with independent datasets. In Raba catchment all other approaches performed significantly worse. When the ability of change detection is considered, the model ensembles performances did not differ significantly from the best of considered individual algorithms. One should note however, that they gave the highest values of all measures for 5% change threshold. For both change detection and change intensity evaluation, model ensembles outperformed also approaches based on the best models selected according to RMSE or MAE.

At least three issues should be discussed based on the achieved results. First of all, the research presented in this paper was designed to re-evaluate the results of sub-pixel imperviousness and imperviousness change intensity assessment reported in Drzewiecki (2016b) using the approaches suggested as more appropriate in machine learning literature. In general the outcomes obtained are in accordance with the previous ones. When individual algorithms are considered the Cubist still may be recommended as the most appropriate for mapping imperviousness in individual time points and Random Forest as the better choice for ISA change assessments. However,

some differences exists as well. The corrected *t*-test for repeated cross-validation applied in present study resulted in higher number of algorithms which cannot be considered as significantly worse from the best ones in by pair comparisons for individual datasets. It does not change the final conclusions substantially. However, in actual study the difference between the performances of Cubist and GBM algorithms for individual time point assessments in cross-validation procedure is much smaller than in previous evaluation. Both algorithms performed also the best for independent validation datasets. Moreover, in this case GBM was a little bit better. Based on these results, when application of single algorithm is considered, it can be recommended as the alternative approach to ISA evaluation for single time-point.

Secondly, in presented study two scenarios were used to compare machine learning approaches for individual time points ISA evaluations. Algorithms were assessed based on multiple datasets using Friedman tests and on individual datasets using by pair comparisons to the best performed algorithm. The latter approach seems to be more powerful, although it was more complicated and time-consuming as well. However, by using pair comparisons on individual datasets, we were able to find more differences in performances than with Friedman test and post-hoc analyses. The reason may be searched in low number of datasets available for comparison. As pointed out by Trawiński et al. (2012) the power and efficiency of nonparametric test of significance is low with small sample sizes. In our case we had only six datasets available to compare.

The last but very important issue is related to the applicability of researched approach for change detection. We tried to detect sub-pixel changes of imperviousness based on comparison of two sub-pixel ISA estimations done for individual points in time. The research reveal the fact that such methodology is hardly applicable if we want to find subtle ISA changes. None of considered approaches was able to assure satisfactory accuracy of change detection for relevant change threshold defined as 1 or 3 percents. For 5% threshold the most approaches failed as well. Despite of very conservative rules applied in presented study (we wanted the results better than the result of random classifier for all cross-validation runs), this shows some limitations of the researched approaches. As showed in study of Bernat and Drzewiecki (2014), detection of completely pervious pixels in additional step of hard (binary) classification may result in more reliable determination of such areas and reduced classification noise, i.e. the lower number of (usually small) errors for completely perviouse pixels. We may expect that it also should improve the ability for detection of ISA changes, especially when the more subtle changes are being searched. We plan to check this in future studies.

5. Conclusions

In this paper we presented the results of thorough statistical comparison of nine machine learning methods (Cubist, Random Forest, stochastic gradient boosting of

regression trees, k-nearest neighbors regression, random k-nearest neighbors regression, Multivariate Adaptive Regression Splines, averaged neural networks, and support vector machines with polynomial and radial kernels) for sub-pixel imperviousness and imperviousness change assessment. The study is the continuation and extension of the research presented in Drzewiecki (2016b). In present study new methodology was applied for comparisons and, additionally, the applicability of selected approaches for detection of relevant ISA changes was taken into consideration. When performances of single algorithms are considered, the Cubist and GBM approaches outperformed the other techniques for imperviousness evaluation in single time steps. However, when the goal is to assess imperviousness change Random Forest would be better choice. Despite lower accuracies for individual time point predictions, it allowed for both more accurate estimation of change intensities as well as more reliable mapping of relevant change occurrences.

The study proven also that heterogeneous ensembles of non-linear regression models allow to obtain the results which are better than or at least as good as the best of the ones obtained with individual models. This is true for individual time points imperviousness assessment as well as for ISA change intensity evaluation and detection of relevant imperviousness changes. It is worth noting, that to construct such ensembles we do not need reference data for ISA change (Drzewiecki, 2016b). As the models are ensembled based on their performances in individual time points, the reference information about imperviousness in single dates is enough. This is very important as in many cases one faces lack of the reference areas for ISA change evaluation (Yang et al., 2003; Dams et al. 2013) which makes impossible to choose the most accurate individual models. Using the approach based on model ensembles one can possibly improve the sub-pixel imperviousness change assessment even without reference change information.

The study also revealed that the methodology of imperviousness change detection based on differences of sub-pixel evaluations done for individual time point has limited ability of reliable detection of subtle ISA changes. In our case we may rely on the change detection results regardless the algorithm used for ISA mapping only for the highest of considered thresholds, i.e. when relevant changes were defined as over 10% increase or decrase of ISA. For 5% threshold model ensembles are preferable as for the most of approaches some risk of too low accurate results occurred.

Acknowledgments

Research founded by AGH University grant No. 11.11.150.949. Datasets used in this research were prepared by Ms. Katarzyna Bernat under Author's supervision for the needs of SaLMaR project co-financed by the Ministry of Science and Higher Education of Poland and the German Federal Ministry for Education and Research.

References

Aleksandrowicz, S., Wawrzaszek, A., Drzewiecki, W. and Krupiński, M. (2016). Change Detection Using Global and Local Multifractal Description. *IEEE Geoscience and Remote Sensing Letters*, 13, 8, 1183–1187. DOI: 10.1109/LGRS.2016.2574940

Amancio, D.R., Comin, C.H., Casanova, D., Travieso, G., Bruno, O.M., Rodrigues, F.A. and da Fountoura Costa, L. (2014). A Systematic Comparison of Supervised Classifiers. *PLoS ONE* 9(4): e94137. DOI:10.1371/journal.pone.0094137

Bernat, K. and Drzewiecki, W. (2014). Two-stage subpixel impervious surface coverage estimation: comparing classification and regression trees and artificial neural networks. In: Proc. SPIE Vol. 9244, 92441l, *Image and Signal Processing for Remote Sensing*. DOI: 10.1117/12.2067308

Bouckaert, R.R. (2003). Choosing Learning Algorithms Using Sign Tests with High Replicability. In: Gedeon T..D., Fung L.C.C. (eds.) AI 2003: Advances in Artificial Intelligence. AI 2003. *Lecture Notes in Computer Science*, vol 2903. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-24581-0_61

Bouckaert, R.R. (2004). Estimating replicability of classifier learning experiments. In Proceedings of the 21st International Conference on Machine Learning Banff, Canada, 2004. DOI: 10.1145/1015330.1015338

Bouckaert, R. R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In: D. Honghua, R. Srikant, and C. Zhang, (eds.), Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26–28, 2004, Proceedings. Springer, 2004. DOI: 10.1007/b97861

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. DOI: 10.1023/A:1010933404324

Coelho, G.P. and Von Zuben, F.J. (2006). The influence of the pool of candidates on the performance of selection and combination techniques in ensembles. In: Proceedings of the International Joint Conference on Neural Networks, 10588–10595.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. DOI: 10.1177/001316446002000104

Daniel, W.W. (1990). Applied nonparametric statistics. Duxbury Thomson Learning, Pacific Grove.

Dams, J., Dujardin, J., Reggers, R., Bashir, I., Canters, F. and Batelaan, O. (2013). Mapping impervious surface change from remote sensing for hydrological modeling. *Journal of Hydrology*, 485, 84–95. DOI: 10.1016/j.jhydrol.2012.09.045

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.

Diedenhofen, B. and Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE* 10(4): e0121945. DOI: 10.1371/journal.pone.0121945

Drzewiecki, W. (2016). Comparison of Selected Machine Learning Algorithms for Sub-Pixel Imperviousness Change Assessment. In: 2016 Baltic Geodetic Congress (Geomatics), 67–72. DOI: 10.1109/BGC.Geomatics.2016.21

Drzewiecki, W. (2016). Improving sub-pixel imperviousness change prediction by ensembling heterogeneous non-linear regression models. *Geodesy and Cartography*, 65, 2, 193–218. DOI: 10.1515/geocart-2016-0016

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 293, 52–64.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 8, 861–874. DOI: 10.1016/j.patrec.2005.10.010

Finner, H. (1993). On a monotonicity problem in step-down multiple test procedures. *Journal of the American Statistical Association*, 88, 920–923. DOI: 10.1080/01621459.1993.10476358

Fisher, R.A. (1937). *Statistical methods and scientific inference*. Hafner publishing Co, New York.

Foody, G.M. (2009). Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113, 8, 1658–1663. DOI:10.1016/j.rse.2009.03.014

Friedman, J. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1–141.

Friedman, J. (2002). Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*, 38(4), 367–378. DOI: 10.1016/S0167-9473(01)00065-2

Friedman, M. (1940). A Comparison of Alternative Tests of Significance for the Problem of *m* Rankings. *The Annals of Mathemathical* Statistics, 11, 1, 86–92.

Garcia, S., Fernandez, A., Luengo, J. and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Information Sciences*, 180 (10), 2044–2064. DOI: 10.1016/j.ins.2009.12.010

Hanley, J.A. and McNeil, B.J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143, 1, 29–36. DOI: 10.1148/radiology.143.1.7063747

Heremans, S. and Van Orshoven, J. (2015). Machine learning methods for sub-pixel land-cover classification in the spatially heterogeneous region of Flanders (Belgium): a multi-criteria comparison. *International Journal of Remote Sensing*, 36, 11, 2934–2962. DOI: 10.1080/01431161.2015.1054047

Hussain, M., Chen, D., Cheng, A., Wei, H. and Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80, 91–106. DOI: 10.1016/j.isprsjprs.2013.03.006

Iman, R.L., Davenport, J.M. (1980) Approximations of the critical region of the friedman statistic. *Communications in Statistics*, 9, 571–595. DOI: 10.1080/03610928008827904

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms*. A Classification Perspective. Cambridge university Press.

Joshi, M.V. (2002). On evaluating performance of classifiers for rare classes. In Proceedings of The 2002 IEEE International Conference on Data Mining, pp. 641–644. DOI: 10.1109/ICDM.2002.1184018

Kircher, J. (2001). Data Analysis Toolkit #5: Uncertainty Analysis and Error Propagation. University of California Berkeley Seismological Laboratory. Available online at: http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_05.pdf

Klaric, M. (2014). Predicting Relevant Change in High Resolution Satellite Imagery. *ISPRS International Journal of Geoinformation*, 3, 1491–1511. DOI: 10.3390/ijgi3041491

Landis, J.R. and Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 30, 1, 159–174. DOI: 10.2307/2529310

Li, S., Harner, E.J. and Adjeroh, D.A., (2011): Random KNN feature selection – a fast and stable alternative to Random Forests. *BMC Bioinformatics*, 12:450. DOI: 10.1186/1471-2105-12-450

Lu, D., Li, G., Kuang, W. and Moran, E. (2014). Methods to extract impervious surface areas from satellite images. *International Journal of Digital Earth*, 7, 2, 93–112. DOI: 10.1080/17538947.2013.866173

Lu, D., Li, G., Kuang, W. and Moran, E. (2014). Current situation and needs of change detection techiques. *International journal of Image and Data Fusion*, 5, 1, 13–38. DOI: 10.1080/19479832.2013.868372

Morgan, M.G. and Henrion, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis.* Cambridge University Press.

Nadeau, C. and Bengio, Y. (2003). Inference for the Generalization Error. *Machine Learning,* 52, 3, 239–281. DOI:10.1023/A:1024068626366

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42–57, DOI: 10.1016/j.rse.2014.02.015

Powers, D.M.W. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2, 1, 37–63. DOI: 10.9735/2229-3981

Quinlan, R. (1993). Combining instance–based and model–based learning. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 236–243.

Ridd, M.K. (1995). Exploring a V-I-S (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: Comparative anatomy for cities. *Int. J. Remote Sens.*, 16, 2165–2185. DOI: 10.1080/01431169508954549

Ripley, B. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 10(3): e0118432. DOI:10.1371/journal.pone.0118432

Santafe, G., Inza, I. and Lozano, J.A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44, 467–508. DOI: 10.1007/s10462-015-9433-y

Shahtahmassebi, A.R., Song, J., Zheng, Q., Blackburn, G.A., Wang, K., Huang, L.Y., Pan, Y., Moore, N., Shahtahmassebi, G., Haghighi, R.S. and Deng, J.S. (2016). Remote sensing of impervious surface growth: A framework for quantifying urban expansion and re-densification mechanisms. *International Journal of Applied Earth Observation and Geoinformation*, 46, 94–112. DOI: 10.1016/j.jag.2015.11.007

Smola, A.J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88

Steiger, J.H. (1980). Tests for Comparing Elements of a Correlation Matrix. Psychological Bulletin*,* 87: 245–251. DOI: 10.1037/0033-2909.87.2.245

Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A. and Fisher, P.F. (2015). A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*, 160, 1–14. DOI: 10.1016/j.rse.2015.01.006

Trawiński, B., Smętek, M., Telec, Z. and Lasota, T. (2012). Nonparametric Statistical Analysis for Multiple Comparison of Machine Learning Regression Algorithms. *International Journal of Applied Mathematics and Computer Science*, 22, 4, 867–881. DOI: 10.2478/v10006-012-0064-z

Turner, II B.L. and Meyer, W.B. (1994). Global Land Use and Land Cover Change: An Overview. In: Meyer W.B. and Turner II B.L. (eds.), Changes in Land Use and Land Cover: A Global Perspective. Cambridge University Press, pp. 3–10.

Wężyk, P., Hawryło P., Szostak, M., Pierzchalski, M. and de Kok, R. (2016). Using Geobia and Data Fusion Approach for Land use and Land Cover Mapping. *Quaestiones Geographicae*, 35, 1, 93–104. DOI: 10.1515/quageo-2016-0009

Wieland, M., Liu, W. and Yamazaki, F. (2016). Learning Change from Synthetic Aperture Radar Images: Performance Evaluation of a Support Vectore Machine to Detect Earthquake and Tsunami-Induced Changes. *Remote Sensing*, 8 (10), 792, DOI: 10.3390/rs8100792

Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics,* 1(6), 80–83. DOI: 10.2307/3001968

Yang, L., Xian, G., Klaver, J. M. and Deal, B. (2003). Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 69, 9, 1003–1010. DOI: 10.14358/PERS.69.9.1003

Yang, Y. and Liu X. (1999). A re-examination of text categorization methods. *In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49. DOI: 10.1145/312624.312647

Zou, G. Y., (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12, 4, 399–413. DOI: 10.1037/1082-989X.12.4.399