On the validation of spatial statistical vegetation models

Igor Makhatkov*

Institute of Soil Science and Agrochemistry SB RAS, Academician Lavrentiev Ave., 8, building 2, Novosibirsk, 630090, Russia

Abstract. The features of coefficients of determination and coefficients of leave-one-out method for spatial vegetation model and spatial models of squared deviations are discussed. The properties of models are illustrated in key area for spatial model of Cladonia stellaris projective cover.

Spatial modeling of the vegetation and its different features can be realised as spatial extrapolation of classification units of vegetation [1] or spatial extrapolation of unclassified variables [2]. The last way makes opportunity to create a spatial model of unclassified information [3]. In both cases, spatially distributed variables e.g. the spectral responses of the surface are used as predictors of extrapolation. The quality of modeling is estimated by the coefficient of determination (R^2, r-squared), which shows the ratio of variation (the sum of the squared deviations) explained by the adopted model [4]. This estimation of the statistical dependence of modified variables and predictors can be increased unlimitedly by increasing the flexibility of the regression functions and by including more predictors. The validity of such a model improvement is evaluated by validation methods, such as the coefficient of determination based on the leave-one-out method, the successive exclusion of each sample from the training data and the use of its deviation from the predicted value [5]. Depending on the completeness of the data and the features of the regression model, its deviations from the observed values in different areas of the model can vary significantly. Spatial modeling of these deviations can provide a detailed assessment of the quality of the model and the quality of area survey. To investigate deviations of spatial models we used the key area with geobotanical relevés. Spatial models of the squared deviations of the vegetation model and its individual variables from the observed data, and the squared deviations of the leave-one-out method are considered.

25 geobotanical relevés of the key area (N 63.379, E 75.865 / N 63.105, E 76.451), Yamal-Nenets Autonomous Region were used as observed data. 88 rare species were excluded. The key area is within the subzone of the northern taiga [6]. Here the good drained loamy habitats are occupied by zonal larch shrub-moss forests, drained sands are occupied by intrazonal pine shrub-lichen forests. Flat watersheds are occupied by frozen and thawed mires. Oligomesotrophic bogs occupy hydrological net of swamps. Floodplain series of vegetation are common near rivers and streams. Forests and frozen mires on the key area have been partly burned by recent fires. The spatial modeling of species abundance was based on the variables of the spectral responses of the composite of 6

* Corresponding author: makhatkov@mail.ru

channels of 3 summer Landsat images with a resolution of 30 m/pixel. For regression analysis, polynomial functions were used. For data handling we used QGis 2.12 (http://www.qgis.org/) with additional modules, for calculations - Python 2.7 (https://www.python.org/) with the necessary libraries (http: // www. numpy.org/, http://scikit-learn.org/, http://www.gdal.org/), and the editor of PyCharm Community (https://www.jetbrains.com).

The vegetation model of the key area was build using principal components values of the geobotanical relevés matrix with values of species projective cover (%) as factors [7]. According to a preliminary estimation of the principal components significance the first three components were significant only. The coefficients of determination were calculated for the vegetation model in total and separately for 3 factors (main components). In addition, as an example, the coefficients for the projective cover models of two species of Cladonia stellaris (Opiz) Pouzar et Vězda and Rubus chamaemorus L werelculated (table).

Table. Coefficients of determination

	Model in total	PCA components			Cladonia stellaris	Rubus chamaemorus
		1	2	3		
Model of vegetation cover	0.66	0.63	0.82	0.53	0.70	0.35
Leave-one-out method for model of vegetation cover	0.52	0.52	0.72	0.32	0.63	0.13
Model of squared deviations	0.26	0.28	0.17	0.30	0.16	0.30
Leave-one-out method for model of squared deviations	0.40	0.38	0.43	0.43	0.19	0.54

The coefficients of determination of the predicted and observed values were relatively low. This indicates a weak statistical dependence thematic values and predictors, in our case with species abundance and imagine values. Nevertheless some regularities of this dependences should be noted. The models of deviations in all cases were than better than models of vegetation worse. In other words the larger the ratios of deviations that the accepted model does not explain, the more closely these deviations are connected with predictors. This property of deviations is probably associated with differences in the regression functions accepted in the main model building and in model of deviations, and can be used to improving the models.

The greatest deviations of the leave-one-out method are associated with local lack in the survey data when the exclusion of some observations strongly changes the model. The coefficients of determination of leave-one-out method are always better than coefficients of models, and the statistical relationship of deviations and predictors does not always depend on the quality of the model. The values of squared deviations of leave-one-out method can be used as an assessment of the surveying of the territory and pay attention to areas with relatively large values of squared deviations to supplement the survey data.

The properties of deviation models are well illustrated by their fragments for Cladonia stellaris for key area (Fig., a). The projective cover model (Fig., b) shows the highest values for pine shrub-lichen forests that are common on sandy habitats along some rivers and low values for thawed mires, the lowest for larch forests, and complete absence on floodplains. The model of deviations (Fig., c) shows the highest quality of abundance model for most of the larch forests and frozen mires, slightly less for pine forests, and the worst for ones impacted by fires and partly for larch forests. The model of squared deviations of leave-one-out method (Fig., d) is generally similar to the previous one. But it shows large differences on fragments of frozen bogs that were affected by fire. This indicates a lack of survey data in similar areas.

In general despite the relatively weak dependence between squared deviations of model, squared deviations of leave-one-out method and predictors, spatial models their values can be used for a detailed assessment of the quality of the main model and its improvement.

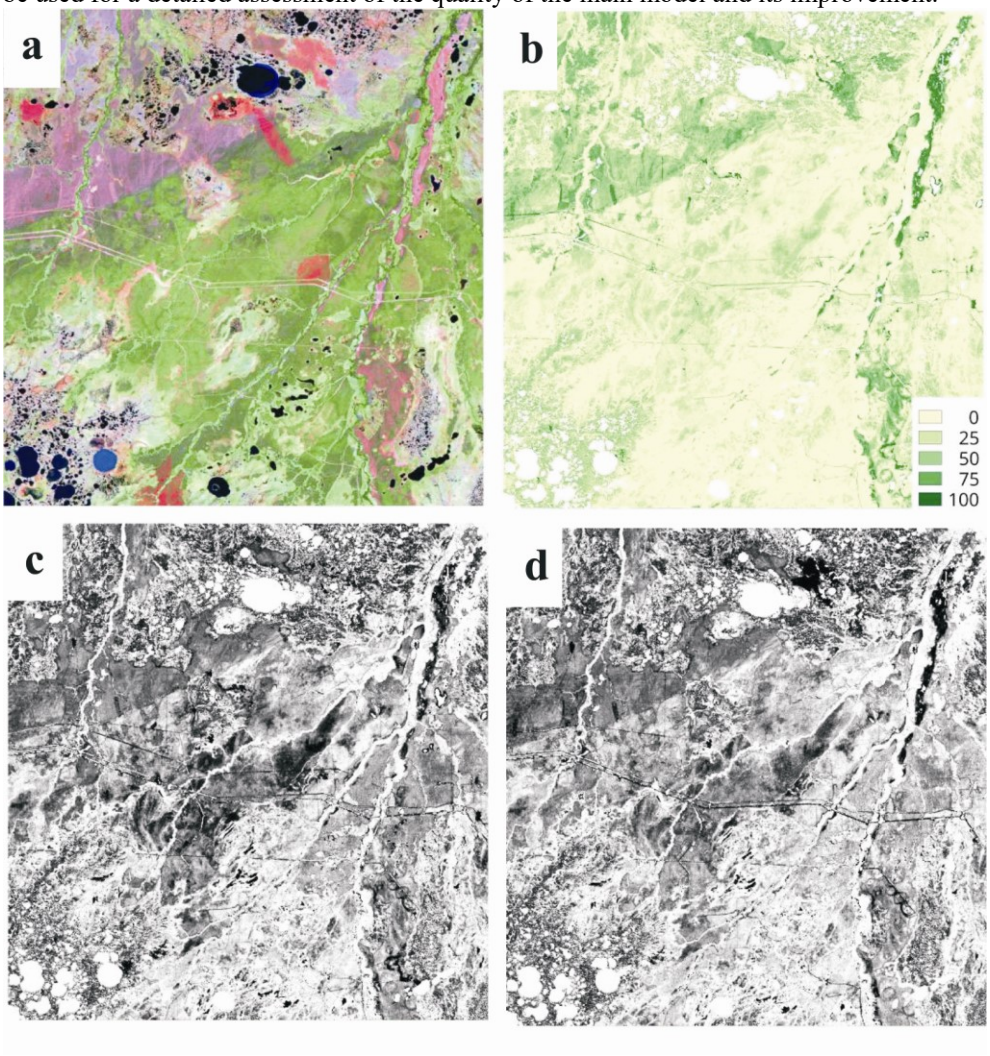

Fig. a-d. The spatial model of the projective cover of Cladonia stellaris and the assessment of its reliability (a - image of a fragment of the key site, 7 4 2 Landsat bands, b - model of projective cover (%), c - model of squared deviations, d - model of squared deviations of leave-one-out method).

References

1. S. Bartalev, V. Egorov., V. Zharko, E. Lupyan, D. Plotnikov, S. Khvostikov, Modern problems of remote sensing of the Earth from space, **12**, 5 (2015)
2. M. Al-Hamdan., J. Cruise, D. Rickman, D. Quattrochi, Remote Sensing, **6 (10)** (2014)
3. J. Franklin, *Mapping Species Distributions: Spatial Inference and Prediction* (Cambridge University Press, Cambridge, UK, 2009)

4. G. Ivchenko, Yu. Medvedev. *Introduction to mathematical statistics* (M., Publishing House of LCI, 2010)
5. B. Efron. *Non-traditional methods of multivariate statistical analysis*. *Finance and Statistics* (M., 1988)
6. I.S. Ilyina, E. I. Lapshina, N. N. Lavrenko, L. I. Mel'tser, E. A. Romanova, B. A. Bogoyavlenskii, and V. D. Makhno, *Plant cover of the West Siberian Plain* (Nauka, Siberian Branch, Novosibirsk, 1985)
7. J. Kim, C. Muller. *Factor Analysis: Statistical Methods and Practical Issues. Factor, discriminant and cluster analysis* (M., Finance and Statistics, 1989)