Bayesian Classification of Digital Images by Web Application

Milan TALICH, Ondřej BÖHM and Lubomír SOUKUP, Czech Republic

Key words: key Bayesian classification, digital images, web application, web map services

SUMMARY

The contribution introduces web application for image classification that has been developed at the Research Institute of Geodesy, Topography and Cartography in the framework of grant project InGeoCalc (supported by Ministry of education of the Czech Republic). The web application is aimed to display, examine and classify digital image data. The data are expected to be obtained from Internet by means of Web Map Services (WMS) or from other sources (possibly non-registered). Image data from different sources can be combined and presented as composition of layers (coverage) with adjustable degrees of transparency.
After gathering the data, Bayesian (supervised) classification is applied to distinguish separate regions in the image. User can choose between several classification methods and adjust pertinent parameters. Furthermore, several subsequent basic analytical tools are offered, namely computation of distances, areas or perimeters related to the classified regions, simple statistical summaries about classification results (e.g. distribution of classes, percentage of non-classified regions, etc.). The classification results and registration parameters can be saved for further use.
The web application is based on common Internet standards (HTML, Javascript, SVG). The only requirement for running the application is an up-to-date Internet browser supporting SVG (Scalable Vector Graphics). Typical usage of the web application can involve land cover mapping based on satellite or aerial images. The application is available free of charge for any Internet user.

Bayesian Classification of Digital Images by Web Application

Milan TALICH, Ondřej BÖHM and Lubomír SOUKUP, Czech Republic

1. INTRODUCTION

In the last decade, digital images have been heavily spread in common life. This trend is caused by immense availability of digital cameras and scanners of different kinds as well as new ways of providing and sharing images by means of up-to-date telecommunication channels, namely Internet based ones. Utilisation of such images has been rapidly growing in vast variety of human activities.

In case of geospatial information, there is significant trend to provide raster spatial data, namely by web map services. Furthermore, additional applications and services have been required to process analytical operations over the given spatial data. Several motivation aspects of this trend can be identified, but decision support processes seem to be the main ones. For such decision processes, collecting correct reference spatial data plays a crutial role. The spatial data are usually required in form of specific regions that satisfies given criteria, e.g. flood areas, areas impacted by human activity or natural disaster, land use areas or another kinds of areas of interest. Common analytical GIS tools for identifying areas of interest in raster images are usually time demanding since they include significant portion of manual interaction. Automation of such operation is therefore required and some classification technique could be applied. Reliability of classification highly depends on the method used, especially on the level of application of mathematical statistics. From this point of view, Bayesian classification seems to be most promising.

1.1 Problem formulation

Regions with some characteristic features have to be localized in a digital image. These features could be uniquely derived from the given attributes of pixels, e.g. color of a pixel. The whole image has to be decomposed into disjoint regions and each region has to be attributed by a unique class according to the prevalent characteristic features. The set of classes has to be given in advance. The decomposition task, i.e. classification, results in assignment a specified class to each pixel in the given digital image.

1.2 The required result

Result of classification has to be in form of a new image that consists of homogenous disjoint regions of different classes. The regions are distinguishable by class labels or colors that are explained in the associated legend.

2 CLASSIFICATION OF RASTER IMAGES

2.1 Review of the main classification methods

Vast number of different classification methods have been designed during short history of development of computer image processing. Two main groups of classification methods can be recognized: deterministic and statistic. Other distinction between classification methods is based on practical circumstances of solution of the classification problem. When some characteristic features of the classes are available, the classification is called supervised. If no preliminary data about classes are known in advance, unsupervised classification (cluster analysis) has to be performed. Statistical supervised classification (see e.g. [2], [1]) presents more powerful tool than the other kinds of classification.

Statistical characteristics of the all admissible classes have to be known at statistical supervised classification. The most common way of gaining characteristic features of classes is to determine training sets in the given image. Training set is a region in the image which well represents certain class. Searching for other regions with similar characteristic features is the task of supervised classification. Principle of statistical supervised classification is based on geometric notion of feature space. Feature space is an Euclidean space of points, whose coordinates are features that characterize each pixel in the image. Typical example of feature space is color space RGB. Each pixel of digital image displays as a point with coordinates that are the features, e.g. color components Red, Green, Blue. Points in feature space create clusters that represent particular classes. Some points of these clusters correspond to pixels that belong to a training set. These points can be labeled by identifier of a corresponding class. With the aid of the labeled points of training sets, other points in feature space have to be labeled to complete the classification. Hence the classification task can be formulated as a rule for labeling pixels displayed in feature space. This rule, called classifier, can be searched for by means of several manners. The most common classifiers are e.g. linear, nearest neighbour or Bayesian classifiers. Bayesian classification that is the most important member of the family of statistical supervised classification will be studied in the sequel.

2.2 Bayesian classification

2.2.1 Input data and assumptions

A digital image is given where training sets are determined. Certain number of classes is chosen to distinguish regions of different characteristics. Let C be the set of the all classes. Each training set is assigned to a certain class. Each class has to be represented by one training set at least. Furthermore a prior probability P(C) has to be known for each class C in C. The prior probabilities describe general preliminary information about presence of classes in the given image.

2.2.2 Solution of the problem

The classification problem is solved by Bayesian classifier in this contribution. The Bayesian classifier stems from Bayes formula (see e.g. [2]). This formula enables to compute the probability that a pixel with feature vector F belongs to class C. It is conditional probability P(C | F). We can estimate opposite conditional probability P(F | C) for any feature vector F and class C with the aid of the training sets. Expression P(F | C) stands for probability that a pixel of class C has feature vector F. Under these assumptions for known prior probabilities P(C) the Bayes formula has form:

$$P(C|F) = \frac{P(F|C)\,P(C)}{\sum_{T \in \mathbf{c}} P(F|T)P(T)} \quad (1)$$

The last step of the classification procedure comprises assignment of class C to pixel with feature vector F to maximize posterior probability P(C | F).

Crucial problem resides in computation of probabilities P(F | C), since it is sensitive to input data in training sets. Three variants of Bayesian classification will be presented to cover most cases of determining training sets.

2.2.3 Basic variant

The simplest way of computation probabilities P(F | C) is based on relative frequency of pixels in the training set. Let us denote n_C the overall number of pixels in training set of class C and $n_{C,F}$ the number of pixels with feature vector F in the same training set. Then the probability P(F | C) can be approximately estimated by

$$P(F|C) = \frac{n_{C,F}}{n_C} \quad (2)$$

2.2.4 Extended variant

The extended variant is based on assumption, that clusters of the same class are normally distributed. Under this assumption each training set could be extended by adding other pixels with similar features as the original pixels selected by the actual training set in the chosen cluster.

Pixels, whose feature vectors are sufficiently close to the center of the cluster, could be treated as members of the actual class C. Such pixels can extend the actual training set to create a new, extended training set. The extended training set is more representative, but there is some risk, that some of its pixels do not belong to the actual class C. If the risk is small (e.g. less then 0.05), it is possible to compute relative frequency (2) with greater numbers n_C , $n_{C,F}$. Better estimation of probabilities P(F | C) could be reached by this way. The problem is in definitions of riskiness and sufficient closeness to the center of cluster.
The distance of additional pixels from the center of cluster is measured by Mahalanobis

distance. The limit distance below which the pixels are considered close has to be determined in accordance to the risk of appending wrong pixels.

2.2.5 Nearest neighbour variant

This variant is based on assumption of normality of clusters as in the previous variant. Indeed, membership of a pixel into a class is computed as a distance of the pixel from the center of the cluster. The pixel is assigned to the class, whose training set is the nearest to the pixel in question. The metrics for measuring the distance is derived from Mahalanobis distance. The distance between a pixel and a training set of class C is a posterior probability P(C | E_h) which is given by Bayes formula in the consequent form.

$$P(C|E_h) = \frac{P(E_h|C)\,P(C)}{\sum_{Tec} P(E_h|T)\,P(T)} \qquad (3)$$

Symbol E_h depends on the risk of appending wrong pixels.

3 PRACTICAL ON-LINE SOLUTION

Web application for practical solution of Bayesian classification was created. The application, named WACLASS, works as a client - server application. It is available at *http://www.vugtk.cz/ingeocalc/igc/classification/*.

The client part of the application supports all the user operations, namely design of classes, definition of training sets and so on. The actual classification runs on the server side of the application. This part of the application was programmed in Python language with the aid of web framework Django and image processing library PIL (Python Image Library). Client side of the application is based on standard up-to-date web technologies such as HTML, Javascript, and SVG (Scalable Vector Graphics). It means that the application can be used on practically any computer that is connected to Internet with any web browser. The only exception is Internet Explorer of older version than 9, since it does not support SVG. Communication client - server is asynchronous.

3.1 Features of the application

The application offers all the necessary tools for supervised classification of digital images. It allows to create classes, display image data, and define training sets.

3.1.1 Classes and training sets

Definition of classes consists in entering necessary information about each class: name, colour, and a prior probability. The colour will be used in the final result of classification. The colour as well as prior probability can be modified during the classification process.
Training sets are formed graphically as rectangles. User points out two opposite corners of the

rectangle for an actual training set. The user can delete selected classes as well as training sets if necessary.

3.1.2 Data sources

Both image files or WMS data (i.e data provided by Web Map Service) can serve as data sources. The both kinds of data sources can be combined. In the case of image files, georeference information can be supplied.

The web application can simultaneously display several data sources. Any data source as well as any result of classification is displayed as a separate layer. Overlays of multiple layers can be created simply by changing transparency of the layers.

The image data, including the georeference information as a world file (see [4]), can be saved to local computer..

3.1.3 Variants of classification

The classification is executed by selecting the desired variant in application menu. Three variants are available: basic variant, extended variant, and nearest neighbour variant. All the three variants use the RGB color space as feature space.

Basic variant
This variant uses relative frequencies of pixels in training sets (see paragraph 2.2.3). Advantage of this variant is simplicity and speed, but the results are not too good. Many pixels remain unclassified.

Extended variant
This variant is based on extension of training sets. Selected additional pixels from a cluster are appended to the original training set. There is some risk of appending wrong pixels. Principle of this variant is described in paragraph 2.2.4.

Nearest neighbour variant
This variant uses clusters in colour space as the previous one. The classification criterion is distance from the center of a cluster. Principle of this variant is described in paragraph 2.2.5.

Results of the classification are displayed as another layer. It is possible to save it as an image (with georeference information if it is available).

3.1.4 Analytical tools

The application provides the user with some analytical tools. First, statistics of the classification lists data on the number of pixels in separate classes including unclassifiied pixels. Furthermore, brief overview of classified areas is presented (these areas are meaningful only for georeferenced data as they are computed from spatial resolution of the

image).

Additionally, some interactive tools are available: measuring of distances, perimeters and areas selected by the user on the screen.

3.2 Application controls

The application user experience strives to imitate desktop aplications. The browser's display area is split between application menu, tab bar and current tab's content. Most of the functions are executed from the application menu.

3.2.1 Application menu

Application menu is located in the upper part of the display screen. Its role is the same as in desktop applications. It provides access to functions and settings of the application.

3.2.2 Tab bar

The application uses tabs similarly to web browsers. There are two tabs - the first displays the image data, classes and training sets, the other tab displays statistics. Users can switch between tabs by means of the tab bar located below the application menu.

3.2.3 Viewport

The viewport shows active image data as layers. Active layers are selected in a sidebar (by means of checkboxes). The layers can be panned and zoomed and individual layers can be assigned different degrees of transparency.

3.2.4 Side panel

The side panel shows a list of the image data and classes. Checkboxes enable to turn on and off the individual items. Turning on a class also displays its training sets in currently displayed images.

4 PRACTICAL EXAMPLES

4.1 Comparison of strip-mined area in north bohemian coal basin

This user case demonstrates coarse estimation of the change of strip-mined area in north bohemian coal basin in the course of about 5 years. Two aerial images from different years were used as data sources. Only one class "mine" was created and the nearest neighbour method was used to detect this class in the images. The figures 1 and 2 show results of the classification (the detected strip-mined areas are indicated with red colour).

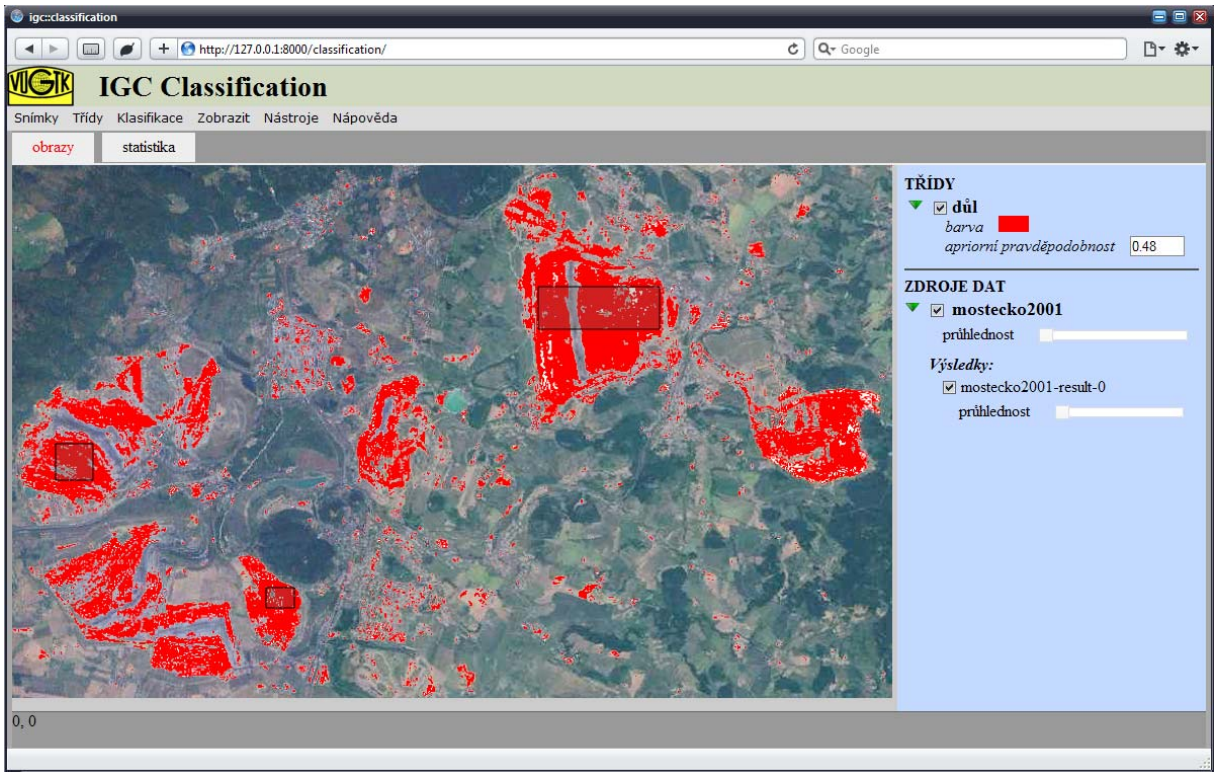


fig.1 - detected strip-mined ares, 1st phase

fig.3, 4 -statistics ofstrip-mines detection in 1st and 2nd phases

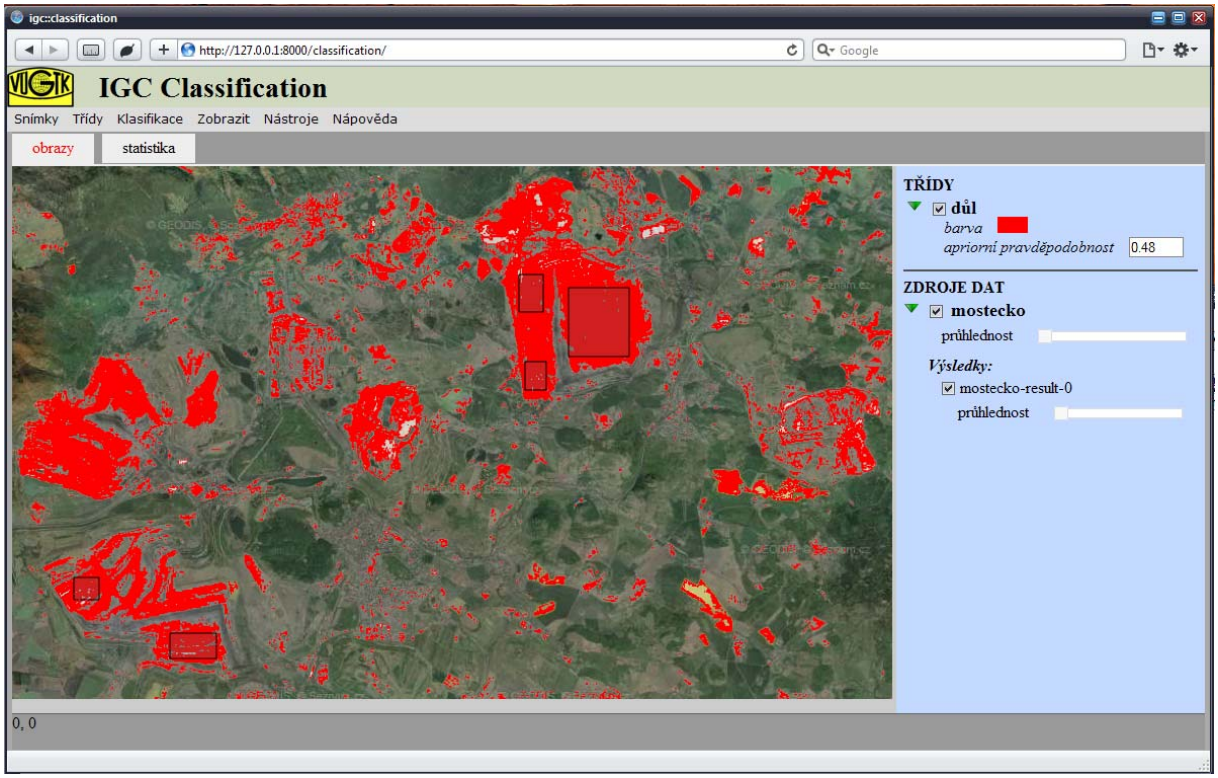


fig.2 - detected strip-mined ares, 2nd phase

The results show that the algorithm detected also areas that obviously aren't strip-mines. The cause is high similarity in colour of these areas to actual strip-mines. If we assume the amount of falsely detected areas is roughly the same in both images, we can ignore these errors.

The statistics (shown on figures 4 and 5) show the size of detected areas in square km and in percents (relatively to the depicted area). We can see growth of strip-mined area of 13.60 km2 or 27.5% (from 49.37 to 62.97 km2). The classification errors should be already eliminated the difference.

4.2 Determining the area of water surface

This example shows the process of determining the size of water areas, here presented by the river Vltava flowing through centre of Prague. The data source are aerial photos provided by CENIA as WMS (Web Map Service). All three classification methods were used to demonstrate the difference in results.

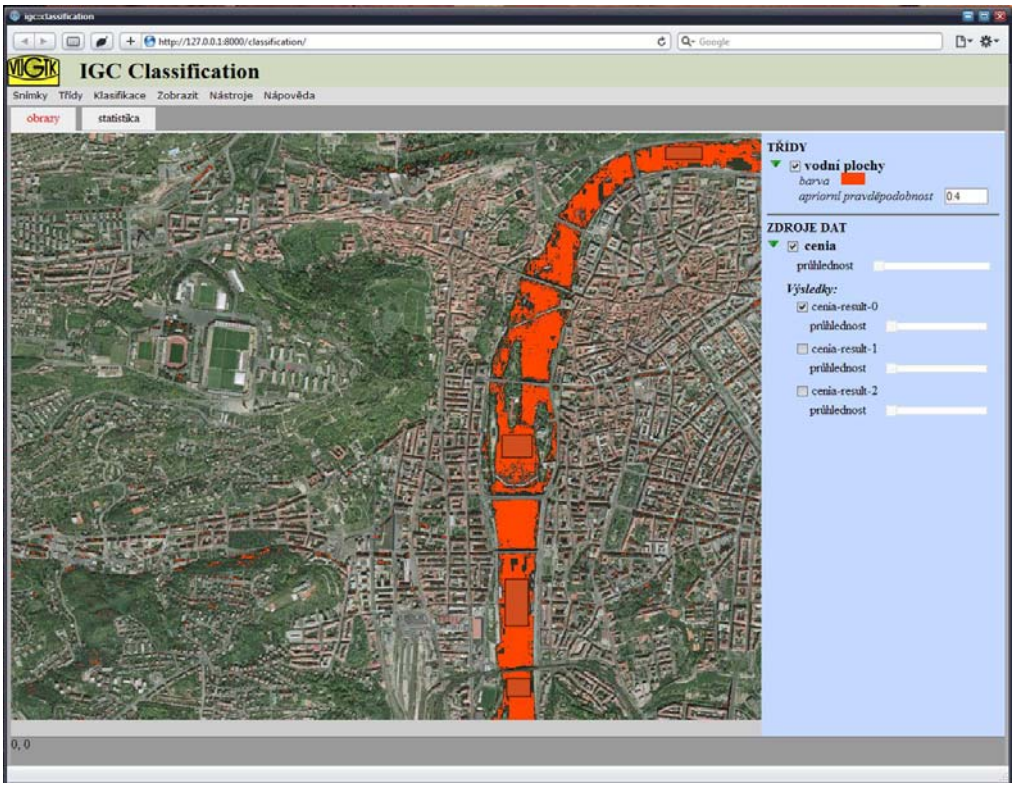


fig.5 - result of water surface detection - relative frequencies method

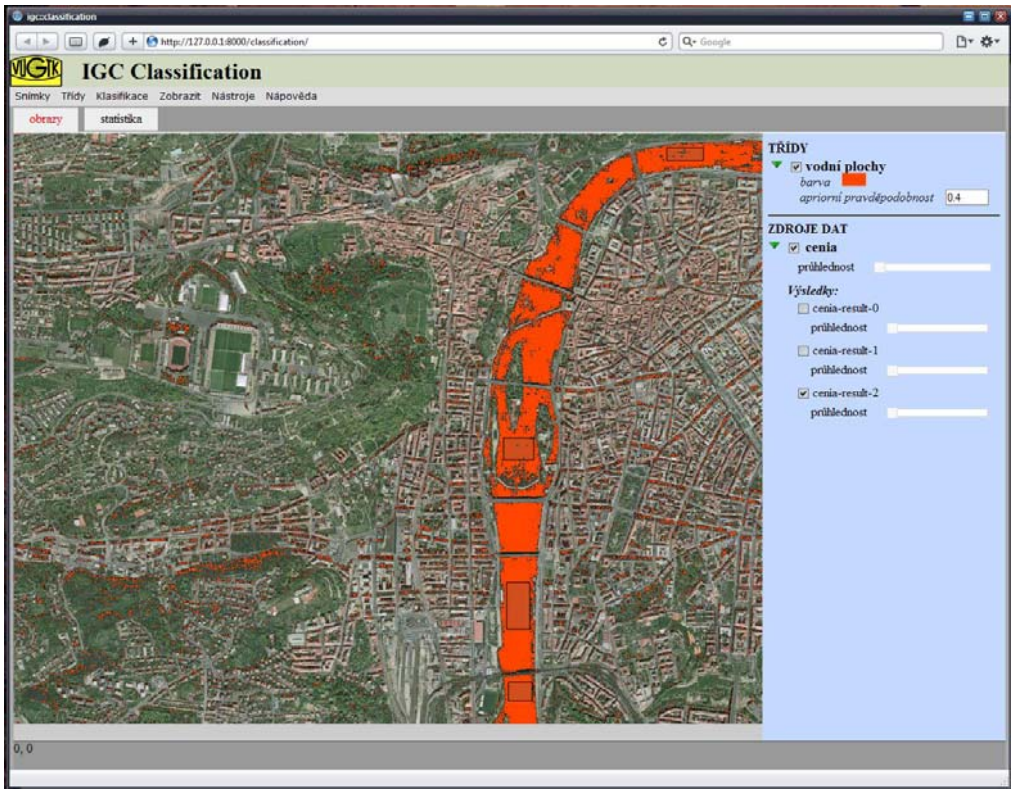


fig.6 - result of water surface detection - training set extension method

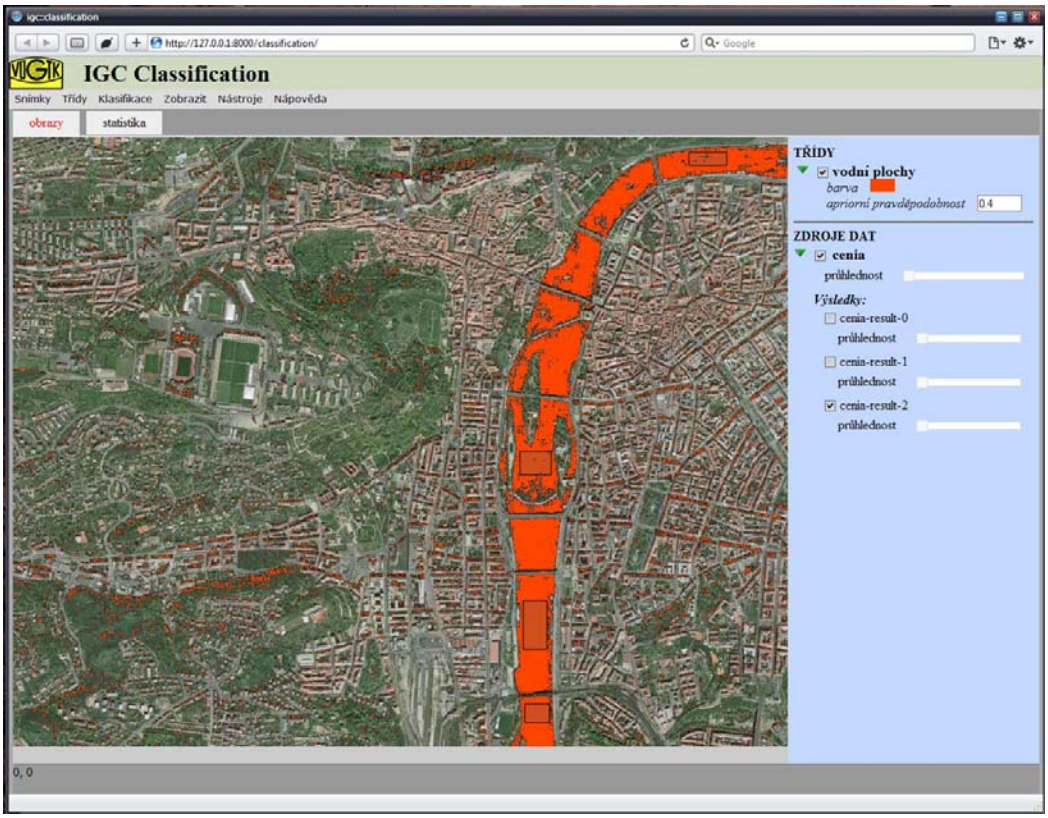
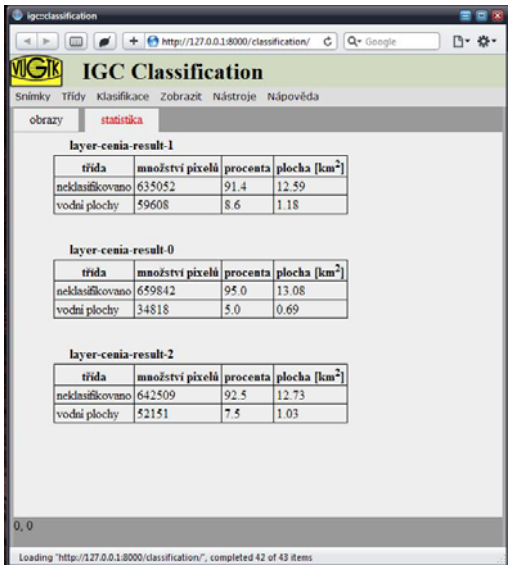


fig.7 -result of water surface detection -nearest neighbour method

We can see that in this case it's the simplest method that gives the best results. The reason is quite high degree of homogeneity and narrow colour range of the water surface. Because of that the result of the simple method isn't so grainy. On the other hand the other two methods, that in some way extend the training sets, include more pixels that do not belong to water areas, but are close to it in colour.



The statistics table documents this - we can see larger classified area for the more complex methods (training sets extension and nearest neighbours). Where the simple method detected only 5% of the image as water, the training sets extension method detected 8.6% and nearest neighbours method 7.5%.

This user case clearly shows the significant role played by method selection. Similarly important is the choice of pertinent parameters of the classification method (which is not documented here for brevity).

fig.8 -statistical overview of the various results of water surface detection

5 CONCLUSION

The objective of the contribution is to present research results in the field of Bayesian classification of digital images. Important part of this presentation is the message about possibility of using web application WACLASS to perform this classification. The results were achieved at the Research Institute of Geodesy, Topography, and Cartography in the framework of project InGeoCalc. The web application is available at *http://www.vugtk.cz/ingeocalc/igc/classification/* and allows the user to apply Bayesian classification on his own images data or on data provided by WMS (Web Map Service).

The test results of the implemented variants of Bayesian classification show significant influence of the selected variant as well as additional parameters. It is necessary to keep this in mind and try more variants using different methods and parameters to achieve satisfactory results.

REFERENCES

[1] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith: Bayesian Methods for Nonlinear Classification and Regression. Willey series in probability and statistics. John Willey & Sons, 2002.

[2] Andrew Webb: Statistical Pattern Recognition. John Willey & Sons, 2003.

[3] OGC Web Map Service Interface, ed. Jeff de la Beaujardiere [http://portal.opengeospatial.org/modules/admin/license_agreement.php?suppressHeaders=0&access_license_id=3&target=http://portal.opengeospatial.org/files/index.php?artifact_id=4756]

[4] ArcGIS Resource Center: What is the format of the world file used for georeferencing images? [http://resources.arcgis.com/content/kbase?fa=articleShow&d=17489]

BIOGRAPHICAL NOTES

Milan Talich (*1961) was graduated from the Czech Technical University (ČVUT) in Prague, Faculty of Civil Engineering, Department of Geodesy and Cartography. Since 1987 he was working at geodetic networks processing and geodynamic problems. At present he is focused on information systems oriented to web applications.

Ondřej Böhm (*1979) was graduated from the Czech Technical University in Prague, Faculty of Civil Engineering, Department of Geodesy and Cartography, specialization Remote sensing. Nowadays he works at processing of image data, creation of web applications and studies of using InSAR data for deformations.

Lubomír Soukup (*1963) was graduated from the Czech Technical University in Prague, Faculty of Civil Engineering, Department of Geodesy and Cartography, specialization

Remote sensing. Nowadays he works at theory of probability and mathematical statistics for geodetic measurements.

All of the authors are staff of the Research Institute of Geodesy, Topography and Cartography (VÚGTK).

CONTACTS

Milan Talich Ph.D.,
Ondřej Böhm,
Dr. Lubomír Soukup

Research Institute of Geodesy, Topography, and Cartography
Ústecká 98,
250 66 Zdiby,
CZECH REPUBLIC
Tel. +420 284 890 515
Fax + 420 284 890 056
Email: Milan.Talich@vugtk.cz
Web site: http://www.vugtk.cz/