Milan Talich*, Ondřej Böhm*, Lubomír Soukup*

Classification of digitised old maps and possibilities of its utilization

Keywords: Digitised old maps; classification; digital images; web application; web map services

Summary
This paper is concerned with classification of digitized old maps in the form of raster images. Examples of this classification and its utilization are presented. Web application of raster image classification is introduced as well. The web application can classify both individual images and raster data provided via Web Map Services (WMS) with respect to OGC standards (Open Geospatial Consortium).

Introduction

Because of their importance as historical sources, old maps are steadily becoming more interesting to researchers and the effort grows to use them with the help of modern and especially digital methods. However, the users are no longer satisfied only by simple digitization and on-line publication. Nowadays it is required to preserve and fully take advantage of all the specific attributes of maps that allow, for example, the measuring of distances, directions and areas. Even this isn't enough. Users demand more, some added value, that would simplify the use of old maps and allow to gain more information than from the classic use of analogue paper maps.

Expected user demands

General basic requirements on digitized old maps

Free on-line accessibility can be without question counted among the basic general requirements users have regarding old maps. The reason is apparent: the user - reader - researcher usually doesn't have to pay for access to paper maps in libraries and archives. The argument about costs of digitization equipment and its operation, servers and internet connectivity holds little weight, because these costs are only a different form of expenses otherwise spent on library staff needed to fetch the maps from depositories and put them back and other tasks connected to physical loans.
Simple browsing of maps can be easily replaced by an online service displaying digitized raster representations of original maps and allowing to pan and zoom the maps in convenient and unobtrusive way. An example of this way of map publication is the Zoomify software.
However, the preservation of basic cartometric characteristics of maps (where applicable of course) can also be considered as a general demand for digitized maps. To preserve these cartographic attributes the digitized maps must be georeferenced, i.e. placed into a coordinate reference system with regard to their map projection. The georeference then must be taken into con-

*Milan Talich Ph.D., Ondřej Böhm, Dr. Lubomír Soukup, Research Institute of Geodesy, Topography and Cartography, Ústecká 98, CZ 250 66 Zdiby, Czech Republic [Milan.Talich@vugtk.cz]

sideration when publishing the maps on-line, so users can measure distances etc. directly on the display. In this way the digitized map keeps its cartographic attributes.

Adding value to digitized maps

When old maps are properly digitized georeferenced and published on-line it's logical that the users attempt to use them as effectively as possible. They expect some kind of added value compared to paper maps that would give them better use of the map data than before. The following paragraphs present some thoughts as to what this added value might be.

Search and display maps available on-line

The most basic utility is a good search tool, that would allow to find on-line available maps pertinent to a given area or problem from as many sources as possible (archives, libraries, museums etc.). A typical use of the application would be for researchers looking for maps depicting certain area with certain scale range at a certain time. These parameters would serve as search criteria with the area specified either as coordinate range or with graphical selection in a general map. The search result would be a list of all found maps with links and metadata.

This problem is currently at the forefront of interest of several institutions and several applications have been created that address it: for example Arnaud (2011), CartoMundi (2011), Geographical Map Search (2011) or Geografické hledání (2011) used in the Moravian Library.

Drawing comparison

Digitized maps offer great possibilities for comparing the drawing of two or even more maps. In contrast with paper maps the digital ones can be placed side by side on the monitor independent of where the paper originals are stored. Also the possibility of changing the scale of individual maps makes comparisons easier.

Even more useful is comparing the maps by stacking them onto each other as several layers and adjusting the opacity of these layers. This method allows for better discerning differences in the various map elements (courses of communication, forested areas etc.). As useful as this method may be it is more demanding on the provided data. In order to be able to combine the maps in this way with any ease, the maps must be georeferenced and served in accordance with common standards. Although there are several standards, by far the most widely used is the Open Geospatial Consortium (OGC 2006) Web Map Service (WMS) specification.

Where map series consisting of more than one map sheets are concerned, it is necessary to perform the so called mosaicking - join the individual map sheets into one seamless whole. From a technical standpoint it isn't necessary (and sometimes even not possible) to create and store one huge file comprised of all the individual map sheets. Instead, the joined maps are cut into regular tiles with georeferencing information. These tiles are then put together as needed on the fly (or cached at convenient scales) to produce a map covering the desired area. Unfortunately mosaicking is far from an easy task. Many problems can be encountered in the process and solving them all usually requires some compromise. The shrinkage and distortion of paper must be taken into account, as well as possible inaccuracy of drawing. Last but by no means least, the map projection of a given map must be considered when joining the map sheets.

The area around Městec Králové can serve as an example of map drawing comparison. When someone is interested in the changes in land use in the last 150 years, she can use e.g. the web application on the map portal of VÚGTK (*http://mapy.vugtk.cz*). The application offers several

map series. Among others it is map series II. Military Survey (also known as Survey of Franz Joseph I.) created between 1836 and 1852 in scale 1:28 800 (for example seeFig. 1). These maps can be displayed together with other maps or a current orthophotomap (example shown in Fig. 2). Each map series is displayed in its own layer and the layers' opacity can be adjusted, which allows for fairly comfortable comparison of both maps as demonstrated in Fig. 3. By inspecting the combined image we can easily see how the large lake Štítarský from 19th century has vanished over the years, leaving only two small lakes (Krčský and Štítarský) in its place. A railway crosses the former lake's area (which can be inconvenient for the railway's quality). By utilising other map sources we can find out that the lake was in place as early as 1720 (as shown on Müller's map) and it has already been greatly reduced in size in 1878 (as documented by maps from the III Military Survey).

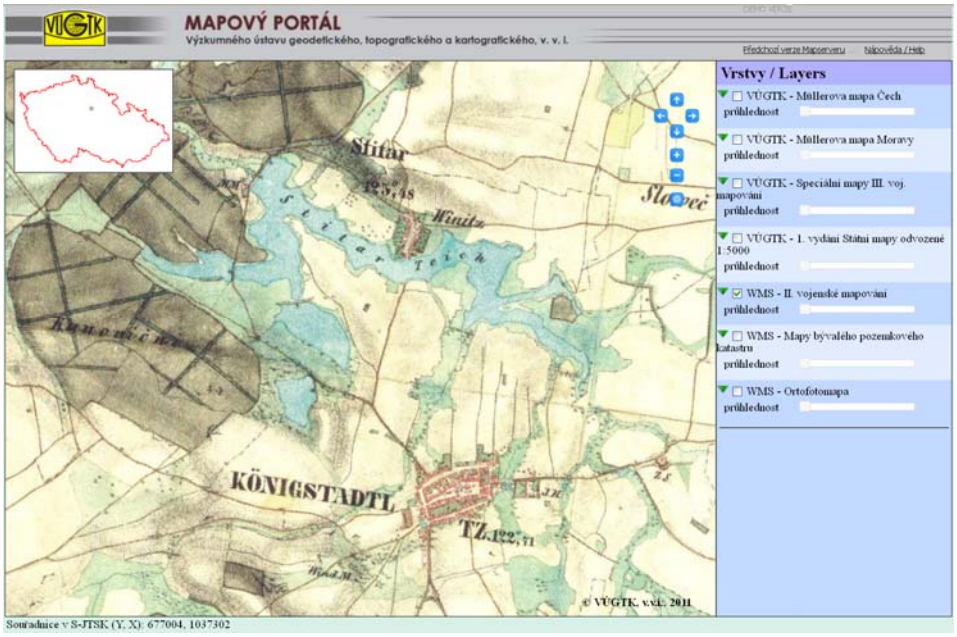

Figure 1. Lake Štítarský near Městec Králové / II. Military Survey.

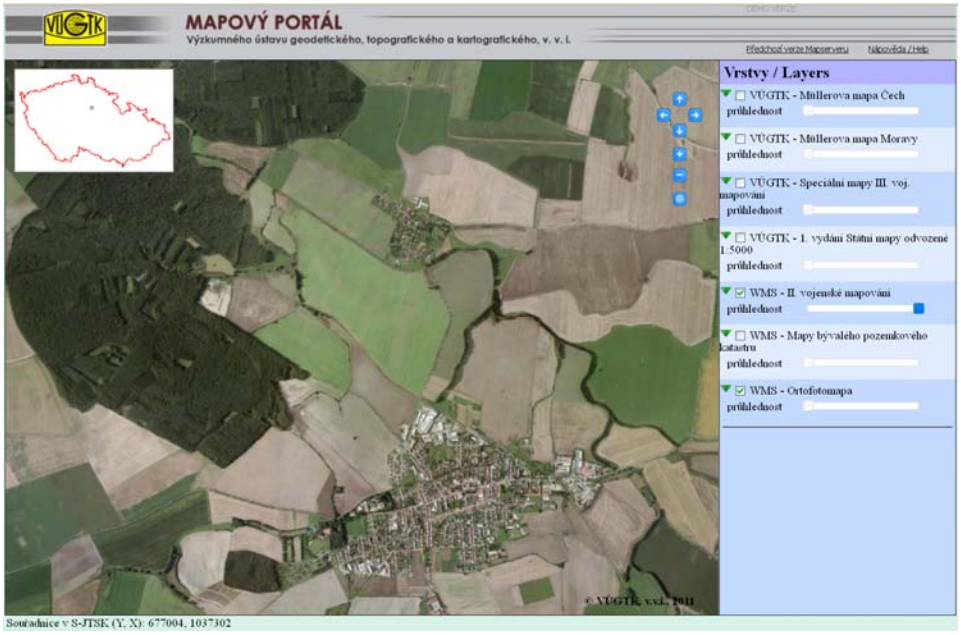

Figure 2. Lake Štítarský near Městec Králové – contemporary othophotomap.

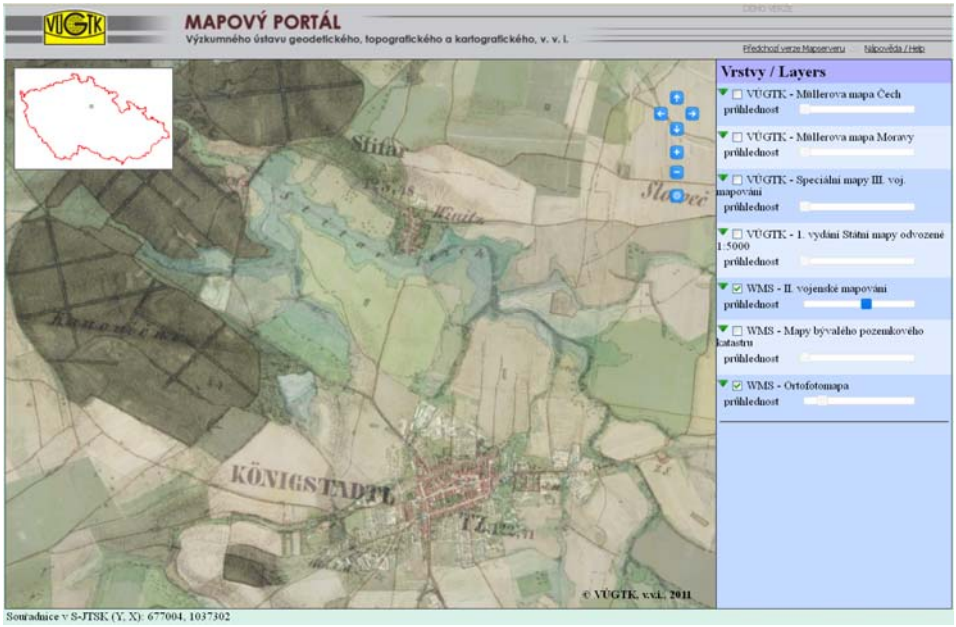

Figure 3. Lake Štítarský on overlay of III. Military Survey and contemporary orthophotomap.

Automatic detection of map elements and classification of raster images

Another important enhancement of digital maps can be the automatic detection and recognition of map objects. These tools would detect with certain level of reliability for example map symbols. It would be useful especially when dealing with large map series with many map sheets. Because these map series have in common the map projection, the method of representing earth surface, the map key and colours, such a tool would make work with these maps much easier for researchers.

Another useful tool is the automatic classification of digital maps. This process allows for automatically detecting areas with common characteristics, i.e. forests, water surfaces, buildings etc. Technically it is a problem of assigning the image's pixels to one of several classes defined in advance. If the map is georeferenced the classified image can be used to determine the surface areas of the classified regions, or otherwise evaluate their position.

Unfortunately quite substantial difficulties can be expected when attempting to apply these tools. The main cause of these difficulties is the varying quality of digitized maps resulting from damage caused to the original maps by time or storage conditions and from varying scanning procedures. Even individual maps from a single map series can differ quite a lot.

The basic prerequisite for processing old maps in this way is to have them scanned and published on-line in standardized way, because only then the maps will be available easily enough to research, develop and try sufficiently robust and efficient solutions to presented problems. Classification of raster images will be further discussed in a later section of this paper with special emphasis on the Bayesian methods of classification.

Using digitized maps in custom applications

The possibility of using digitized maps in custom applications is probably the most useful added value. If the digitized old maps are accessible free of charge and in accordance with agreed upon standards then individual users are able to use them in their own application serving their specific needs.

The point is that the set of potential users of old maps is very large and maps can be utilised in almost every field of human activity. It is therefore impossible to anticipate every possible use case and create the appropriate application. It is much more efficient to provide the old maps as a data service and allow users to create their own applications and tools suited for their needs.

Classification of digitised old maps

Problem formulation

Regions with some characteristic features have to be localized in the digital image of a digitised old map. 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 the assignment of a specified class to each pixel in the given digital image.

The required result

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

Review of the main classification methods

A vast number of different classification methods has been designed during the short history of development of computer image processing. Two main groups of classification methods can be recognized: deterministic and statistic. Another distinction between classification methods is based on the practical circumstances of the solution of a 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. Webb, 2003, Denison et al., 2002), presents more powerful toosl than the other kinds of classification.
Statistical characteristics of 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. A training set is a region in the image which represents well a certain class. Searching for other regions with similar characteristic features is the task of supervised classification. The principle of statistical supervised classification is based on the 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 the colour space RGB. Each pixel of a digital image displays as a point with coordinates that are the features, e.g. colour 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 labelled by an identifier of a corresponding class. With the aid of the labelled points of training sets, other points in feature space have to be labelled to complete the classification. Hence the classification task can be formulated as a rule for labelling pixels displayed in feature space. This rule, called classifier, can be searched for by means of several man-

ners. The most common classifiers are e.g. linear, nearest neighbour or Bayesian classifiers. Bayesian classification, which is the most important member of the family of statistical supervised classification will be studied in the following.

Bayesian classification

Input data and assumptions

A digital image is given where training sets are determined. A 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.

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. Webb, 2003). 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 the 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 the probability that a pixel of class C has feature vector F. Under these assumptions for known prior probabilities P(C) the Bayes formula is:

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

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

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

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

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

Extended variant

The extended variant is based on the 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 centre 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 , $n_{C,F}$. A better estimation of probabilities P(F | C) could be reached by this way. The problem is in the definitions of riskiness and sufficient closeness to the centre of cluster.
The distance of additional pixels from the centre of the 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 the wrong pixels.

Nearest neighbour variant
This variant is based on the 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 centre 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.

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

where E_h depends on the risk of appending wrong pixels.

Practical on-line solution

A web application for the 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 the design of classes, the 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). The 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 the Internet with any web browser. The only exception is Internet Explorer of a version older than 9, since it does not support SVG. Communication client - server is asynchronous.

Features of the application

The application offers all the necessary tools for supervised classification of digital images. It allows creating classes, displaying image data, and defining training sets.

Classes and training sets
Definition of classes consists of entering necessary information about each class: name, colour, and a prior probability. The colour will be used in the final result of the classification. The colour as well as the prior probability can be modified during the classification process.

Training sets are formed graphically as rectangles. The 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.

Data sources

Either image files or WMS data (i.e. data provided by Web Map Service) can serve as data sources. 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 the 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 (ArcGIS Resource Center, 2001), can be saved to a local computer.

Variants of classification

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

Basic variant

This variant uses relative frequencies of pixels in training sets. The advantages of this variant are 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 the wrong pixels. The principle of this variant is described in a previous paragraph.

Nearest neighbour variant

This variant uses clusters in colour space as the previous one. The classification criterion is the distance from the centre of a cluster. The principle of this variant was described previously.

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).

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 unclassified pixels. Furthermore, a brief overview of classified areas is presented (these areas are meaningful only for georeferenced data as they are computed from the spatial resolution of the image).

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

Application controls

The application user interface attempts to imitate desktop applications. The browser's display area is split into application menu, tab bar and current tab's content. Most of the functions are executed from the application menu.

Application menu
The 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.

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.

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.
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.

Practical example

This section presents an example of the automatic classification for estimating the surface area of former lake Štítarský in the first half of the 19th century. The lake is situated near the present day village of Vinice, near Městec Králové. The lake can be found on II. Military Survey maps, but it no longer exists today in its former size. By comparing the old maps with contemporary maps we can see that only small remnants of the original water body are left. The original size of the lake can be found out by using a web application for raster image classification accessible on the URL *http://www.vugtk.cz/ingeocalc*. This application is part of a knowledge system for decision support based on geodata created in VÚGTK between years 2006 and 2011. This application can display (among others) the maps of II. Military Survey (provided as WMS) and use it as data source. When the map is displayed it is necessary to select training areas – representative samples of the areas of interest. In this case, the area of interest is water surface and it is best represented by several rectangular areas inside the lake Štítarský (Fig. 4). Based on these training areas the application classifies the image, i.e. marks pixels with satisfiying degree of similarity with pixels of training areas as water surface. Based on the characteristics of the processed image it might be necessary to try different classification methods and/or adjust the parameters of these methods. When the result of the classification is satisfactory Fig. 5), the surface areas of classified regions can be computed by using the „Statistics“ utility from the „Tools“ menu. This tool generates a table listing the total surface areas of all classified regions in the image. In this case the area of water surfaces - the area of lake Štítarský in the first half of the 19th century - is approximately 112 ha. For better demonstration a layer with a contemporary map source can be displayed, for example the Basic Map of the Czech Republic 1 : 10 000 (as shown in Fig. 6). The classified region then clearly designates the area where the lake used to be and changes in the area can be easily discerned.

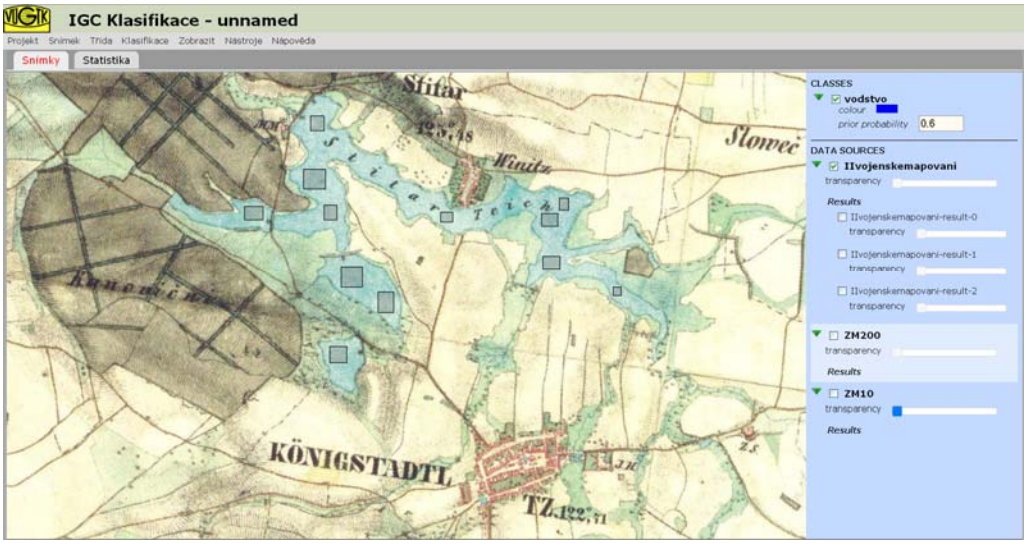

Figure 4. Lake Štítarský, training areas selection.

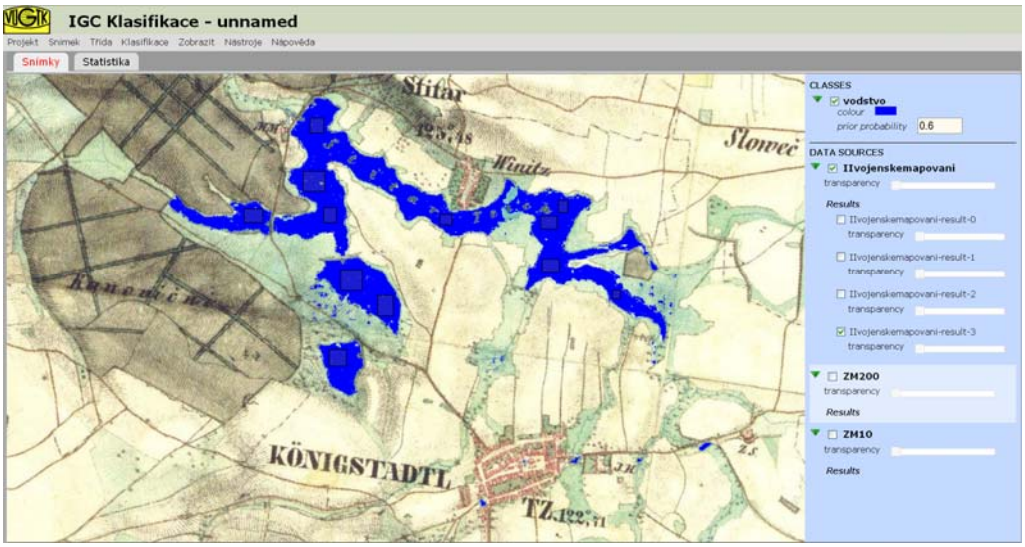

Figure 5. Lake Štítarský, result of classification.

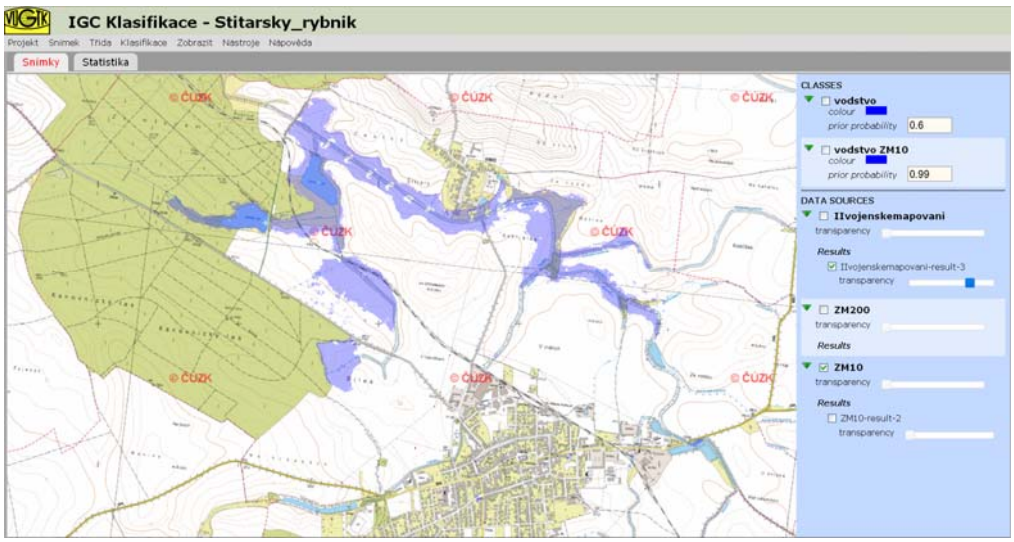

Figure 6. Lake Štítarský, classification result displayed over contemporary map source.

Conclusion

The goal of this contribution was to point out the fact, that today users – readers from libraries, archives etc. are not content with bare on-line accessible maps any more. Nowadays an added value is required – tools that allow to use the digital maps more efficiently, easily and offer the possibility to get more information from them than from the original paper maps. Additionally this article highlights several possible means of adding this value and attempts to point out pre-requisites and potential problems of these tools. It discusses in greater detail the methods for automatic classification of raster images and presents a practical example of use for classification of old maps.

References

Arnaud J.L. (2011). Cartomundi, a new interface to find maps by geo-localisation. In *Proceedings of the 25th International Cartographic Conference.* Paris: 3–8 July 2011, ISBN: 978-1-907075-05-6
http://icaci.org/documents/ICC_proceedings/ICC2011/Oral%20Presentations%20PDF/C3-Digital%20technologies%20and%20cartographic%20heritage/CO-266.pdf

CartoMundi (2011). Online Enhancement of Cartographic Heritage - Find maps by geographic location. http://www.cartomundi.fr/site/CDxx.aspx

Denison D. G. T., Holmes C. C., Mallick B. K. and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Willey series in probability and statistics. John Willey & Sons. ISBN: 978-0-471-49036-4

Geographical Map Search (2011). http://kartenportal.mapranksearch.com/en/

Geografické hledání (2011). http://mapy.mzk.cz/hledat/

OGC (2006). OpenGIS Web Map Service (WMS) Implementation Specification
http://www.opengeospatial.org/standards/wms

Talich M., Böhm O. and Soukup L. (2011). Bayesian Classification of Digital Images by Web Application. In: *FIG Working Week 2011*, 18-22 May 2011, Marrakech, Morocco, 1-13, ISBN 978-87-90907-92-1.
http://www.fig.net/pub/fig2011/papers/ts05i/ts05i_talich_bohm_et_al_4827.pdf

Webb, A. R. (2003). *Statistical Pattern Recognition.* Second Edition, John Wiley & Sons, Ltd, Chichester, UK. ISBN: 9780470845134