Image Processing in Cooperation with GIS

Dana Klimešová, Jakub Konopásek

***Abstract* -** Our world and its phenomena are highly dynamic in space and time. Keeping temporal and spatial information precise and up-to-date is of much importance to the planning and designing of our activities in the landscape. In this paper, the method and the potential of integration of image data and GIS information layers is addressed to refine and update information for the use of temporal GIS. The modelling and monitoring of our environment and society is mainly based on the fact that our surrounding is continually changing. Temporal data processing is in our fast changing world crucial for the directing of disaster prevention, demographic development observation, management of military operations, large fires, and traffic problems and so on. This paper contemplates the contribution of object – pixel context to improve and refine temporal data and support automatic way of processing because manually update of geographical data is unpractical and highly time consuming.

***Keywords* -** Spatial-temporal modelling, object - image integration, sub-object resolution, contextual modelling.
.

I. INTRODUCTION

A. Time and Space

Images created by remote sensing are now widely used for various purposes in different fields and they have many different applications. One of the most typical use for these images nowadays is as a raw data for spatiotemporal analysis. Basic example of this is detection of changes using high resolution satellite imagery. With right tools and methods we can discern and document space-time changes including object recognition.

And of course we can also document change in attributes of respected objects. [1]. Urbanization, rapid land use and change of land cover has taken place in the world in the last decades. In this context, comparison of obtained results from new high definition raster imagery and existing vector data structures is one of the most important issues. Time and space are the main dimensions of our lives. To describe and map our environment we use the mathematical model with spatial or spatiotemporal coordinates. And in geographical information systems we use them for spatial and spatial-temporal modelling [1], [4].

Not only government and state administration, but a lot of industries and companies never considered GIS and satellite images to be the source of prosperity and efficiency.

But that is slowly changing. Spatially referenced data, stored as data in GIS are now needed for use in numerous applications like internet map services or for analysis and administration in private and public services. For such applications it is of highest importance that data that are used are up-to-date and highly correct. Manually updating data each time is very time consuming, therefore support through automatic systems is required [2].

B. Raster - Vector

Many GIS users face the problem of acquiring accurate and timely suitable data at a cost effective price. Finding features in high definition remotely sensed imagery can be a time-saving way to define and update geographical layers.

Moreover, the ways of

- ♦ managing and distributing data,
- ♦ data and information sources,
- ♦ data management tools and techniques

are continually changing.

Space technology has been gaining more and more ground in our life. Earth observation images sometimes turn into an irreplaceable source of independent and up-to-date information about the condition of an area.

Satellite data as well as aerial data is used to monitor fire situation, seasonal and flash floods, construction activities, forest management as well as environmental and navigation situation, etc. [3].

Manuscript received in July 26, 2012. Revised version received in August 26, 2012. Dana Klimešová is with the Czech University of Life Sciences, Prague, Faculty of Economics and Management, Department of Information Engineering, Kamýcká 129, 165 21 Prague, CZECH REPUBLIC, (e-mail: klimesova@pef.czu.cz), and with the Institute of Information Theory and Automation, Czech Academy of Sciences, Department of Image Processing , Pod vodárenskou věží 4, 182 00 Prague 8, email: klimes@utia.cas.cz).
Jakub Konopásek is doctoral student on Czech University of Life Sciences, Prague, Faculty of Economics and Management, Department of Information Engineering, Kamýcká 129, 165 21 Prague, CZECH REPUBLIC, (e-mail: konopasek@pef.czu.cz).

Satellite imagery gives us much needed data to use in:

- efficient expansion and modernization in economic sector (application in logistic, investment, large industrial projects, transport of goods, communication),
- operational services for emergencies monitoring, control and response (strategy planning, fire situation, seasonal floods and flash floods, water areas),
- creating solutions for various tasks concerning nature protection, including prevention and control,
- integration of satellite data, GIS and web technologies.

When high resolution raster images with spectral information are available, so called multispectral images (usually satellite data with multiple wavelength channels), we can use many different methods for classification pixel by pixel into specific classes [1], [4].

In addition to spectral features we can use also textural and in general structural features to refine and optimize the classification. Raster data classified in this way can then be converted to vector layers and than exported to a variety of vector formats.

This process is usually much faster and easier than manually digitizing from the raster image. It is a cost effective way to update our GIS with accurate and up-to date layers or add new vector layers to geo-databases [5].

Frequently used raster classifiers in image processing include standard approaches such as Bayesian classification, maximum likelihood and minimum distance classifiers as well as more sophisticated methods.

For example ESRI together with ENVI has in disposal the package focused on less and more sophisticated classification methods, including neural network and decision tree classifiers.

C. Statistical Pattern Recognition

Recognition (classification) is assigning a pattern/object to one of pre-defined classes.

There are two basic approaches:

- supervised classification - training set is available for each class to give us good orientation in feature space,
- unsupervised classification, with the set of additional useful parameters (clustering), which is primarily used when training set is not available.

Desirable properties of the training set are that [6]:

- it should contain set of typical representatives of each class including intra-class variations,
- to be reliable and large enough,
- should be selected by domain experts.

After that, it is necessary to setup the classification rule. That means the way of the feature space partitioning.

Each class is characterized by its discriminate function g(x) and classification means maximization of g(x), where feature x is a point in n-dimensional metric space (usually Euclidean space).

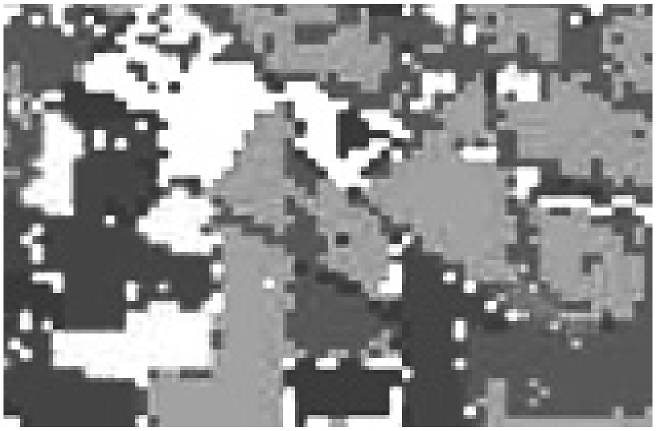


Fig.1. Pixel-based classification.

We assign x to class i if

$$g_i(x) > g_j(x) \tag{1}$$

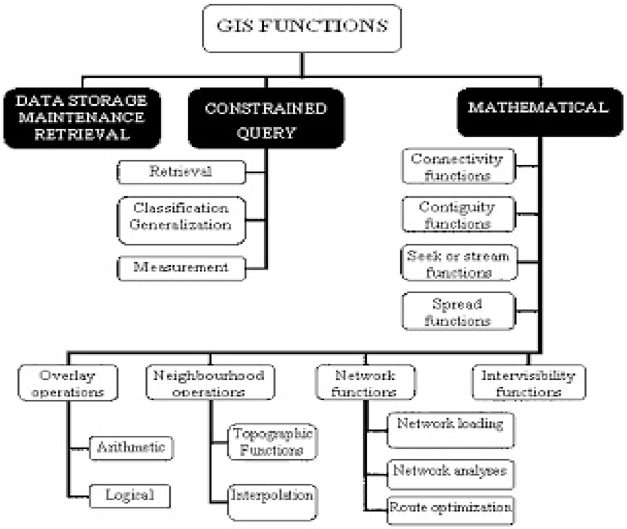


Fig.2. GIS functions.

Discriminate functions define decision boundaries in the feature space. We need a geographic object-based image analysis to bridge the gap between the image processing and the geographic vector database.

We can do it with the support of GIS analytical functions and geographical vector database that contains a lot of useful information including contextual.

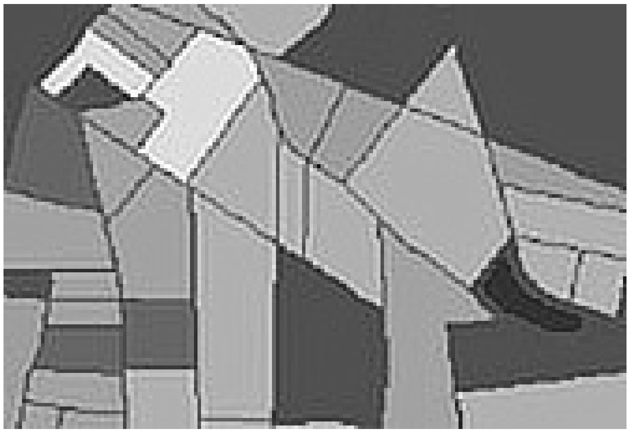


Fig.3. Object-based classification.

II. GIS-BASED TEMPORAL ANALYSIS

A. High Resolution Data

In this research an integration of remote sensing (RS) and GIS data was used to identify and map the ability to support spatial-temporal analysis and dynamics modelling [6].

Remote sensing has made some great improvements in last few years - especially in spatial resolution, spectral resolution (including hyper-spectral bands) and nearly temporal resolution on demand.

All these advancements including aerial data play a significant role in detection of changes in observed area. Change can be in shape, location and of course existence of the object. But as mentioned above, high resolution data also allows generating many non-spatial characteristics and attributes and combinations of above mentioned.

From the GIS point of view data model, as basically any model, must represent the real world - which means data needs to be consistent and up-to-date. We can speak about quality of the model [6], [7]. There are three basic measures for quality control:

- correctness,
- accuracy,
- completeness.

All these quality measures can be derived by comparing the GIS data to the real world. Especially in temporal GIS the regular update is integral part of the job. It means, we need a high degree of automation to reduce the amount of manual work [8].

For this, one of the possible ways is to use a combination of suitable aerial data with satellite imagery like IKONOS to detect, monitor and document temporal changes.

The data, information and knowledge management is an essential process improving competitive advantage.

In this way we are able compensate:

- the loss of procedural knowledge,
- the loss of know-how and customer-related or project-related experiences,
- the loss of the management information [9].

Our decisions are becoming increasingly dependent on understanding of complex relations, on deep context and understanding of the dynamics of phenomena in the world around.

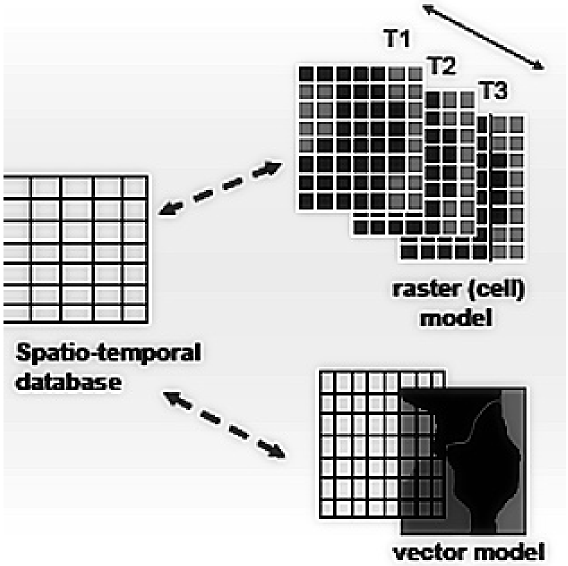


Fig.4. Integration of RS and GIS data is the way to support spatial-temporal analysis and dynamics modelling.

.

B. Integration of RS and GIS data

The main goal of this paper is to show the advantages and possibilities of proposed method of integrating high resolution imagery with GIS object data. For example integrate IKONOS images obtained by fusion of high resolution PAN and MS images with vector object data for the purpose of temporal object- oriented image analysis [10].

In general, we can speak about GIS-based analysis using raster and vector data of the tested area. For growing amount of geographic vector databases it is necessary to quickly produce up-to-date geographic objects and only object-based image analysis has enough strong potential to offer viable solution.

The key problem is that we move and interconnect geographic objects – vector and pixel-oriented raster data. This of course has its advantages and disadvantages [10], [11].

First problem is that a geographic object is not necessarily the same as classification class (see fig. 1 and 3). Different regional and state administration purposes influence the objects definition. On the other hand the image analysis classification results are in disposal in sub-objects resolution level. We can take advantage of this in several ways:

- to create, define and redefine the geographical objects – fig. 5, 6.
- to clean the training sets by a non-parametric procedure (to keep them most relevant).

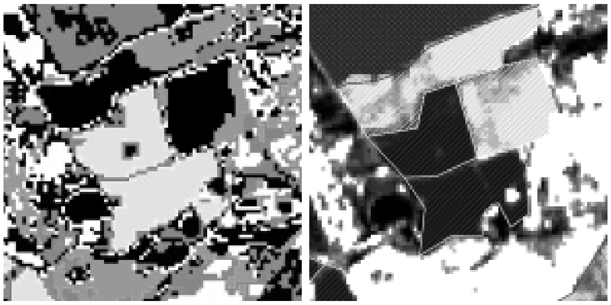


Fig.5. The use of sub-object resolution level to define and redefine objects.

In this context we can take a closer look at object change detection evaluation. In the initial step we can associate each GIS vector layer with the corresponding region of classified image objects from the image analysis data using common location in the map [10], [12].

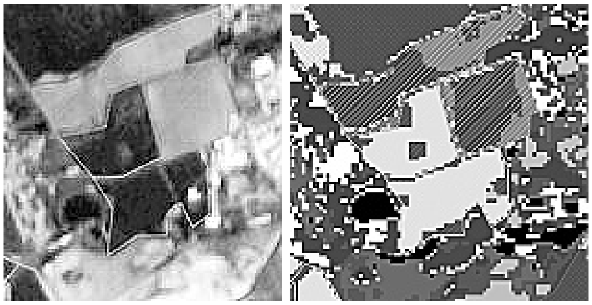


Fig.6. The use of sub-object resolution level to clean data sets.

High correctness of GIS data is of highest importance. Manual updates are very time consuming and support through automatic systems is required. Integration and comparison with existing vector data structures is the most important issue.

C. Statistics and measurement

High resolution images like IKONOS form the input for combined object and pixel based classification and change detection. And they also improve classification accuracy.

Thanks to the incorporation of evaluation step we are able to achieve sub-object resolution for update changes.

To clean training sets we can use the comparison of vector and image data sets and provide training samples being able to keep obtained results to be more robust. Under the idea of robustness we understand a stable system that is not affected by changes in its structure and surroundings [8], [20].

From the point of view of the mathematical statistics, the robustness means the low sensitivity of the system to the rough errors in the input data. Whatever we understand by robustness, if we can, we apply robust techniques as good tool for GIS data refinement. Only then we can provide solution that will work, even when we are assuming that there can be some insignificant changes in the input or in the implementation.

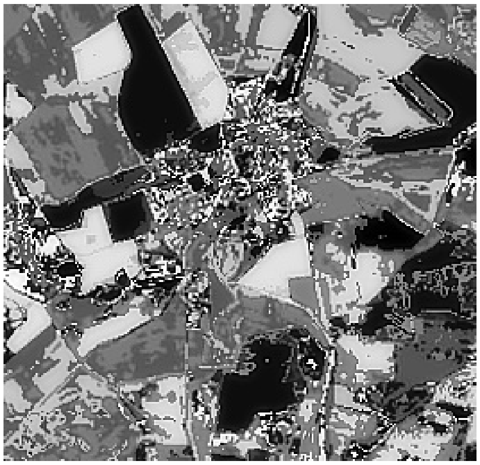


Fig.7. Each image-object is classified based on a maximum likelihood and new classes are created by decision rules.

The mathematical statistics deals with situations where we have finite number of available measurements data from which we estimate important system parameters.

The use of remote sensing data and GIS data provide more advantages. We can quickly extract many of statistical information from the GIS database and use them to set up factors, measure the distances and calculate useful parameters.

If the measured data are loaded with the errors and statistics as the discipline always assume it, then we can expect that the resulting estimate of the random variable will be biased in a similar manner [10], [11].

In practice if we are able to recognize the remote (poor) values then we can delete the values, and prepare the input data for the statistical analysis. The theory of mathematical statistics rejects such procedures as unclean.

On the other hand there exist a number of so-called robust statistical methods in mathematical statistics. These methods allow to some extent that the data is represented by very remote values, but the final calculation of the estimate is made so that remote values do not have adverse effects at all or only partially [14].

III. TO BUILD AND UPDATE GIS DB

A. The robust methods

Despite the importance of data collection and analysis, data quality is always problem. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of sophisticated approaches. As a result, there has been a variety of research over the last decades on various aspects of data cleaning.

We can take a statistical view of data quality, with an emphasis on intuitive outlier detection and exploratory data analysis methods based in robust statistics. In addition, we stress algorithms and implementations that can be easily and efficiently implemented and which are easy to understand.

We meet so called distillation errors. In many applications, raw data are pre-processed and summarized before they are entered into analysis. The data distillation is done for a variety of reasons [13]:

- ♦ to reduce the complexity or noise in the raw data,
- ♦ to perform specific statistical analysis,
- ♦ to emphasize and aggregate properties of the raw data,
- ♦ to reduce the volume of data being stored.

All these processes have the potential to produce errors in the data.

Data integration errors – for example to contain data from a single source, collected and entered in the same way over time as well as data collected from multiple sources via multiple methods over time.

Moreover, in practice many databases evolve by merging in other pre-existing databases and this merging task requires some attempt to resolve inconsistencies across the data. Any procedure that integrates data from multiple sources can lead to errors.

A lot of typical statistical methods are very sensitive to even the smallest failure of assumptions, particularly the assumption of normal distribution. Therefore, we are interested finding the method that would eliminate the mentioned problem. That means method that is insensitive to small deviations from the assumptions of the model. These requirements are meeting by the robust methods [14], [18].

B. Data description

It is not possible to use only summary statistics like means and standard deviations when we analyse data sets for outliers we need more metrics. And more, this assumes a model based on the normal distribution, which is the foundation for much of modern statistics outlier detection techniques.

The dispersion shows the spread of values around the mean value. The most familiar metric of dispersion is the standard deviation, or the variance, which is equal to the standard deviation squared. Usually we assume the normal distribution (Gaussian), see fig. 8.

The normal distribution is defined by a mean value μ and a standard deviation σ

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2}$$

And has the probability density function

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \tag{3}$$

Robust statistics considers the effect of corrupted data values on distribution and develops estimators that are robust to such corruptions. Robust estimators can capture important properties of data distributions in a way that is stable in some sense to many types of corruptions of the data.

Characteristic of variability tells you how the values are concentrated by file or rather dispersed around the average. Measures of variability can be expressed absolutely or relatively. The relative expression means that it is related to the relative magnitude indicating the position distribution.

In relative terms - relative margin – *R/x*, the value is determined as a fraction or in %, where

$$R=x_{max}-x_{min.} \qquad (4)$$

We use also characteristics of skewness and kurtosis. Skewness expresses how the values are symmetrically or asymmetrically distributed around the center of the measured values (different values inflated below the middle and above middle).

Characteristics of the arithmetic mean, variance, skewness and kurtosis coefficients belong to a group called moments. Moments can be general, and these are calculated from the sum of the squares of the measured values [18].

It may also be the central moments (mean absolute deviation, variance, skewness and kurtosis) when the moment is calculated from the sum of deviations from the average measured values (center), either on the sum of squared deviations (mean absolute deviation) or squares (dispersion).

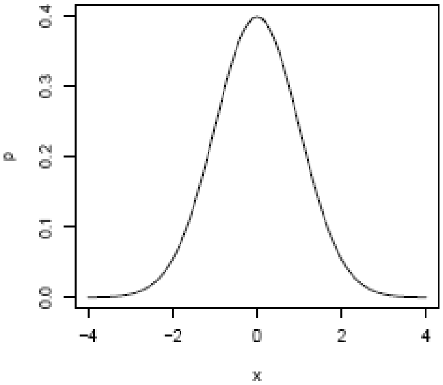


Fig.8. Probability density function for a normal distribution with mean 0 and standard deviation 1. The y axis shows the probability of each x value - the area under the curve sums to 1.0 it means 100%.

C. Median and residuals

Beyond the mean value and dispersion, a third class of metrics, that is often discussed and used, is median which describes how symmetrically the data is dispersed around the centre. Quantile, specifically p-quantile percentage is the value that divides the decreasing number of order statistics in two parts so that one contains p% quantile values less than or just the same, and the second contains a 100-p% larger or just the same.

The median is fifty percent quantile, the value that divides ordered statistics into two halves (the file has a 50% smaller values and 50% of values greater than the median).

When using the median as a center metric, a good robust metric of dispersion is the Median Absolute Deviation (MAD)

$$\mathrm{MAD} = \mathrm{median}_i \left(\left| X_i - \mathrm{median}_j (X_j) \right| \right), \qquad (5)$$

which is a robust analogy to the standard deviation. It measures the median distance of all the values from the median value.

Robust measure of scale quantifies the statistical dispersion in a set of numerical data. The outlier is a value that is too far away (considering defined robust dispersion metric) from the defined the robust center metric [14], [19].

The most frequently we meet in practice the occurrence of distant observations that can be caused either by inaccuracies in measurement or in the transmission of measured data. There is a real and motivating need of good statistical model.

Each model is affected by data so that there is no model that best fits to all the data [21], [22].

Another approach is to choose a model other than a normal distribution to model the data, for example multimodal distribution. Given such a model, the distribution of the data can be compared to the model.

The differences between the empirical data and the model are called residuals. Residuals of selected models are very often normally distributed. It means that standard outlier detection techniques can be applied rather than the data to the set of residuals.

D. Mahalanobis Distance

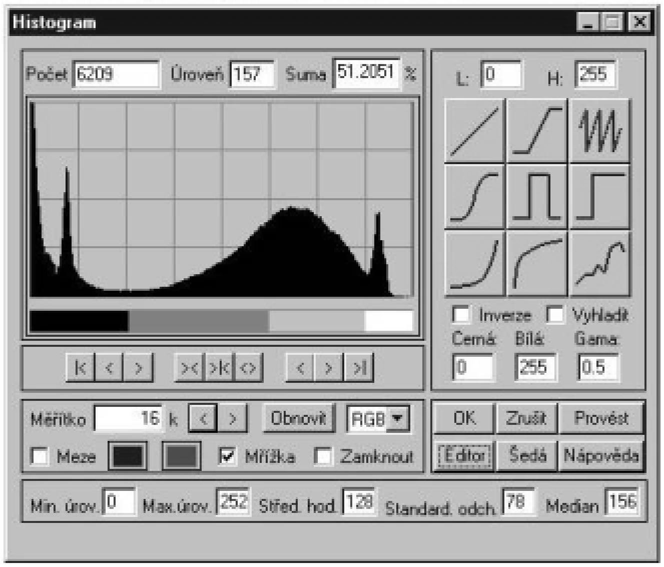
In the particular case it seems like the best solution to remove outliers and further work with the remaining data set. However, it is difficult to correctly determine that the exclusion of such data is correct step.

Often used Mahalanobis distance resp. another type of distance, can in this case serve as a measure of normality that determines the power to contribute and to enhance the quality of classification.

Except the vector of mean value μ we take into account also the variation that may not be the same in all dimensions of feature space. It is necessary to modify the evaluation metric distance of vector x from the centers of gravity

$$D^2 = (x - \mu)^T \sum^{-1} (x - \mu), \qquad (6)$$

where ∑ is dxd covariance matrix that represents the variance. It is positive-definite matrix. If we substitute for all ∑ the unit matrix we obtain the Euclidean distance.



Histogram

Počet 6209 Úroveň 157 Suma 51.2051 %

L: 0 H: 255

☐ Inverze ☐ Vyhladit

Černá 0 Bílá 255 Gama 0.5

Měřítko 16 k

RGB

☐ Meze ☑ Mřížka ☐ Zamknout

Min. úrov. 0 Max.úrov. 252 Střed. hod. 128 Standard. odch. 78 Median 156

Fig. 9. Histogram can be used to compare the relations of mean value and median and analyse the residuals.

E. Discussion

In fact the relationship between GIS objects and image analysis data provides a great flexibility to manipulate with image analysis results, optimize the currently used approaches to build and update GIS databases.

The key component of the method is the automated sample selection refined by a statistical outlier [15, 16]. Each class is cleaned by a non-parametric elimination procedure to keep only the most relevant training samples.

A preliminary image labelling is performed based on the intersection between a geographic vector database and the result of image segmentation. Each class is then cleaned by a non-parametric elimination procedure that iteratively excludes outliers in order to keep only the most relevant training samples.

Eventually, each image-object is classified based on a maximum likelihood and new classes are created by decision rules.

V. CONCLUSION

Time and space are the main coordinates of our living style and the base for spatial and spatial-temporal modelling.

The gathering capabilities of satellite data including high resolution data are in present time very well established for regular continuous temporal analysis, monitoring of changes and management of the major changes and natural resources.

Remote sensing data is used for various purposes in different applications and it is irreplaceable for dynamic analysis.

It is clear that the use of high resolution satellite images is essential, especially when we need distinguish detailed features. Using together the information gained from remotely sensed images with vector database to update the GIS database provide more advantages.

Many of the statistics are quickly extracted from the GIS database for analysis. Moreover, the ability to overlay many features, measure distances, and calculate requested parameters is significant support of data analysis. High correctness of GIS data is of highest importance.

Manual updates are very time consuming and therefore the support through automatic systems is required. Integration and comparison with existing vector data structures is the most important issue. Government, state administration, industries, companies never considered GIS and satellite images to be the source of prosperity and efficiency.

REFERENCES

[1] P. Helmholz, F. Rottensteiner, 2009. Automatic verification of agricultural areas using IKONOS satellite images. *IntArchPhRS* XXXVI - 1-4-7 (on CD-ROM).

[2] S. Borza, C. Simion, I. Bandrea, GIS application with geospatial database for improving the waste management in Sibiu surrounding area, *NOUN International Journal of Energy and Environment* , Issue 5, Volume 5, 2011, pp. 653-660.

[3] M. D. Ashraf, Yasushi, Y., 2008. Remote sensing and GIS for mapping and monitoring the effect of land Use/cover change on flooding in Greater Dhaka of Bangladesh. Department of Earth and Environmental Sciences, Nagoya University, Japan.

[4] P.Y. Hung, S.M. Shen, and Chen, P.H. 2006 "Interpreting shorelines on the large-scale orthogonals, a case study of the sand-gravel beaches, Taitung", Journal of Geographical Research, Vol. 44, pp.89-105.

[5] A. Muslim, G.M. Foody and P.M. Atkinson, 2007, "Shoreline mapping from coarse-spatial resolution remote sensing imagery of Seberang Takir, Malaysia", Journal of Coastal Research, Vol. 23, No. 6, pp.1399-1408.

[6] D. Klimešová , E. Ocelíková, GIS and image processing, International *NOUN Journal of Mathematical Models and Methods in Applied Sciences*, 2011, 5 (5), pp. 915-922.

[7] F. Shen, Y.X. Zhou and J. Zhang, 2006, "Remote sensing analysis on spatial - temporal variation in vegetation on Jiuduansha wetland", aceanlogia et Limnologia Sinica, Vol. 37, No. 6, pp.498-504.

[8] S. Jain, Technical note: Use of IKONOS satellite data to identify informal settlements in Dehradun, India. *Int. J. Remote Sensing*, 2007, **28**, 3227–2333.

[9] H. Yamano, H. Shimazaki and T. Matsunaga, 2006, "Evaluation of various satellite sensors for waterline extraction in a coral reef environment: Majuro Atoll, Marshall Islands", Geomorphology, Vol. 82, pp.398-411.

[10] Yang Zhang: Environmental monitoring of spatial-temporal changes using remote sensing and GIS techniques in the Abandoned Yellow River Delta coast, China, IJEP,Vol. 45, No. 4, pp 327-341.

[11] S. Arnold, 2009: Digital landscape model DLM-DE – Deriving land cover information by integration of topographic reference data with remote sensing data. *IntArchPhRS (38)*, Part 1-4- 7/WS, Hannover, (on CD-ROM).

[12] D. Klimešová, E. Ocelíková, Study on Context Understanding, Knowledge Transformation and Decision Support Systems. *WSEAS Transactions on Information Science and Applications,* 7 (7) 2010, pp. 985-994.

[13] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proceedings of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C. USA, pages 29–38, 2003.

[14] C.V. Stewart, Robust parameter estimation in computer vision, SIAM Review, 41(3):513–537, 1999.

[15] P. M. Atkinson, C. Jeganathan, J. Dash, Analysing the effect of different geo-computational techniques on estimating phenology in India," 10th International Conference on Geo-Computation, University of New South Wales, Sydney, Australia, 2009.

[16] Ch. Chalkias, A. Faka , GIS-based spatiotemporal analysis of the exposure to direct sun light on rural highways, *WSEAS Transactions on Information Science and Applications,* Issue 1, Volume 7, January 2010.

[17] D.Klimešová, E. Ocelíková, Spatial-temporal modeling and visualisation, NOUN *International Journal of Mathematical Models and Methods in Applied Sciences* , 2012, 6 (1) , pp. 149-156.

[18] P. Sun, S. Chawla, and B. Arunasalam. Outlier detection in sequential databases. In Proceedings of SIAM International Conference on Data Mining, 2006.

[19] Ellis, E.; Robert P. and Cutler J. (2009). "Land-cover". In: Encyclopedia of Earth. Eds. Cleveland (Washington, D.C.: Environmental Information Coalition, National Council for Science and the Environment). First published in the Encyclopedia of Earth April 14, 2009. *http://www.eoearth.org/article/Land-cover*.

[20] G. Fernandez-Auilés, J. M. Montero, J. Matern, Spatio-temporal modelling of carbon monoxide. Are ecologists right? Proc. of WSEAS DEEE'10, Puerto De La Cruz, Tenerife.

[21] D.Klimešová, E. Ocelíková, Study of Uncertainty and Contextual Modelling. *WSEAS International Journal of Circuits, Systems and Signal Processing,* Issue 1, Volume 1, 2007, pp. 12-15.

[22] S. Borza, C. Simion, I. Bandrea, GIS application with geospatial database for improving the waste management in Sibiu surrounding area, *NOUN International Journal of Energy and Environment ,* Issue 5, Volume 5, 2011, pp. 653-660.