Optimised Pattern Size for land cover–land use information conversion

Györk Fülőp*

Department of Mathematics and Informatics, Faculty of Horticultural Sciences, Corvinus University of Budapest, Doctoral School of Landscape Architecture and Landscape Ecology, Villányi út 29-43, 1118, Budapest, Hungary
*Corresponding author, e-mail address: gyork.fulop@uni-corvinus.hu

Abstract
Can we keep a lonely tree in the middle of a plowed parcel to be a forest? Or a glade, opening in the place of a fallen tree, to be a meadow? Surely not. In this case emerges the question, where starts the forest, and where the meadow? OPS (Optimised Pattern Size) represents the unit size of a land cover, with which it can be mapped optimally, based on the classification of MS VHR satellite images. By the described case study, I demonstrate the OPS computation method, which can be totally automated and integrated into recent classification programs, in order to reduce land cover–land use conversion problem onto the level of semantics, and to let regional and local authorities to be supplied with timely land use information.

Keywords: Optimised Pattern Sizes (OPS), land cover, land use, control plan, MS VHR satellite images.

Introduction
Land cover map shows what is there. Land use map of the same area shows what it is used for. Structural and master plans of regional and local authorities shows what the area should be used for, and how. However, the first two maps are in strong relation with each other (giving a good basis for the preparation of the last two ones), automated conversion of them is still not solved. Land cover information today is just an input for a human interpreter, who is able to integrate additional information as well, producing real land use map, the link between of them is not defined properly. Besides sub-pixel data [Gianinetto et al, 2010] need, between-pixel relations must be either taken into consideration. Causes, from which this conversion problem derives, can be grouped into two main classes: semantics and technical causes.

In this paper, first, I will introduce the practical circumstances of an application definition that brought on the need of Optimised Pattern Size (OPS), a new indicator in land cover–land use...
use conversion. Then I will describe momentarily the semantical and technical bottleneck factors of land cover–land use conversion. After that I will provide a definition and short description of OPS, and finally a modelling case study of OPS computation from setup of analysis through modelling experiments to presentation of results of computation. The case study aims to introduce OPS computation in its process with the main factors of the Optimised Pattern Size.

**Problem emerges in practice – The need for OPS**

The need for OPS (Optimised Pattern Size) emerged during the validation phase of a KEO (Knowledge-centred Earth Observation) application definition in 2011. The development of OPS aims to answer one of the two challenges that rose due to the validation process, and provides an iterative feedback in the application definition process (Fig. 1).

![Figure 1 - Application definition process in KEO system, OPS as iterative feedback.](image)

The Knowledge-centred Earth Observation (KEO) is the largest, public modular and scalable component based processing environment of our world. The admin of the system is ESA (European Space Agency). The KEO system is the tool of decentralised research and development in the field of Earth Observation in Europe.

Within the confines of an ESA PECS (Plan for European Cooperating States) project a new application was defined in KEO: Automated Knowledge-centred Assessment of Tourism Adequacy (AKATA) [Fülöp, 2011]. The aim of the application development was dual. First, we would like to provide the local decision makers from the level of LAU-2 (the smallest Local Administrative Unit of the EU: settlement) to NUTS-3 (3rd level of Nomenclature of Territorial Units of Statistics: counties) with a free access tool which can be a real support of the territorial decision making in the field of territorial touristic developments. It is very common nowadays that every territorial strategies, programmes and plans puts great emphasis on touristic development, but usually the well reasoned spatial references are missing from these documentations. AKATA aims to fill in this territorial planning gap.
Secondly, during the application definition we automated an analogue territorial assessment, published by Hans Kiemstedt in 1967 [Kiemstedt, 1967]. The original Kiemstedt-model measured landscape variability (which is in a very strong correlation with territorial tourism adequacy of the landscape) on 1:25 000 scaled topographic maps in 1 km² tiles with manual analysis. Due to the defined KEO application, nowadays, the assessment is based on VHR MS (timely) data and it is fully automated – only one click is enough to gain the 1 km raster tourism adequacy output from the input VHR image (Fig. 2). The metadata and the reference dataset of the application definition can be found at GeoNetwork [Geonetwork 1].

During the validation phase [Fülöp and Szikszainé Szigeti, 2011] exciting results were gained. The validation consisted of the comparison of the output test results of the application (Fig. 2) and the validation results, extracted from airborne DOP (digital orthophoto) images of a smaller plot of the same test area by manual interpretation (Fig. 3). While the whole application fulfilled the requirements of decision support, two main discrepancies emerged:

I. The final re-categorization of the 1 km² tiles was executed by simple interval
division, thus the final results do not follow the tendencies of background data (if there is an outlier tile, it affects the whole assessment of the territory).

II. As it in the correlation table can be seen (Tab. 1) land use factor (describing area) of test and validation datasets had a quite good correlation, but the forest-edge factor (describing perimeter) could be characterized only with poor values. This effect derives from the fact that the automated application made too much “speckles” while trying to extract land use information from the VHR image, while the human interpreter did not made the mistake of keeping a single tree to be forest. That is why the assumed areas of land uses correlated well, while assumed perimeters differed from each other in the test and validation results.

| Table 1 - Correlation matrix of KEO application validation process. |
|---------------------------------------------------------------|
| **Correlation matrix of factors of test and validation data** |
|                  | test Vsz | test Esz | test R | test T |
|------------------|----------|----------|--------|--------|
| valid Vsz        | corr.    | 0.843    |        |        |
|                  | SL       | 7.557E-10|        |        |
| valid Esz        | corr.    |          | 0.249  |        |
|                  | SL       |          | 0.163  |        |
| valid R          | corr.    |          | 1      |        |
|                  | SL       |          | 0      |        |
| valid T          | corr.    |          |        | 0.766  |
|                  | SL       |          |        | 2.048E-7|

As it can be felt, the second discrepancy rose the need for OPS (Optimised Pattern Size). The validation process showed that the extraction pixel size of a land cover should be optimised in order to preserve the coherence of the extracted information, thus to support the land cover–land use conversion on a higher standard. However, as it has already been mentioned, land cover–land use conversion is bottlenecked by two factors: semantical and technical drawback factors.

**Semantical and technical drawback factors**

Before introducing semantical and technical drawback factors of land cover–land use conversion, general practical impacts of them must be illustrated first. “Landscape is a cultural construction expressing itself in images, associations and imaginations” [Strohmeier, 2007]. Land use is visualized by these images, for instance by satellite images, providing a complex data source about the environment and about the use of it: the landscape itself. In the case of obstructed (automated) extraction of land use information from VHR MS satellite images (recently the most available raw data sources), maps and plans from higher level (master plans of local authorities, management plans of Natura 2000 sites, strategic development plans of regions etc.) will have an implicit lack of objectivity and/or timeliness. As a result, from the side of regional and spatial planning and development, functional and administrative regions may differ from each other because of inaccurate land use mapping. As some of these plans have effect on definition of rights (building activities, agricultural management etc.) quality of land use mapping has a direct impact on the land use and even the land user itself. Semantic problems emerge mainly in the decision making process. For instance, how should we decide about forested pastures? That is a rare, but very valuable land use in Central-
Europe and can be imagined as a pasture with grown up trees standing on it homogenously but loose (earlier these trees were supposed to give shadow for the flock or cattle at noon). Who can decide it to be a pasture, or a forest, or a new entity – a forested pasture – if not a human person? These semantic questions will have to be answered in the future, however great research process had already started on the level of semantics [Durbha et al., 2009; Barb and Shyu, 2007]. Now, I am dealing with the technical part.

By using any (even partially) automated land use classifier [Persello, 2010], we will find the results to be affected by some discrepancies. The best example for this phenomenon is the lonely tree, standing in the middle of the arable land, kept to be a forest. Raw land use maps, extracted from land cover information, contain usually too many “speckles”. These speckles are removed by generalization process using between-layer operators. This practice resulted the GIS tradition that land cover is raster based, while land use map is produced in a vector model, coupling land cover–land use conversion with raster–vector conversion. However, as we know it from experiences of translator program development, the number of translations (or changing of describing model) affects significantly the contained information quality. Measuring this aspect, it is considerable to develop land cover–land use conversion within the confines of raster models by reducing the role of generalization of land use patches.

OPS aims to help land cover–land use conversion within the confines of raster models. With the use of OPS, creation of speckles in patches can be avoided (more precisely: optimised in quantity and quality) already in the phase of extraction of land cover information from VHR MS images. By this process we can gain a special land cover map, a “pre-land use map” that still not contains the semantics which are the specificity of land use maps. The whole land use extraction process is still under research, in the followings I would like to introduce the computation method of OPS by a modelling case study as a first milestone.

**Optimised Pattern Size (OPS)**

Definition: OPS, Optimised Pattern Size is a remote sensing data-analysis indicator, showing which output pixel size is optimal for the extraction of the targeted land cover, measured in the dimension of the input image pixel size.

Practically: OPS shows the optimised (geometrical) generalisation level during extraction/categorization of a land-cover (land cover of interest) described on a VHR MS remote sensing (RS) image. It is essential to understand, that OPS (being an objective indicator) is derived directly from the RS data itself!

Each land cover has its own OPS value that refers to how small the land cover patch can be, and how big part of the surface can already have the properties of the required land cover. The unit of OPS is the consisted number of pixels of the input VHR MS image. It must be emphasised, that OPS means the pixel size for extraction, and not for the map visualization – as extraction pixels may overlap on land cover map. OPS can be calculated from VHR MS data sets at first, as spatial resolution of HR images already does not allow statistically adequate sampling pixel number.

**Set up of analysis**

In the modelling experiments (described in the following two chapters) I used QuickBird MS VHR image (geometric resolution: 0,67 m, pansharpened, RGBN) from Saarland (Germany). I chose a homogeneous (mature Pinus nigra forest) (Fig. 4) and a heterogeneous (young Carpinus
betulus growing-up) land cover (Fig. 5) and delineated 5-5 training samples areas, each measured 75 times 75 pixels. Empirically these training samples resulted good land cover categorisation.

I cropped the images with the AOIs of training samples, and exported the numeric pixel values of each bands – generating pixel value matrixes. After this, I used PASW Statistics 18 software for producing describing statistics, probes and multivariable analysis.

The analysis aimed to define OPS values of the two land covers: the homogeneous and the heterogeneous one. These two types were chosen to demonstrate the effect of land cover heterogeneity on OPS values. This consideration was necessary because the definition of the OPS values was assigned by two factors: size driven factor and heterogeneity driven factor.
Possible Pattern Size (size driven factor)
I began the analysis with assigning a new code to the pixels of the input image, thus segmenting the matrixes into patterns from 3 to 37 pixel size in sequence (Fig. 6). My computation syntax of the new code is detailed in [1].

\[
\text{PaCo} = \text{TRUNC}((R \ast P + C + P - 1)/P, 1) \times \text{TRUNC}((R + P - 1)/P, 1) \ast P - 1 \quad (1)
\]
\[
+ \text{TRUNC}((R + P - 1)/P, 1) \ast 10000
\]

where: \( \text{PaCo} \) is the new code of pixels – pattern code; \( R \) is the number of row in which the pixel can be found; \( C \) is the number of column in which the pixel can be found; \( P \) is the pattern size into which we would like to segment the pixel value matrixes of the training sample areas (3, 4, 5, ..., 35, 36, 37); \( \text{TRUNC}(x, y) \) is a function that returns the value of first argument truncated towards 0, and where second argument specifies that the result is an integer multiple of this value – for example \( \text{TRUNC}(6.564, 0.1) = 6.5 \).

This way I gained a sequence of segmentation of the same 75 by 75 pixel sized training samples from 3 by 3 pixel sized patterns to 37 by 37 pixel sized patterns (possible pattern sizes – PPS). These sequences helped in the demonstration of size driven factor of OPS (“\( a \)” on Figure 7).
Figure 7 - Method of seeking OPS after computing regression of information loss in relation of pixel size.

Sequence of experiments (heterogeneity driven factor)

It has to be remembered, that training samples are used to gain spectral information which will describe the land cover during the classification of the whole image. By using pixel value matrices, I compared multivariate (RGBN) expected pixel values of training sample areas by ANOVA analysis (expected values were equal on the significance level of 2.5E-11 for homogeneous and 2.8E-10 for heterogeneous values). I produced after ANOVA Post Hoc tests (Scheffe) for homo- and heterogeneous samples separately. According to this information I defined the two most similar homogeneous samples (no. 3 and 5).

To demonstrate the heterogeneity driven factor of OPS I set up a sequence of 5 experiments:

I. Experiment: I assessed with the use of homogeneous training sample area no. 3, how the expected pixel values of sequenced possible pattern sizes (PPS) approached the expected pixel value of the whole sample area no. 3.

II. Experiment: I assessed with the use of homogeneous training samples area no. 3 and 5 how expected pixel values of sequenced PPS approached the common expected pixel value of the two most similar sample areas.

III. Experiment: I assessed with the use of homogeneous training samples area no. 3 how expected pixel values of sequenced PPS approached the common expected pixel value of homogeneous sample areas.
IV. Experiment: I assessed with the use of heterogeneous training sample area no. 3 how the expected pixel values of sequenced PPS approached the expected pixel value of the whole sample area no 3.

V. Experiment: I assessed with the use of heterogeneous training samples area no. 3 how expected pixel values of sequenced PPS approached the common expected pixel value of heterogeneous sample areas.

Experiment results target the characterization of OPS in relation with heterogeneity of training sample areas as average difference between expected values of training samples rises from experiment I. to experiment V (heterogeneity driven factor of OPS – “b” on Figure 7).

Results

The analysis aimed the characterisation of two factors influencing OPS. As a method I chose to measure the difference between the expected pixel values of possible pattern size segments (e) and training samples (E) (Fig. 6). For the measure of distance I used RMSE (root mean square error) and MaxSE (maximum square error) both. For the computation of OPS I draw up two relations: 1) pixel size–RMSE/MaxSE, which aims to describe the loss of the spectral information in PPS segments, which is contained by the training samples; 2) pixel size–pixel area, which aims to describe the land cover specific geometrical resolution of information extraction (Fig. 7).

I built regression models (Tab. 2) from the experimental RMSE/MaxSE values to get descriptive curves for information loss in each experiment. As it was expected the strongest regression could be built with the use of an inverse model (worst R-square was 0.78). It is obvious that additional constant cannot participate in the model, the asymptote must be at y=0, as if we would continue the segmentation over pixel size 37 we would get nearer and nearer to 75, which is the size of the training sample – that means 0 information loss (measured to the categorization); and if we would not use any pixel values (0 pixel possible pattern size) – we would lose infinite (all) information.

| Table 2 - Regression models of information loss in relation with pixel size. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Measure of distance | R square | b1 coeff. | Modell |
| EXP I. | RMSE | 0.994 | 903.366 |
| | MaxSE | 0.907 | 7809.024 |
| EXP II. | RMSE | 0.994 | 903.222 |
| | MaxSE | 0.906 | 7932.516 |
| EXP III. | RMSE | 0.994 | 903.131 |
| | MaxSE | 0.902 | 8461.064 |
| EXP IV. | RMSE | 0.900 | 1785.878 |
| | MaxSE | 0.798 | 16394.965 |
| EXP V. | RMSE | 0.895 | 1782.832 |
| | MaxSE | 0.778 | 19299.625 |
It means that the magnitude of coefficients measures size driven factor. For the extraction of heterogeneity driven factor the quotient of MaxSE and RMSE values would provide a solution. The sequence of these quotients (in relation with the number of experiments) can describe how the heterogeneity of the training samples may influence OPS – however now, with the lack of entropy values this curve could not be estimated in the lack of reference values.

Figure 8 - OPS values in the cross of information loss and pixel area curves – using RMSE values in distance measurement. Blue: homogeneous, green: heterogeneous land cover.

Results of the case study demonstrate the meaning of OPS as well. By measuring information loss with RMSE homogeneous land cover (mature Pinus nigra forest) can be classified with 9 OPS value. It means that the smallest pattern size during the classification will be 9 times 9 pixel huge – on the surface that means 36 sqms (Fig. 8). The young Carpinus betulus heterogeneous land cover has a 12 OPS value (64 sqms on the surface). MaxSE distance measurements resulted much more strict values: 20 OPS for homogeneous, 27 OPS for heterogeneous land cover (Fig. 9).
Figure 9 - OPS values in the cross of information loss and pixel area curves – using MaxSE values in distance measurement. Blue: homogeneous, green: heterogeneous land cover.

Summary
OPS, Optimised Pattern Size is a remote sensing data-analysis indicator, showing which output pixel size is optimal for the extraction of the targeted land cover, measured in the dimension of the input image pixel size. It is needed to ease the land cover–land use conversion in the phase of land cover extraction from VHR MS images.

The demonstrated OPS computation can be totally automated and built in any classifier programme. OPS can be a basis for a new classifying methodology of VHR MS images, which aims the elimination of raster-vector conversion in land cover–land use translation. OPS can be computed, first of all, from VHR MS images, as these datasets provide statistically enough pixel value, and have adequate spatial resolution. Thus, OPS can easily integrated into nowadays practice, allowing more self-conscious land cover information extraction.

OPS values describe the land cover during classification process, preventing creation of land cover patches that have no sense. Each land cover has its own OPS value, which derives only from its spectral characteristics, presented in a RS dataset. This value means that due to the spectral specificities of the land cover, how many pixels of the origin image...
should be considered to represent the smallest patch of the subjected land cover. However, it must be emphasised, that OPS is not the same as the visualization patch size of the output land-cover image, as OPS patches are to be used during the extraction process, and may overlap in the output thematic map.

Outlook
The optimal one from possible pattern sizes (PPS) could be chosen by the respect of two factors – size driven factor and heterogeneity driven factor. In my next research I will quantify heterogeneity driven factor by the use of SAR datasets. At the same time I will characterize the semantic difference of measuring spectral information loss with RMSE or MaxSE. By the use of OPS, “pre-land use maps” can be extracted from VHR MS images. These “pre-land use maps” are products that still lack the semantic solution of land cover–land use conversion, but can be extracted almost automatically, making a step towards real decision supporting services of EO and RS.

In the near future OPS extraction module will be integrated into Automated Knowledge-centred Assessment of Tourism Adequac KEO application (Fig. 1). The vision of only-raster based land cover–land use conversion means that support vector machines (SVM) [Bruzzone et al., 2009] might be used after and not before conversion of information. This way soundness of information might be kept, which is a very important factor in the monitoring of land use changes, and in extracting timely land use information for regional and local authorities to prepare territorial concepts, strategies, programmes, structural and master spatial plans.

References
Barb A. S., Shyu C.-R. (2007) - User-specific semantics for modelling content-based information in geospatial knowledge. IEEE International Geoscience and Remote Sensing Symposium, pp. 330-333. doi: http://dx.doi.org/10.1109/IGARSS.2007.4422797.

Bruzzone L., Marconcini M. (2009) - Toward an Automatic Updating of Land-Cover Maps by a Domain Adapatation SVM Classifier and a Circular Validation Strategy. IEEE Transactions on Geoscience and Remote Sensing, 47 (4): 1108-1122. doi: http://dx.doi.org/10.1109/TGRS.2008.2007741.

Durbha S. S., King R., Shah V. P., Younan N. H. (2009) - A framework for Semantic Reconciliation of disparate Earth Observation thematic data. Computers and Geosciences Elsevier, 35 (4): 761-773.

Fülöp Gy. (2011) - Opening for people - Automated landscape variability assessment in KEO system. VII. Conference on Image Information Mining: Geospatial Intelligence from Earth Observation, Ispra, Italy, April 2011 ISBN 978-92-79-19708-6 doi:10.2788/69291.

Fülöp Gy., Szikszainé Szigeti I. (2011) - Automatikus raszteres adatbányászat eredményének ellenőrzése VHR űrfelvételen többváltozós statisztikai eszközökkel. IX. Magyar Biometriai, Biomatematikai és Bioinformatikai Konferencia, Budapest, Hungary June 2011. pp. 57.

Kiemstedt H. (1967) - Zur Bewertung der Landschaft für die Erholung. Beiträge zur Landespflege, Stuttgart, Sonderheft 1.

Persello C. (2010) - Advanced Techniques for the Classification of Very High Resolution and
Hyperspectral Remote Sensing Images. PhD thesis, University of Trento, March 2010.
Strohmeier G. (2007) - Werkmaterialen zur Landschaftswahrnehmung. Österreichisches Portal zur Umweltbildung und nachhaltigen Entwicklung. URL: http://www.umweltbildung.at/LBL/wahrnehmung/hintergrund/index.htm (last accessed: 12.02.2008).
Gianinetto M., Rota Nodari F., Maianti P., Tortini R., Lechi G. (2010) - Optimal spectral band configuration for forest land-cover classification of hyperspectral data: a study for the Italian-Canadian Joint Hyperspectral Mission. Italian Journal of Remote Sensing, 42 (1): 3-12. doi: http://dx.doi.org/10.5721/ItJRS20104211.
Geonetwork 1 - http://geonetwork.dimeola.esrin.esa.int/geonetwork (last accessed: 25.01.2012).

Received 12/09/2011, accepted 09/05/2012

© 2012 by the authors; licensee Italian Society of Remote Sensing (AIT). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).