AUTOMATIC CAL/VAL METHODS LOWERING PRODUCTION EFFORT OF UPDATING LARGE CORE SPATIAL DATA

A. Peled

Department of Geography and Environmental Studies, Canter for Spatial Information Research, University of Haifa, Israel
peled@geo.haifa.ac.il

Commission III/7

KEY WORDS: Supervised Classification, Data-Base Driven, Autocalibration, Type-classes, Automatic change detection, Ad-hoc type-class identification, Automatic Updating, Big core spatial Data Bases, Munchhausen method

ABSTRACT:

There are basically two levels of calibrations and validation of digitally acquired spectral and other information via sensors carried on space-borne or airborne platforms. The basic level is carried out by the data producers executed by comparison made of results taken over test fields for example. The second level, more a part of a supervised classification effort are carried by the data users and value added spatial information users or providers to edge users. The latter is quite typical for supervised classification protocols. This is either for establishing libraries of spectral signatures for each relevant class-type or for ad-hoc classification where no previous information or specific knowledge wee kept. Such methods indicate and support even strongly the need of the basic Cal/Val step of the sensors made by the original data providers. The paper is reviewing the method of database-driven concept that allows for automatic recognition of detected features within the digital spatial 2-D (yet) realm to its identification within the digital 2.5D spatial vector information within existing large Big-data national core spatial data bases to be updated. These Large data bases are Big enough to operate the resourceful Munchhausen method of self-pulling information out of the huge abandon of data resources.

1. INTRODUCTION

Optimal and meaningful Land management and other geo-spatial related decision making, relies on the availability of a various sources of current information sources. The availability of accurate, up-to-date databases that digitally represent the full actual reality is of essence. Thus, the maintenance of such databases turned to be one of the main tasks of any developed or developing country. Furthermore, this was one of the very first manifestations of the than new thought of e-Gov. The last two or three decades brought with them a global growth that even with some financial drawbacks every 10 years or so, is still amazing in its capacity. Furthermore, the growing demand for information of all sorts including the use of many applications based on geospatial information, brought with themselves a growing need of fast methods and protocols for updating huge spatial information databases throughout the world. The worldwide rapid civil infrastructure development we witness in these passing decades is influence the integrity of our stored data. This as data stored in Geospatial databases are usually subjected to an intensive change processes that diminish their relevance. Also, as some traditional definitions of type-classes used when establishing these information sets were (and still are!) based on semantic descriptions, these data bases include different types of discrepant information. In order for a National Geospatial database to be used for sustainable decision making on public and private levels, the thematic land-use data need to be consistent and prompt.

Remote sensing is the primary source for many types of thematic data critical for GIS analyses, including information on land use, land-cover characteristics and surface elevation. Compared to more traditional mapping approaches such as terrestrial surveys and basic interpretation of aerial photographs, the integration of spatial information from GIS databases with remotely sensed spectral data has the advantages of low cost, area-wide coverage, and easily effected frequent updating. Consequently, additional geo-spatial information products have become an essential tool in many operational programs used in land-resource management. Nowadays these considerations are relevant even more also when using drones and other UAVs as the platforms of very small and light sensors to detect and identify changes in local and very small projects, in term of area. The use of a large variety of sensors, digital cameras and other spectral and non-spectral apparatuses for data acquisition has even a stronger impact of the question at hand on how to deal with Calibration and Validation of the integrity of such sources and of the data they produce. This step in the highest level of Cal/Val efforts is important to ensure consistency or to detect anomalies and non-conformity in order to establish the consistency of the data provided. This level is not discussed here at all. Yet, it is postulated here that such step is important only when a set or an array of sensors is used to capture the new digital description of Earth or any set of objects monitored by the data source.

1.1 The Munchhausen approach

The GIS-driven approach [Peled, 1993; 1994] was proposed as an automatically supervised classification methodology to the Survey of Israel (SOI) some 25 years ago [Peled & Haj-Yehia, 1998]. It is based on training the system, in the spectral characteristics of Geo-spatial objects and phenomena by superimposing the GIS polygons on the newly acquired remotely sensed image. This approach has similarities to training human interpreters to detect of unfamiliar yet known objects.

The basic idea here was that any existing geospatial database will carry enough "samples" of major, if not all, of type-classes as saved within the very same database to be inspected or updated. At the time of the very first proof of concept test there were about 1,200,000 buildings of sort, in Israel. At the very

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-W10-167-2020 | © Authors 2020. CC BY 4.0 License.

same time, (1991), the UK national geospatial database maintained by the Ordnance Survey, held about 30,000,000 such polygons. In both instances, even in the smaller Israel, there were enough polygons that could serve in order to investigate how such Buildings polygons (perimeters of construction) may be characterized by a spectral radiometric signature of each band for all bands. That is translated to the possibility of automatically establish a radio9metric signature for any class-type for any band of any sensor. That is, no prior external knowledge or supervised information is needed for the new classification effort. This is the "old Munchhausen case" where the old database itself is the source of the new updating calibration. It is further postulate that even the atmospheric and other meteorological conditions or time of the data collection are of essence. The automatic calibration as a replacement of the supervised learning step are done regardless of all these parameters. As there is no need for an image to image comparison, there is no need to ensure images taken in two different epochs will be converted as if taken in standard conditions. There is no need for standard conditions. There are no standards conditions whatsoever.

In the first major effort of proof of concept, used RGB bands of colour air-photographs as the newly acquired source. Originally it was done using only B/W air-photographs, but success was meaningful statistically for the Transportation network and some moderate success was found with buildings. Yet, the understanding that the SOI will use eventually coloured air photography expedited the shift of the research and development from B/W to Colour. These experiments were carried on several geographical zones of Israel. It was based on using three bands namely RGB and applying the method to 4-6 representative class-types.
The training step was carried out separately on each band, generating radiometric signature for each polygon on the relevant type-class. In order to normalize the size of polygons the resulted counts of pixels within slices of 16 grey levels were kept in values representing the percentage relative to the size of the polygon. Thus, each polygon, or instance of the class-type could be compared to other such normalized radiometric signatures. For each section of 16 grey levels out of 256 (a total of 16 such aggregated grey level groups an average and a set of upper and lower values were set (see fig. 1), that depict such set of triplet signatures within one class-type. Polygons which its signature were out of the range for a specific radiometric set, were marked. A pre-process cut-off decision of the amount of allowed such non-conformity instances used to accept or reject a specific polygon within the class-type. The result was a very accurate set of radiometric signatures for almost all class-type; the tagging of old polygons that had to be redefined due to changes or non-conformity found within their radiometric signatures; and (3) a set of new polygons, found within the newly acquired data that needed to go through a classification process.

2. AUTOMATIC CLASSIFICATION AND UPDATING

2.1. GIS-driven analysis of land-use

The heterogeneity of land-use usually results in high spectral variation within the same class in satellite imagery. The straightforward pixel-wise solely spectral-based methods (e.g. Maximum likelihood) cannot overcome the high spectral variation of intensity within the same class and retain its spatial distribution. In order to address land-use classification, one critical issue is utilization of information inherited in the existing land-use classification. The GIS-driven approach offers for satellite image analysis the spatial information about land-use that captured in the thematic classification. This may reduce the within-class spectral variation and improve the spatial proximity of classification results (Peled, 1994, Peled& Haj-Yehia, 1998). In order to provide the analysis of the time series with the meaningful image objects, proposed is the utilization of land-use polygons from the existing spatial database land-use classification.

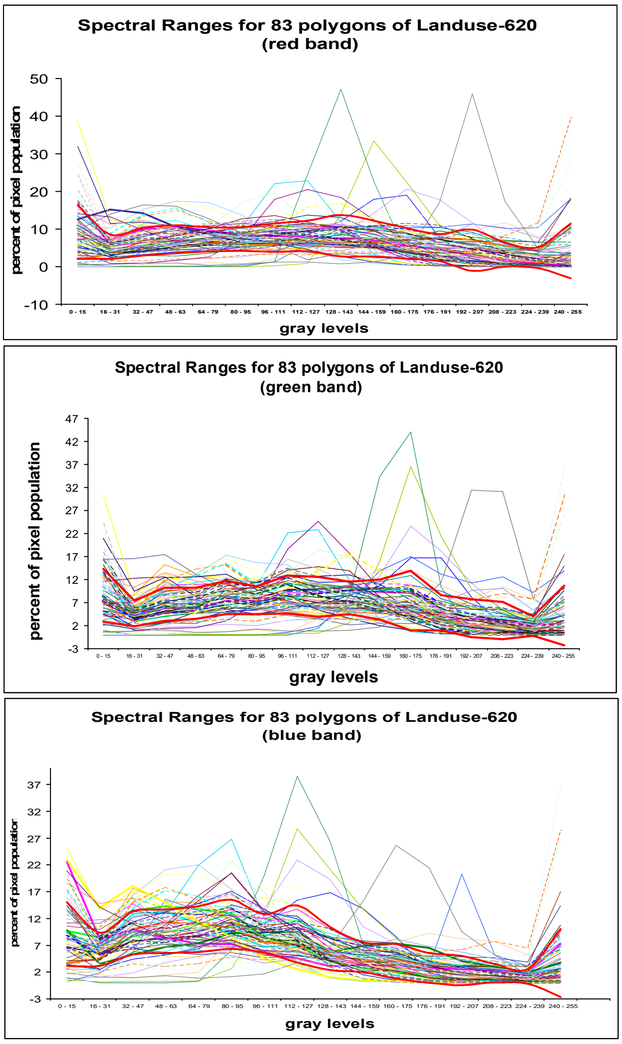


Figure 1. GIS-Driven spectral radiometric signature of RGB bands respectfully of the Forest Class-type code=620

The underlying original assumption was that the knowledge about land-use is that most classified polygons, correctly represent the land-use reality of the image. In such a case the discrepant polygons should be reclassified in respect to available land-use classes. In later versions, this concept was translated to an effective set of algorithms that governed a set of computer subroutines that automatically tagged all non-conformal polygons. This in itself provided a "cleaner" set for the automatic data-base driven classification step of all tagged polygons and of new entities detected within the realm of a newly provided spatial information spectral and other sources.

2.2 Iterative discriminant analysis

The iterative discriminant analysis (IDA) implemented in Exelis ENVI/IDL 5 was applied to determine the spectral bands which separate on statistically significant level between the land-use classes. The spectral characteristics for the existing classification were captured for each polygon as average intensity values and then used in IDA processing. The detailed description of IDA algorithm implemented in signal recognition is a common knowledge The method as adopted in the study generalizes standard linear discriminant analysis and attempts to use the spectral bands as independent variables to discriminate between the land-use classes through a series of iterations.

At the beginning of the IDA process, the discriminant functions are generated from polygons for which land-use classification is known when the class-types as were determined within the spatial database to be updated are accepted as initial classification. Then the functions applied iteratively to new classification cases with measurements on the same set of spectral variables. On each iteration step the combination of the statistically significant variables are determined and the classification is corrected as according to the posterior probabilities provided by current discriminant functions. The iterative process continues until no change is found between previous and current classification. This situation represents the best discrimination between the spectrally homogeneous land-use classes.

In order to define the spectral characteristics for the existing classification, the land-use classes were related to the reflectance. Then the stepwise linear discriminant analysis was used to determine the discriminant functions that distinguish the land-use classes on a statistically significant level (estimated with Wilk's lambda test). Discriminant function is a linear combination of the discriminating variables (spectral bands) given for each spectral band in form of following equation:

$$F_{km} = u_1 X_{1km} + u_2 X_{2km} + \dots + u_n X_{nkm} \qquad (1)$$

Where,

F_{km} = the score value for case m in land-use class k;

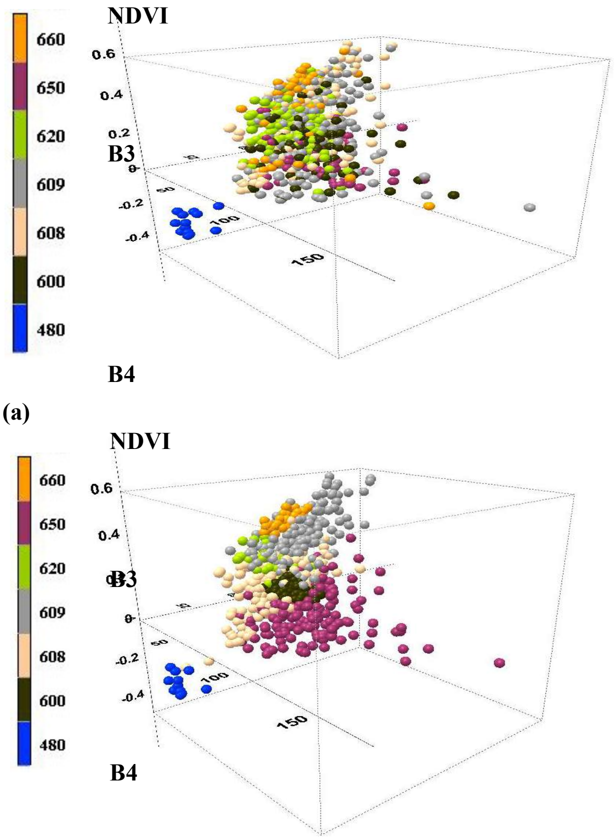
X_{nkm} = the value on discriminant variable Xn for case m in land-use class k; and

u_n = coefficients which produce the desired characteristics in the function.

The coefficients for the first discriminant function are derived so as to maximize the differences between the land-use class means. The coefficients for the second and following discriminant functions are derived to maximize the difference between the land-use class means, subject to the constraint that the values of the latter and former discriminant functions are not.

2.3 Detection of discrepancies

In our study the discrepancies are detected as according to the difference between the initial classification and the IDA classification, assuming that the best spectral discrimination between land-use classes represents the most correct land-use classification. Therefore, in naturally homogeneous land-use class, the amount of discrepant polygons is considered to be substantially less than the total number of polygons in the same class. The assumption was that the classes subjected to rapid land-use changes as "cultivated fields" or "residential areas" will comprise more discrepancies in comparing with classes of "water bodies" or "natural forests". The resulting IDA classification was compared to the initial land-use classification in form of error matrix. The contingency matrix has summarized the distribution of polygons between the initial and the IDA classifications. The polygons re-classified to a different class were assigned as discrepant. Polygons that remained in the same land-use class were assigned as consistent. The accuracy of the detection was assessed in terms of false and true discrepancy/consistency using the validation dataset.



(a)

Figure .2. Within-class distribution of land-use polygons in according to the selected spectral bands. (a) Initial classification as existing in the GIS database and (b) final IDA classification. The class-codes are displayed as according to the legend bar at the right and correspond to the land-use class.

2.4 Rule-Based System

The rule-based system which was developed includes sets of rules which supply a unique description for each type of objects. These sets of rules integrate radiometric, geometric, textural and topological parameters. The radiometric parameters include the distribution and other statistical parameters of the grey level values of each band for the pixels within each object (see figure 2). These parameters were computed only for objects within "no-change" regions. The geometric parameters include descriptors which define the geometrical characteristics of the object, such as area, perimeter, elongation, compactness, moments of inertia, etc. The textural parameters describe the textural template of grey level values for each band, such as: contrast and homogeneity. The topological parameters include topological and spatial relations between the objects from different types. These relations take into consideration instances such as if the object is within urban, rural, industrial, flat or mountainous zones.

2.5 Knowledge-based Classification

2.5.1 Texture properties

Image classification is the process of assigning a pixel (or homogenous groups of pixels) of remote sensing image to a land-use class. The objective is to classify each image pixel into only one class (crisp or hard classification) or to associate the pixel with many classes. Poor class definition, gradual transition zones or fuzzy boundaries, mixed pixels, and incomplete or imperfect data give rise to uncertainty in remotely sensed image classification results. Homogeneous regions with similar reflection can easily be identified as objects on a remote sensing image. The presence of texture makes it more complicated but carries a vast new field of describing of texture may be based on measurements which characterize a texture and provides a classification in terms of correct labelling of each set of measures. Texture measures can be extracted in spatial space or spatial frequency space. The extracting way depends on the specific application. The landscape image is complex and contains various land cover types. These land cover types have versatile shapes and appearances. Even the same kind of land cover type appears different. There are three main categories for texture identification: statistical, structural and model-based. Additional category is a signal processing and concerns to structural measures of texture. The most well known statistical approach to texture description is the grey level co-occurrence matrix. The co-occurrence matrix contains elements that are counts of the number of pixel pairs for specific brightness levels, when separated by some distance and at some relative inclination. A relatively new and model-based texture measure is the local binary pattern operator]. It is a theoretically simple yet efficient approach to grey scale and rotation invariant texture classification based on local binary patterns.

2.5.2 Decision trees

Decision tree is non parametric classifier. A decision tree classifier is a hierarchical structure where at each level a training set is applied to one or more attribute values that may have one of two outcomes. The result may be a leaf, which allocates a class, or a decision node, which specifies a further test on the attribute values and forms a branch or sub-tree of the tree. Classification is performed by moving down the tree until a leaf is reached. The advantages of decision tree classifier over traditional statistical classifier include its simplicity, ability to handle missing and noisy data, and non-parametric nature i.e., decision trees are not constrained by any lack of knowledge of the class distributions. The automatic construction of decision trees with CART (Classification and Regression Trees). The main difference between the different algorithms used, is the condition followed to carry out the partitions of training sets.
In presented work, used CART-like algorithms (QUEST and CRUISE) of automatic decision tree generation. Because of decision tree classifiers differ in the ways they partition the training sample into subsets and thus form sub-trees. That is, they differ in their criteria for evaluating splits into subsets. QUEST (stands for Quick, Unbiased and Efficient Statistical Tree) produces binary decision trees, where each node may only have two children (may only result in a 'yes' or a 'no'). CRUISE (stands for Classification Rule with Unbiased Interaction Selection and Estimation) produces decision trees that split each node into as many sub-nodes as the number of classes in the response variable where each pair of numbers represents a valid range. Implementation of these algorithms for ENVI RSI software provides non-parametrically determination of statistical relationships between given data layers in order to produce knowledge-based multistage classification by using a series of binary decisions to place pixels into classes.

3.3 Image classification

Once the classification rules are created using decision tree classifier, they can serve as a knowledge base. This knowledge base can be used for classification of the data. The beginning of image classification includes the modelling of local binary pattern in accordance to local neighbourhood of a grey scale image as the joint distribution of grey levels. At the same time, texture values are calculated by co-occurrence measures, according to grey-scale spatial dependence matrix. Then, the further steps of classification process include composition of synthetic images which contains stacked textural bands along with spectral data. These images serve as a basis for non-parametrically determination of statistical relationships between many data layers in order to create a binary decision tree. Training data were collected for each image epoch separately. The GIS-Driven training data was used for execution of both algorithms of decision tree classification to extract knowledge in the form of IF-THEN classification rules from the satellite and ancillary GIS data. Thus, decision tree generated using QUEST/CRUISE decision tree classifier was converted to classification rules and then, classification rules were used directly using knowledge base classifier to classify the input image.

2.4 Secondary objects classification

As the simple approaches for automatic yet fast updating methods were proving themselves some further upgrading were tested and introduced to the production software. These were among other, Image segmentations, changes detection and identification of whole original polygons; change detection and identification of partially changed original polygons and cluster segmentation to result with new polygons covering part of one or several original land use polygons. Another type of such enhancements was the secondary objects classification. In a 20 years old research effort on the various Neiborhood's within the jurisdiction of Haifa Municipality, one of the results was an algorithm and a method to define or describe the "basic type" of various Neiborhoods. This was based on a statistical evaluation of objects within the neighbourhoods. 5 major types were defined in the very beginning to serve for a statistically analyses of the percentage of 10 major class-types in terms of area. Thus, a rich suburb depicted higher rate of water (due to presence of swimming pools. The downtown region depicted a very low percentage of trees, or any "green" class-types. After the percentage ranges were defined for each of the 1o parametric class-types of the 5 representing types of neighbourhoods, the whole city went through a basic object classification and then this particular secondary classification. This was immensely helpful to the municipality efforts of re-zoning the city both in terms of taxing properties and also by using a more qualitative

evaluation as a basis for the rearranging the development plan for the city.
This successful research of a MA graduate student helped also, concept wise, in for the cluster segmentation mentioned above. Again here, there was no need for calibration. The existing objects, after the GIS-Driven first step of tagging non-conformal objects, served for the basic step of the secondary classification, namely cleaning the data set from erroneous objects, whether due to an error in the original interpretation or due to changes made in the civil development and build up/

3. SUMMARY AND CONCLUSIONS

The original method of GIS-driven updating of spatial databases, proposed by Peled 91993} in the early 1990s proved to be working even on B/W photographs by line following and skeletonizing the raster footprints of the transportation network. Since then simple pixel wise statistics within the perimeters of existing polygons, to be updated, within the old spatial databases proved even efficient enough for automatic 2.5-Dimensional updating of large core Big data national spatial data bases. Developments and additions of various methods such as IDA or simple textural conditions and other indices to expedite the automatic updating were adopted in order to facilitate for faster, more accurate and lowered the operational pressure of maps and digital spatial data bases. The major conclusion is that any sensor or newly acquired spatial sources whether spectral or other regardless of being carried by satellites, aircrafts drones or any type of UAVs, are optional for implementation. No prior knowledge in terms of spectral radiometric, phase or intensity are important. As long as the basic information to define each particle of the newly acquired data in space and sometime in time whether correlated or not, one will be able to verify whether the existing information is correct and to update new changes or correct past interpretation mistakes and errors.

REFERENCES

Peled, A., (1993) "Remote Sensing in Israel - From Change Detection to GIS Generation." International Symposium on Operationalization of Remote Sensing, Enschede, the Netherlands, 19-23 April 1993. Vol. 6 - Remote Sensing and GeoInformatics, pp. 117-126.

Peled, A. (1994) Revision of Digital Maps and GIS Databases. Symposium on: Mapping and Geographic Information Systems, Athens, Georgia, USA, In: ISPRS International Archives of Photogrammetry Vol. 30, Part 4, pp. 268-272.

Peled, A and Haj-Yehia, B. (1998) Toward automatic updating of the Israeli National GIS Phase II, D. Fritsch, M. Englich, M. Sester (Eds.), International Archives of Photogrammetry & Remote Sensing, Volume XXXII, Part 4.

Michalis, P., Dowman, I., 2008: A Generic Model for Along-Track Stereo Sensors Using Rigorous Orbit Mechanics. *Photogrammetric Engineering and Remote Sensing* 74(3), 303-309.
456-469.