Gimme my vegetation map in an hour! Towards operational vegetation classification and mapping: an automated software workflow

Adam Kania[1], András Zlinszky[2]

[1] ATMOTERM S.A., ul. Łangowskiego 4, 45-031 Opole, Poland
yuia@op.pl

[2] Vienna University of Technology, Department of Geodesy and Geoinformation, Research Groups Photogrammetry and Remote Sensing, Gußhausstraße 27 – 29, 1040 Vienna, Austria, az@geo.tuwien.ac.at

1. Introduction

Classifying vegetation based on remotely sensed measurements (ALS) is still largely a topic of rather research interest, but not a widely available commercial service.

Part of the difficulty to bring operational – that is fast, accurate enough for indended use and cost effective – vegetation classification to widespread use in the market, is that in order to produce a quality vegetation map, a number of highly qualifed experts (biologists/ecologists, remote sensing, machine learning, data processing and IT specialists) need to collaborate, often in similar time and/or place, which puts extra demands from organizational and logistical point of view, at the same time preventing any radical drops in costs of such a service (or actually preventing it from becoming a readily available service in the first place).

Another constraint is that the the state-of-the-art algorithms developed for various scientific tasks can only be assimilated very slowly by the producers of major market software packages, keeping them out of reach for everyone except experienced programmers who can code these tools themselves. Most of the large software packages are oriented towards general GIS and image processing tasks, limiting their applicability specifically for vegetation mapping.

This study presents a successful implementation of an integrated suite of software tools called Vegetation Classification Studio, that makes a large part of the process of producing vegetation maps highly efficient by automating what can be made automatic, and by supporting human experts in tasks require their decision and knowledge.

The main principles driving the design of the tool were:

- automating as much as possible of all tedious and repetitious data processing tasks, especially those requiring a lot of precision and care when performed by human operators, and their essence does not require taking creative decisions, or when the quality of their results can be assessed and verified in an objective manner, allowing to substitute human labor with an optimisation algorithms
- implementing enough expert knowledge in the software form so that most decisions required during the processing can be made automatically with a reasonable quality, resorting to human expert knowledge rather only at the very beginning and very end of the process
- try to limit the amount of input data required to bare mininum, taking care of producing of all the necessary intermediate data products automatically, as part of the internal workflow
- support creating and evaluation of many variants and alternatives of vegetation maps, being smart in suggesting reasonable alternative parameters of algorithms at different

processing steps, quantifying their impact on accuracy and quality of the final results, and making the multiple processing results available in a standard form suitable for convenient evaluation by the expert
- processing time matters, so result should be produced fast! — and it's not because of computing resources – but because producing high quality vegetation map requires mutliple cycles of consecutive correction and improvement, and any time that a human waits for a machine to process results impact directly the interactivity of the work and lengthens the final schedule

2. Method

We typically tested the Vegetation Classification Studio on Laser Scanning data, therefore the first step of the process was to calculate point cloud products. This was implemented using OPALS modules: In the first step of the algorithm, subsets of the sensor data are created within the calibration and validation plot outlines. Then, within these patches, a very large number of point cloud attribute derivatives is calculated, representing geometric, radiometric and roughness properties of the surface. Noise reduction and texture filtering is also applied automatically in order to further enhance the information content of the data. The input data can be complemented with any kind of information that can be converted to a raster: standard single- and multi-channel .tif files are used for this intermediate step, allowing direct interoperability with image data sources, vector maps, or other kinds of field information.

A vegetation classification task is essentially a learning problem: from a set of points in n-dimensional space (the data pixels with their respective attribute values), the user-defined groups have to be produced, based on a limited sample of training data. For this general setting, a machine learning algorithm was implemented in Python language, based on the scikit-learn library, accessing the sensor data through GDAL. The learner selected the data products with the highest information content for the respective task, and automatically creates a decision tree for classification. The result is for each pixel a set of probabilities that the pixel belongs to a particular vegetation class.

The accuracies of the classification are tested on the validation dataset, a standardized text report is output, and if this is accepted, a reduced-resolution graphical rendering of the whole study area is produced. For this, only the data derivatives corresponding to the optimum settings are calculated. Visual checking of this output rendering allows detection of any artefacts or classification errors, adjusting the algorithm if necessary, and finally rendering the full-resolution product raster.

3. Results

The Vegetation Classification Studio was tested on a range of sensor data products and habitats: forests, grasslands and wetlands were all tested, mostly with laser scanning data but in some cases also including imaging spectrometry. In practically all cases, the number of detectable classes and their accuracy and reliability were far superior to those achievable with mainstream software. Up to thirty classes could be accurately represented from sensor data, and the reporting and rendering scheme allowed quick feedback to the ecologists of the team, who could focus on extending field data or identifying problems with the output.

One lesson learned was that the system is very sensitive to imperfections of sensor data georeferencing and systematic errors. However, the artefacts these produce are easy to recognize and can often be corrected. The speed of processing allowed several hundred square kilometers to be rendered within hours, allowing for rapid checking of multiple options and final selection of the best products. The output data is ready for immediate viewing in a GIS and colour schemes can easily be adjusted to allow intuitive understanding (including the blending of colours for representing fuzzy class membership), facilitating dialogue between biologists, programmers and remote sensing scientists.

4. Discussion

The Vegetation Classification Studio has proven to be suitable for a wide range of input data and studied vegetation. It has allowed a new way of collaboration between ecologists, remote sensing scientists and programmers, and took remote sensing classification to the very edge of what was believed possible.

The promise from the title – to produce a quality vegetation map in an hour – still cannot be considered realistic for real-world situations, especially when we take into account the time required to prepare field and sensor data. In optimal conditions it did actually happened, that a whole process of making a new version of a vegetation map – after some field data adjustments or with new ALS derivatives or algorithm parameters – resulted in a biologist being able to evaluate a new version of a classified vegetation map of a large area in his GIS viewer in just about 1 hour, which makes that goal seem more and more realistic in a near future.

Future plans are to test the Studio on integrated data from different sensors, apply it further to new habitats and finally to make it commercially available.

Acknowledgements

The work presented in this study was partially funded by ChangeHabitats2: an IAPP Marie Curie project of the Seventh Frame Programme of the European Commission (FP7-PEOPLE-2009-IAPP – Grant Agreement Number 251234).