Implementation of a new algorithm for Clapas in R language in order to improve efficiency of soil survey

Sébastien Lehmann [A], Olivier Vitry [B], Dominique Arrouays [A].

[A]INRA INFOSOL, Orléans, France, Email sebastien.lehmann@orleans.inra.fr, dominique.arrouays@orleans.inra.fr
[B] Université d'Orléans, UFR de Mathématiques, Orléans, France, Email vitryolivier@hotmail.fr

Abstract

The growing development of information technologies and the breakthroughs made in signal and image analysis allow automating many protocols. Digital Soil Mapping is one of the numerous applications illustrating this ability to engineer and innovate computed, automated and repeatable treatment and analysis methods. The clapas algorithm developed in 1994 by J.-M Robbez-Masson is within this framework. It allows the segmentation of an image using reference areas and ancillary variables chosen to discriminate different soil mapping units. The map resulting from this segmentation can then be used as a basis for a traditional soil mapping as it simplifies and eases the field work. We have implemented this algorithm in the 'R' language in order to generalize its use and to enhance its functionalities.

Key Words

Clapas, soilscape, Digital Soil Mapping, spatial segmentation, spatial analysis.

Introduction

Digital soil mapping is computer-assisted production of a digital map of soil type and/or soil properties. The generic framework of Digital Soil Mapping has been defined by McBratney *et al.* (2003) as scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method. Most classification algorithms do not integrate spatial relationships within the model. However, environmental attributes at neighbouring locations can be of great interest in predicting soil pattern. Therefore, some authors investigated the potential of incorporating local neighbourhood information into the training pixels using convolution filtering operations (Grinand *et al.* 2008). Another way to integrate spatial relationships is to analyse the global patterns of soil forming factors over a window centred on the pixel of interest (Lagacherie *et al.* 2001). This kind of method has also been used in remote sensing for segmentation and landscape delineation. A method for calculating distance between soilscapes, named Clapas, was initially designed for describing quantitatively, comparing and classifying soilscapes in small-scale soil surveys (). It has then been adapted by Lagacherie *et al.* (2001) to map the representativity of a reference area.
In this study, the Clapas method is implemented in a free software environment and with recent spatial exploration advances, namely implementing moving windows that can be oriented according to the main relief characteristics.

Methods

Calculating distances between soilscapes with Clapas

Clapas was initially designed for describing quantitatively, comparing and classifying soilscapes in small-scale soil surveys (Robbez-Masson 1994). The procedure has three steps: (i) defining a soilscape variable at a given point from available soil forming factors, (ii) using this soilscape variable for a quantitative synthetic description of the site soilscape, and (iii) comparing sites regarding their soilscapes.
In the following we detail these steps in succession.

Let $v(x)$ be a variable describing the elementary landscape at each site x, i.e. the point-to-point combination of soil forming factors. $v(x)$ is a categorical variable taking its values in $\{v_1, v_2, v_i, v_\mathrm{p}\}$, the set of the p elementary landscape classes of the region. Each elementary landscape class vi is defined by a unique combination of soil forming factor classes on a point-to-point basis. Soil forming factor classes are either mapping units (e.g. geological units or land use classes) or derived from a pre-classification of quantitative variables (e.g. elevation, slope gradient of a DEM). p classes can be identified, p being less or equal to the product of the numbers of classes of each soil forming factor.

The soilscape of site x is then defined by a "cover-frequency vector" (Wharton 1982)
$\boldsymbol{l}(x) = (f_1 (x), f_2 (x), . . . , f_i (x), . . . , f_p (x),)$ where $f_i (x)$ is the relative frequency of class v_i within an area

delineated around *x* to include the set of neighbouring sites which must be taken into account for describing the soilscape at *x*. The size and the shape of this area are user-defined. In this new version of Clapas, that we have developed in R language, the neighbourhood area is elliptic, which requires the setting of the two main radii of the ellipse, and can be automatically oriented according to the slope aspect.

The soilscape of site *x* can then be compared quantitatively with the one of a reference site *x* by computing the distance $d(x, x_0)$ between the vectors $\boldsymbol{l}(x)$ and $\boldsymbol{l}(x_0)$. The following Manhattan distance was preferred among distances dealing with qualitative variables because of its robustness:

$$d(x, x_0) = \sum_{i=1}^{p} |f_i(x) - f_i(x_0)| \quad . (1)$$

The distances calculated with Eq. (1) range between 0, i.e. same composition of classes within the explored areas, and 2, i.e. disjoined cover-frequency vectors with no classes in common. These distances can then be used for allocating individuals to pre-defined reference soilscapes. In Clapas, allocation is performed on a nearest neighbour basis. Figure 1 provides a simple example of a Clapas application. First, four elementary soilscape classes are defined by combinations of soil forming factor maps (Figure 1a).

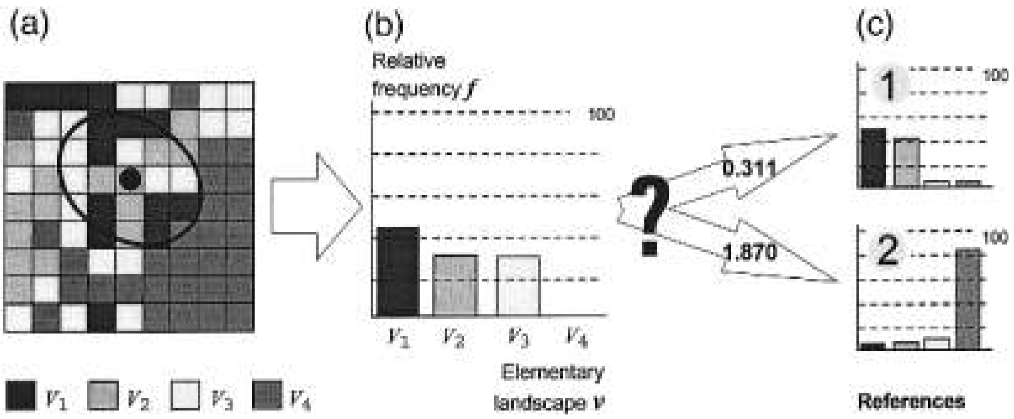


Figure 1. An example of application of the Clapas Procedure.

Then, the soilscape of the black dotted site of Figure 1a is quantitatively described by a cover frequency vector (histogram Figure 1b), which is computed from an elliptic neighbourhood of the site. Finally, the soilscape description of the black dotted site is compared with both soilscape descriptions of reference sites (1 and 2, Figure 1c). In this example, the computed Manhattan distances reveal that the studied soilscape is closer to reference site 1 (distance = 0.311) than reference site 2 (distance = 1.870).

In practice, the Clapas procedure deals with raster cells of a landscape scale image, each cell representing a site of the region. Prior to Clapas processing, the input maps of soil forming factor are thus converted into a grid structure (Lagacherie *et al.* 2001). Preparation and postprocessing of the data are done within a free statistic software (R) and Geographical Information Systems (ArcGis^{TM}).

The region studied and the data

The region studied is situated in the south west of France, at the piedmont plain of the Pyrenees, in the Gers department. It corresponds to a soil map of Mirande at 1:50 000 scale (Figure 2).

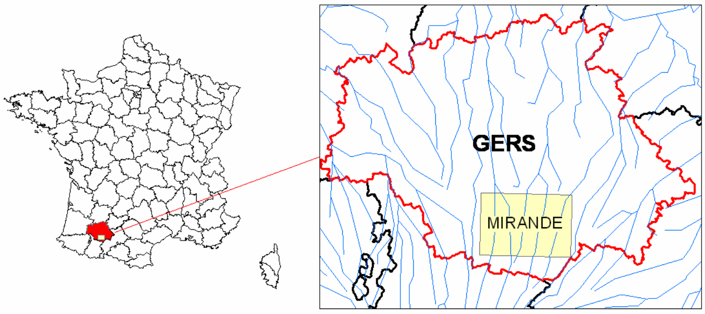


Figure 2. Study area: The map of mirande is located in south west of France, in the Gers department.

Between the dissymmetric valleys, the soft hills are made of a stack of layers corresponding to sedimentary cycles. The rhythmic sedimentation begins with pudding stone or conglomerate, then molasse and limestone beds covered by a thick layer of marl.

The pedological context is characterized by a general phenomenon of decarbonatation and eluviation which leads quickly to a reduction phasis followed by planosolization.

Those conditions have produced soils like Fluvisols, luvic, eutric, calcaric and hypereutric Cambisols, Luvisols and Albeluvisols.

Our prior experience in surveying the region, and statistical tests (Lehmann *et al.* 2007) led us to select three landscape parameters for identifying soil associations: parent material classes, Beven index (Beven *et al.* 1979) and Relief Index. Those two last indices were calculated with ArcGis^{TM} tools from a 50x50 m Digital Elevation Model distributed by the French National Geographic Inventory. The Beven and Relief Index were discretized into 3 and 4 classes respectively. They have been further combined with the parent material map containing 8 classes, into a new image (Figure 2) as required by the procedure for calculating soilscape distances.

The soil map of Mirande was used for running and validating the procedure of soil prediction from the reference area to the rest of the map.

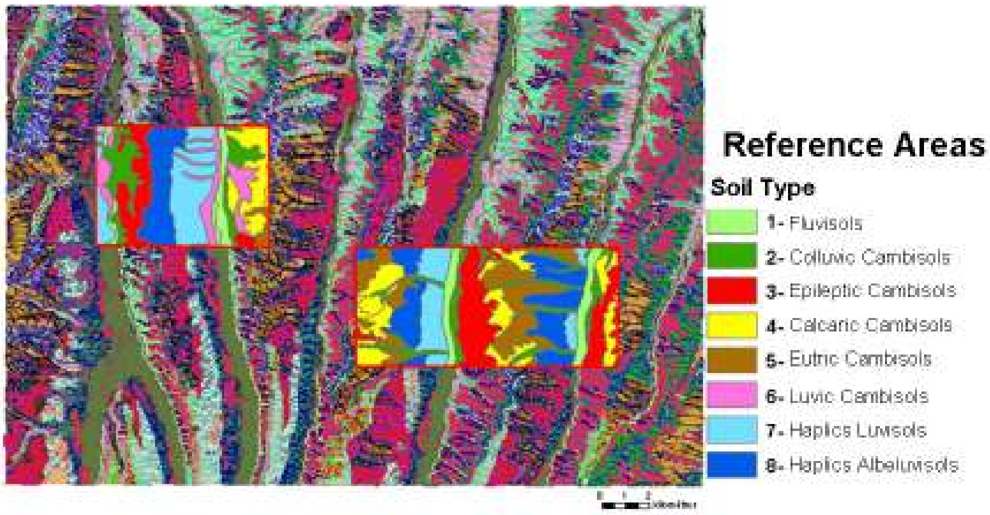


Figure 3. Image of the combined parameters covered by the reference areas.

Results

A simple visual comparison of the simplified soil map with the one generated by Clapas (Figure 4) is enough to judge the strong resemblance of the two maps.

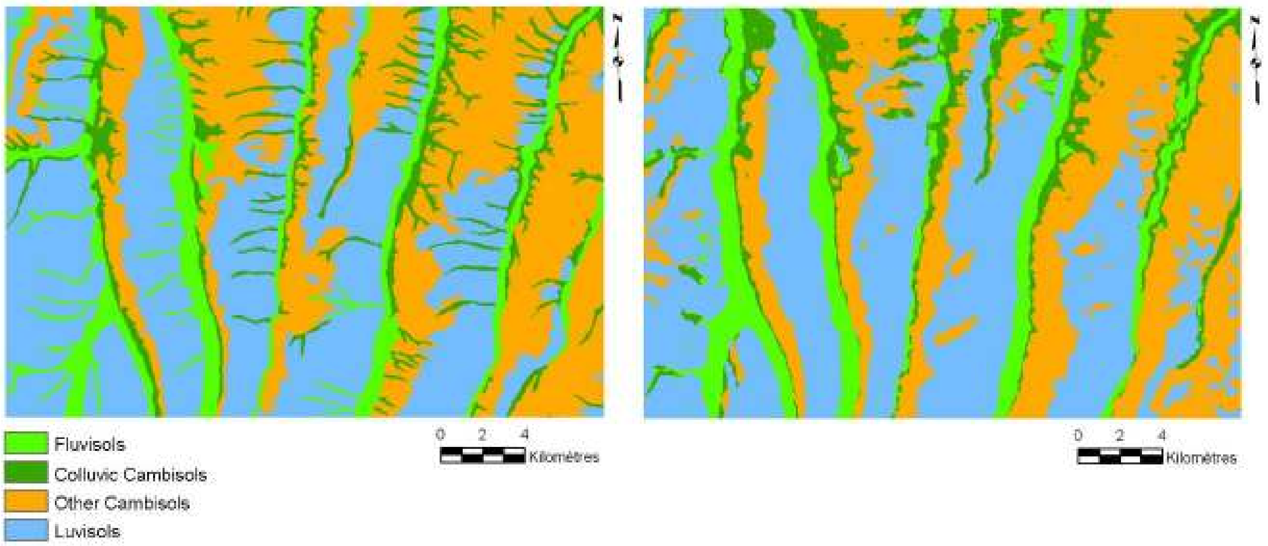
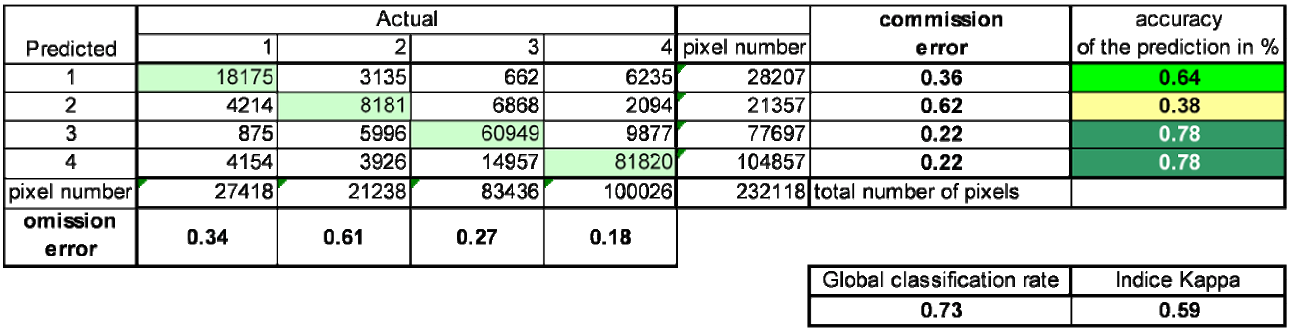


Figure 4. Comparison of the simplified soil map (left), with the one generated by Clapas (right).

The confusion matrix (Table 1) confirms it. The classifier has a 73% global recognition rate. It is also ensured by an acceptable kappa rate of 59%. Looking in detail at the results of the classification, we can see from the matrix that the separation between Colluvic Cambisols and the other Cambisols is difficult. Fluvisols have a 64% recognition rate despite a relative strong confusion with the 4th class: the Luvisols, which represent 1/5th of the prediction for that class.
The Luvisols and the Cambisols are quite well predicted with the same rate of 78%.



	Actual					commission error	accuracy of the prediction in %
Predicted	1	2	3	4	pixel number		
1	18175	3135	662	6235	28207	**0.36**	**0.64**
2	4214	8181	6868	2094	21357	**0.62**	**0.38**
3	875	5996	60949	9877	77697	**0.22**	**0.78**
4	4154	3926	14957	81820	104857	**0.22**	**0.78**
pixel number	27418	21238	83436	100026	232118	total number of pixels	
omission error	**0.34**	**0.61**	**0.27**	**0.18**			

Global classification rate	Indice Kappa
0.73	**0.59**

Table 1. Confusion matrix of the prediction with 4 classes.

Conclusion

The implementation of Clapas in R shows exactly the same results than the results produced with the old version of Clapas (Lehmann 2007). However it has now two new improvements that must be tested in the future and compared with other procedures like Mart Classification (Grinand *et al.* 2008). The first is that the number of combination in the image of entry is no more limited to 255 which enable to create more classes and will likely improve the accuracy of the prediction. The second is that the orientation of the elliptic window that considers the neighbourhood of a pixel can now be automatically oriented according to slope aspect.

References

Beven K, Kirkby M (1979) A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* **24**, 43-69.

Grinand C, Arrouays D, Laroche B, Martin MP (2008) Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* **143**, 180-190.

Lagacherie P, Robbez-Masson JM, Nguyen-The N, Barthès JP (2001) Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma* **101**, 105-118.

Lehmann S, Bégon JC, Eimberck M, Daroussin J, Wynns R, Arrouays D (2007) Utilisation du logiciel Clapas pour l'aide à la délimitation de pédopaysage. *Etude et gestion des sols* **14**, 135-151.

McBratney AB, Mendonça-Santos ML, Minasny B (2003) On digital soil Mapping. *Geoderma* **117**, 3-52.

Robbez-Masson JM (1994) Reconnaissance et délimitation de motifs d'organisation spatiale. Application à la cartographie des pédopaysages. PhD Thesis, Ecole Nationale Supérieure Agronomique de Montpellier.

Wharton SW (1982) A contextual classification method for recognizing land use patterns in high resolution remotely sensed data. *Pattern Recognition* **15**, 317–324.