Cokriging for large spatial datasets: mapping soil properties at a region scale from airborne hyperspectral imagery

Emily Walker, Pascal Monestiez, Cécile Gomez, Rossano Ciampalini and Philippe Lagacherie

Abstract Recent developments in soil sensing technologies, initially oriented towards soil mapping at the field scale for precision agriculture, show high potential for digital soil mapping (DSM) of large areas. We present here a spatial statistical model that combines hyperspectral remote sensing, field measurements and, potentially soil types from existing pedological maps, to predict soil properties as clay or calcium carbonate contents at increasing resolutions from 5m to 100m over large regions. Methodological difficulties arise from dimensional aspects. From a spatial point of view, the geostatistical model have to be inferred from rare field soil samples and remote sensing data that are patchy - only informative on bare soils - and very numerous - several thousand records at fine resolution. From a multivariate point of view, soil properties have to be predicted using PLS from high dimensional - 256 bands - hyperspectral data. To illustrate the proposed approach, a 25-square-km area located in the vineyard plain of Languedoc were surveyed with both airborne hyperspectral remote sensing data at a 5-m resolution and complementary field survey. Various maps of clay and calcium-carbonate content were produced by cokriging and represent different compromises between prediction accuracy and spatial resolution.

Emily Walker and Pascal Monestiez
INRA, Unité de Biostatistique et Processus Spatiaux (BioSP), Domaine Saint Paul, Site Agroparc, 84914 Avignon cedex 9, France.
e-mail: emily.walker@avignon.inra.fr, monestiez@avignon.inra.fr

Cécile Gomez
IRD, Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH), IRD-INRA-SupAgro, Campus SupAgro, Bat.24 - 2 place Pierre Viala - 34060 Montpellier, France e-mail: gomez@supagro.inra.fr

Rossano Ciampalini and Philippe Lagacherie
INRA, Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH), IRD-INRA-SupAgro, Campus SupAgro, Bat.24 - 2 place Pierre Viala - 34060 Montpellier, France.
e-mail: ciampali@supagro.inra.fr, lagache@supagro.inra.fr

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

1 Introduction

Given the relative lack of, and huge demand for, quantitative spatial soil information to be used in environmental management and modelling, digital soil mapping (DSM) has been proposed as an alternative to classical soil surveys for the quantitative mapping of soil properties over regions at intermediate (20-200 m) spatial resolutions (McBratney et al., 2003). Among the available soil sensors, visible near infrared (Vis- NIR) imaging spectrometry looks to be one of the most promising. In laboratory studies, the capability of Vis-NIR spectroscopy (450-250 nm) to accurately quantify soil property contents has been already proven (Viscarra Rossel et al., 2006). More recently, spatial predictions of some usual soil properties for bare soil surfaces were obtained from high-resolution airborne hyperspectral images with uncertainties ranging from $R-2$ = 0.53 to 0.75 depending on the study areas and their properties (Selige et al., 2006; Gomez et al., 2008; Lagacherie et al., 2008; Stevens et al., 2010; Schwanghart and Jarmer, 2011). Although these results revealed a decrease in precision because of atmospheric effects and the signal to noise ratio (SNR) of the instrument (Lagacherie et al., 2008), imaging spectrometry provided correlations with soil properties of bare surfaces that out-performed most of the soil covariates usually considered in DSM applications. The DIGISOL-HYMED project (funded by ANR- French National Research Agency) examines how this new input can be used for the DSM of soil properties over large spatial areas.

This study aims to map soil properties (clay and $CaCO_3$ as examples) over large regions from airborne hyperspectral images of bare soils. The complexity lays on the limited number of soil samples and the high number of hyperspectral images. The objective is to interpolate from a small set of exact measurements (for calibration purpose) using large amount of data which induces uncertainty due to vegetation, a patchy spatial coverage, and a fine resolution.

2 Case study and data set

The study area is a 23.9 km^2 vineyard area, Peyne catchment (Figure 1) in the south of France (43°29'N and 3°22' E).

This area is characterised by vineyards which form the primary land-use, and present a great variability of top soils (different alluviums or sediments).

An earlier ground sampling made in the study region (Lagacherie et al., 2008) showed that these complex soil patterns correspond to a great variability of clay content at the soil surface (from 65 to 452 $g.kg^{-1}$). A study area of 24.6 km^2 (Figure 1) was defined by intersecting this region of interest with the hyperspectral image used in this study.

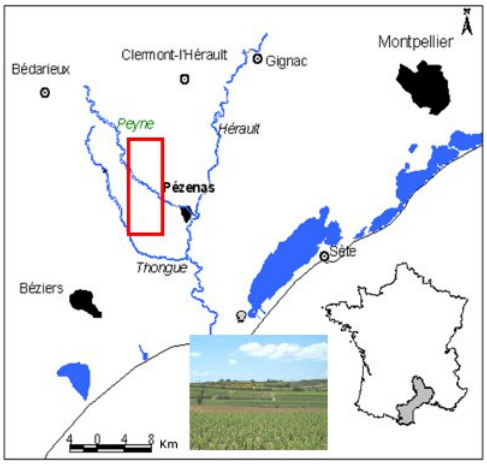

Fig. 1 Location of the study area (red rectangle), near Montpellier in the South of France, with a vineyard picture.

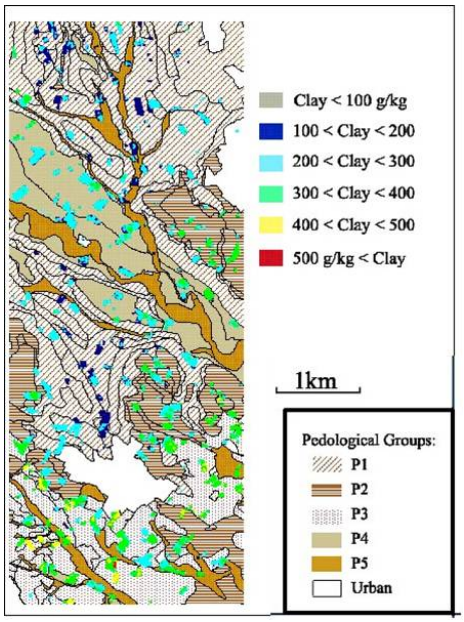

Fig. 2 Prior geological information with 5 pedological groups, and field soil samples with colors corresponding to clay levels. P1: Miocene alluvial deposits, P2: Pliocene alluvial deposits, P3: rocky basement, P4: old alluvial deposit from la Peyne river, P5: recent alluvial deposit from la Peyne river.

2.1 Field soil data

The available soil data consist on 95 sites with laboratory measurements of soil properties on bare soils and 53 sites on vegetated soils. The locations of soil data are plotted in Figure 2.

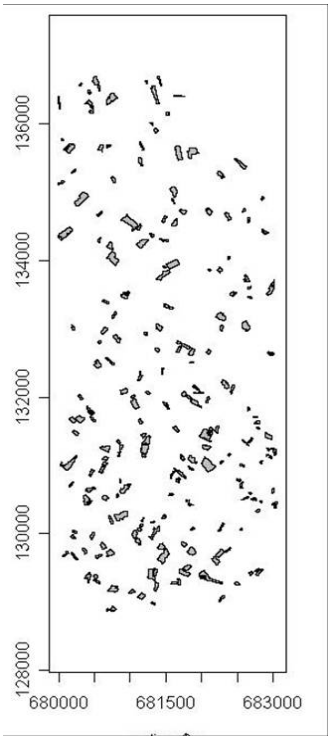

Fig. 3 HyMap image (37230 pixels at 5x5m resolution).

2.2 Hyperspectral covariates

Airborne Vis-NIR imaging spectrometry differs greatly in spatial resolution and extent from those currently handled in DSM. Airborne Vis-Nir sensors provide data at very fine spatial resolutions, including less than 5 m, which is much finer than the resolutions of the usual spatial covariates and target resolutions of DSM (see Introduction). Also, the applications of imaging spectrometry are limited in space because of clouds and vegetation that mask the soil surface. These disruptions can result in scattered spatial data with isolated measured areas separated by non-measured areas.

The mapping was carried out over a 24.6 km^2 area located in the vineyard plain of Languedoc with usable hyperspectral data scattered over only 3.5% of this area.

The hyperspectral images obtained have 37 230 pixels for 5mx5m resolution (Figure 3). The spectrum is comprised from 400nm to 2500 nm and has 126 bands.

2.3 Geological classes

The studied area is characterised by 5 kinds of geological classes of alluvial deposits (Figure 2). Because we assume that soil properties can be influenced by belonging of measurement or predicted sites to one geological class, they will be considered as categorial covariates in the cokriging.

3 Methods

In order to reduce the dimensionality of the problem, the interpolation was decomposed in two stages : (1) we first used a Partial Least Square regression (PLSR) procedure on the 126 bands of the Hymap data to predict for each pixel a value of the soil property at this site (one for each 37 230 pixels) and (2) then we mapped the soil property everywhere using a bivariate Universal Cokriging (UC) based in soil samples (measured values of the soil property) and Hymap PLSR predictions (predicted variables). Geological classes introduced constrains in the UC to account for nonstationarity in mean among the classes.

3.1 PLS Regression

To reduce dimension of spectral data, the objective was to predict Clay content (or $CaCo_3$) from 126 bands spectra data using regression. As studied in Gomez et al. (2008), soil properties predictions from laboratory and Hymap spectra were computed for several properties using Partial Least Square Regression (PLSR). Gomez et al. (2008) used a small data set with only sites with a soil sample and Hymap data on bare soil, i.e. 95 sites. More spectral bands (126 variables) than measured sites (95 individuals) were available, which is essential to avoid overfitting. But it is important to note that spatial dependences were ignored.

Then, the objective of the study is to map Clay content (and $CaCO_3$ resp.) from field soil samples and predictions of clay (and $CaCO_3$ resp.) by PLSR (considered as covariate in the cokriging).

3.2 Variographic analyses

Simple variograms were fitted respectively from 95 sample sites (and from 148 sample sites, for bare and vegetated samples) and 37230 pixels with PLSR prediction. Cross variogram was fitted on 95 sites only with both variables. Linear models of coregionalisation were fitted for clay and $CaCO_3$ (Wackernagel 1995).

3.3 Neighbourhood selection

A moving neighbourhood was defined to select Hymap values (from PLSR results), and to avoid a higher dimension of variance-covariance matrix for cokriging calculation.

3.4 Cokriging with external drift

The variable of interest, i.e. the soil property, is modelled by a random function $Z(x)$ where x denotes the location index (vector of coordinates). $Z(x)$ is decomposed into a deterministic unknown drift $m(x)$ and a stationary zero-mean random function $Z_R(x)$. In the kriging with external drift approach, $m(x)$ is modelled as a linear function of a deterministic external variable. In the version proposed by Monestiez et al. (1999; 2001) and used here, $m(x)$ is modelled as a set of values $e_k, k = 1,\ldots,p$, corresponding to the five geological classes ($p = 5$). The values e_k may be unknown, but the spatial partition of the domain in geological classes must be known everywhere. The model can be written as

$$Z(x) = \sum_{k=1}^{p} \mathbb{1}_{\{k\}}(x)\, e_k + Z_R(x)$$

where e_k is a mean effect for class k to be estimated and $\mathbb{1}_{\{k\}}(x)$ is the indicator function of the class k: it is equal to one if x is in class k, and it is equal to zero otherwise. The variable Z was sampled at n_i sites x_i, for $i = 1,\ldots,n_i$. ($n_i = 95$)

The second variable $Y(x)$, i.e. the covariate of the bivariate cokriging which is here the predicted property by PLSR, is modelled the same way.

$$Y(x) = \sum_{k=1}^{p} \mathbb{1}_{\{k\}}(x)\, e'_k + Y_R(x)$$

The variable Y was sampled at n_j sites x_j, for $j = 1,\ldots,n_j$ and where n_j is the number of neighbours selected among the 37230 hymap pixels.

To symplify notation in the following, the covariance function of Z for a pair of points $C_{ZZ}(x_i - x'_i)$ is noted $C^{(ZZ)}_{i,i'}$ and the cross-covariance between Z and Y, $C_{ZY}(x_i - x_j)$ is noted $C^{(ZY)}_{i,j}$. Covariances and cross-coavariances are directly derived from fitted variograms and co-variograms.

The cokriging predictor is formally the same as an Universal Cokriging:

$$Z^*(x_0) = \sum_{i=1}^{n_i} \lambda_i\, Z_i + \sum_{j=1}^{n_j} \lambda'_j\, Y_j,$$

where the λ_i's and λ'_j's solve the following cokriging system with $n_i + n_j + 2p$ equations to ensure unbiasedness and minimisation of the MSE:

$$\begin{cases} \sum_{i'=1}^{n_i} \lambda_{i'} C_{i,i'}^{(ZZ)} + \sum_{j=1}^{n_j} \lambda'_j C_{i,j}^{(ZY)} - \sum_{k=1}^{p} \mu_k \mathbb{1}_{\{k\}}(x_i) = C_{i,0}^{(ZZ)} & \text{for } i = 1,\dots,n_i \\ \sum_{j'=1}^{n_j} \lambda'_{j'} C_{j,j'}^{(YY)} + \sum_{i=1}^{n_i} \lambda_i C_{i,j}^{(ZY)} - \sum_{k=1}^{p} \mu'_k \mathbb{1}_{\{k\}}(x_j) = C_{j,0}^{(ZY)} & \text{for } j = 1,\dots,n_j \\ \sum_{i=1}^{n_i} \lambda_i \mathbb{1}_{\{k\}}(x_i) = \mathbb{1}_{\{k\}}(x_0) & \text{for } k = 1,\dots,p \\ \sum_{j=1}^{n_j} \lambda'_j \mathbb{1}_{\{k\}}(x_j) = 0 & \text{for } k = 1,\dots,p \end{cases}$$

Compared to the ordinary cokriging system, $2p-1$ constraints are added so that the prediction error is free from class effects: the sum of the weights for the class at x_0 must be one, and the sum of weights in all other classes must be 0. As a consequence, the unit sum constraint on the λ_i's is directly obtained by summing the p constraints $\sum_{i=1}^{n_i} \lambda_i \mathbb{1}_{\{k\}}(x_i) = \mathbb{1}_{\{k\}}(x_0)$ for $k = 1,\dots,p$. There are $2p$ Lagrange parameters μ_1 to μ_p and μ'_1 to μ'_p. Only one term μ remains in the kriging variance whose expression is :

$$\sigma_K^2(x_0) = C_{0,0}^{(ZZ)} - \sum_{i=1}^{n_i} \lambda_i C_{i,0}^{(ZZ)} - \sum_{j=1}^{n_j} \lambda'_j C_{j,0}^{(ZY)} + \sum_{k=1}^{p} \mu_k \mathbb{1}_{\{k\}}(x_0).$$

4 Results

4.1 Fitted linear models of coregionalisation

The simple and cross variograms were plotted in Figure 4. The linear model of coregionalisation was fitted on experimental variograms and calculated only from the 95 colocated samples and Hymap values. The fitting was realised from two spherical models with intermediate and long ranges (around 250m and 2200m).

4.2 Cokriging maps

The universal cokriging maps were calculated for clay and Carbonate calcium (Figure 5).

5 Discussion

This study is the following step after the work presented in Lagacherie et al. (2012), integrating external drift with the categorical covariates from geological classes. To add categorical covariates improved the properties estimations, assuring geological coherence in results. Moreover the cokriging maps were calculated for several properties as pH, silt, sand, cation-exchange capacity. Only clay and $CaCO_3$ results are presented here.

The characteristics of the issue are:

- high dimension datasets with one variable of interest, 126 variables from spectra data, and categorical covariates
- 95 field sites and 37200 Hymap sites (which are aggregated sites)
- cokriging systems needed to be simplified

A interesting idea to improve this work would be to build a global model intergrating PLSR and cokriging calculations both. The presence of a PLS predicted value at the interpolated site induced that collocated cokriging would be sufficient.

Hymap points selection in the neighborhood was choosen to avoid a selection of more than 2000 points in a reasonable distance, and to avoid redondancy in points. This methodology is also applied to Tunisian issues (Cap Bon datasets) (Ciampalini et al.,2012; Gomez et al., 2012).

Acknowledgements The authors are indebted to UMR LISAH (IRD, France) and to CNCT (Centre National de Cartographie et de Télédétection, Tunisia), for providing the AISA-Dual images for this study. This hyperspectral data acquisition was granted by IRD, INRA and the French National Research Agency (ANR) (ANR-O8-BLANC-284-01).

References

1. Ciampalini, R., Lagacherie, P., Hamrouni, H., 2012. Documenting GlobalSoilMap.net grid cells from legacy measured soil profile and global available covariates in Northern Tunisia in DSM2012, Minasny et al (eds). Sydney. Preti, G., Cisbani, A., De Cosmo, V., Galeazzi, C., Labate, D., Melozzi, M., (2008). Hyperspectral Instruments for Earth Observation. International conference on Space Optics, October 14-17 Toulouse, France.
2. Gomez, C., Lagacherie, P., Bacha, S., 2012.Using Vis-NIR hyperspectral data to map topsoil properties over bare soils in the Cap Bon region, Tunisia . in DSM2012, Minasny et al (eds). Sydney.
3. Gomez, C., Lagacherie, P., Coulouma, G. 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma, 148, 141148.
4. Lagacherie, P., Baret, F., Feret, J.B., Madeira Netto, J.S.,Robbez-Masson, J.M. 2008. Clay and calcium carbonate contents estimated from continuum removal indices derived from laboratory, field and airborne hyper-spectral measurements. Remote Sensing of Environment, 112, 825835.
5. Lagacherie, P., Bailly, J.S., Monestiez, P., Gomez, C., 2012. Using scattered hyperspectral imagery data to map the soil properties of a region. Eur.J. Soil Science ; 63 :110-119

6. McBratney, A.B., Mendonca Santos, M.L., Minasny, B. 2003. On digital soil mapping. Geoderma, 117, 352.
7. Monestiez, P., Allard, D., Navarro Sanchez, I., Courault, D., 1999. Kriging with categorical external drift: Use of thematic maps in spatial prediction and application to local climate interpolation for agriculture, in *geoENV II: Geostatistics for Environmental Applications*, Gomez-Hernandez J., Soares A. and Froidevaux R. Eds, Kluwer Academic Publishers, Dordrecht, 163-174.
8. Monestiez, P., Courault, D., Allard, D., Ruget, F., 2001. Spatial interpolation of air temperature using environmental context: application to crop model. Environmental and Ecological Statistics ; 8 :297-309.
9. Ouerghemmi, W., Gomez, C., Nacer, S., Lagacherie, P. (2011). Applying Blind Source Separation on hyperspectral data for clay content estimation over partially vegetated surfaces. Geoderma, 163(3?), 227-237.
10. Schwanghart, W., Jarmer, T. 2011. Linking spatial patterns of soil organic carbon to topography a case study from south-eastern Spain. Geomorphology, 126, 252263.
11. Selige, T., Bohner, J., Schmidhalter, U. 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. Geoderma, 136, 235244.
12. Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma, 158, 3245.
13. Stuffler, T.; Kaufmann, H.; Hofer, S.; Förster, K.-P.; Schreier, G.; Müller, A.; Eckardt, A.; Bach, H.; Penne, B.; Benz, U.; Haydn, R.,2007: The EnMAP hyperspectral imager- An advanced optical payload for future applications in Earth observation programmes,Acta Astronautica, 61,1-6,115-120.
14. Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. 2010. Proximal Soil Sensing, Progress in Soil Science, Volume 1. Springer, Dordrecht, Heidelberg.
15. Wackernagel, H., 1995. Multivariate geostatistics. Springer- Verlag. 255 pp.

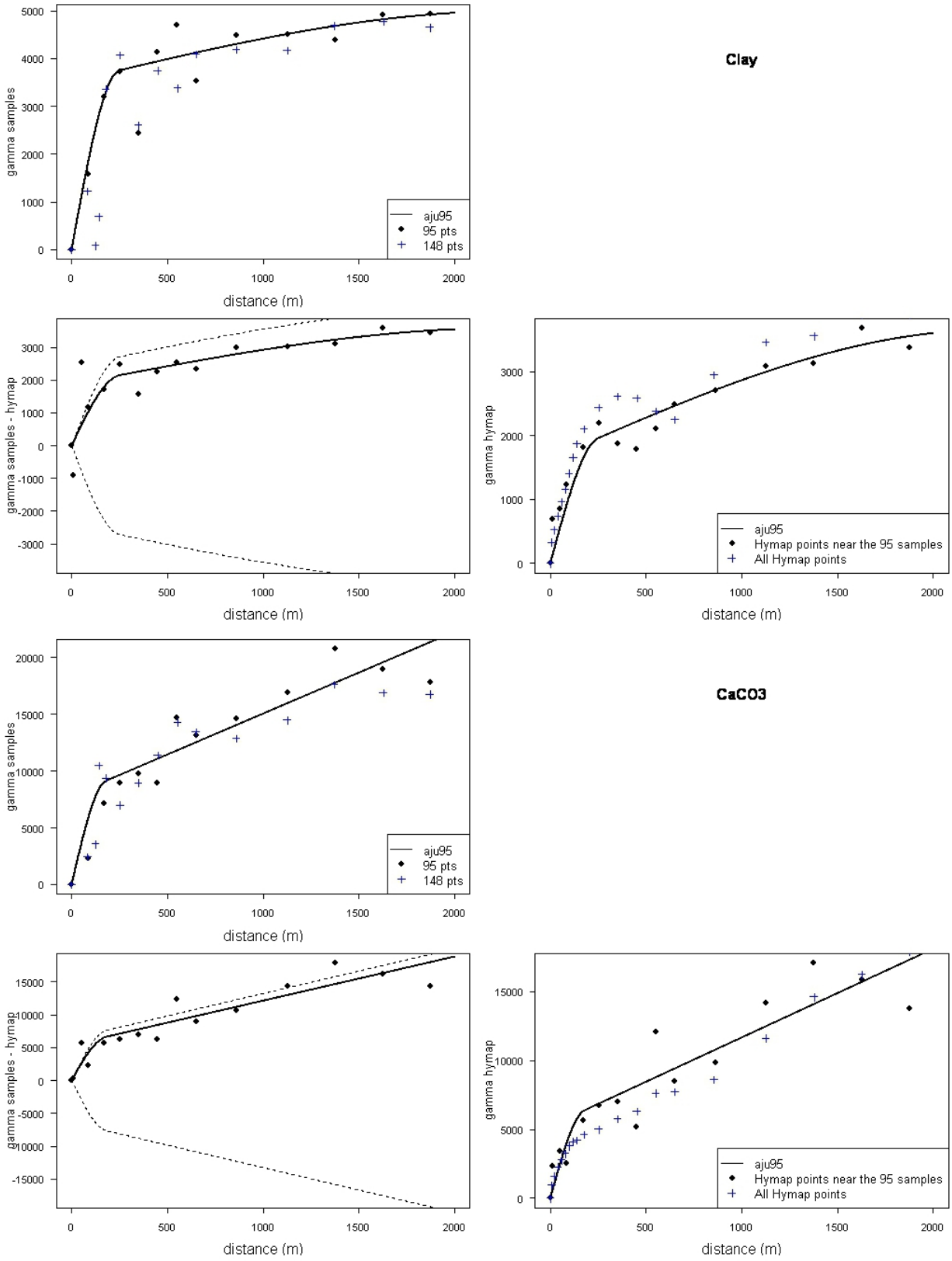

Fig. 4 Clay and $CaCO_3$ variograms, with fitted linear model of coregionalisation in bold lines.

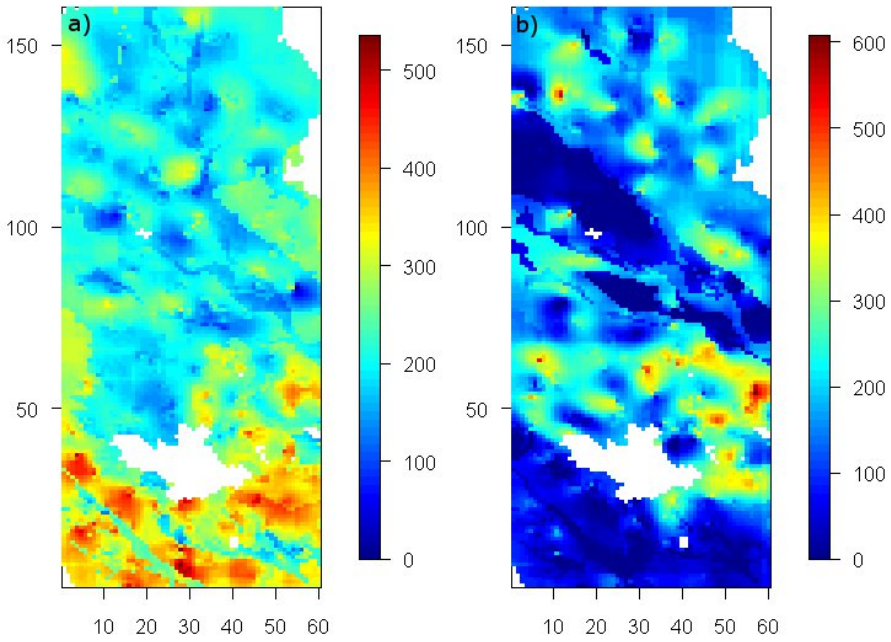

Fig. 5 Cokriging maps of clay (a) and $CaCO_3$ (b) (units: g/kg)