

# Geostatistical prediction of clay percentage based on soil survey data

Ari Talkkari, Lauri Jauhiainen and Markku Yli-Halla

*MTT Agrifood Research Finland, Environmental Research, FIN-31600 Jokioinen, Finland,  
e-mail: [ari.talkkari@mtt.fi](mailto:ari.talkkari@mtt.fi)*

In precision farming fields may be divided into management zones according to the spatial variation in soil properties. Clay content is an important soil characteristic, because it is associated with other soil properties that are important in management. Soil survey data from 150 sampling sites taken from an area of 218 ha were used to predict the spatial variation of clay percentage geostatistically in an agricultural soil in Jokioinen, Finland. The exponential and spherical models with a nugget component were fitted to the experimental variogram. This indicated that the medium-range pattern could be modelled, but the short-range variation could not, due to sparsity of sample points at short distances. The effect of sampling density on the kriging error was evaluated using the random simulation method. Kriging with a spherical model produced a map with smooth variation in clay percentage. The standard error of kriging estimates decreased only slightly when the density of samples was increased. The predictions were divided into three classes based on the clay percentage. Areas with clay content below 30%, between 30% and 60% and over 60% belong to non-clay, clay and heavy clay zones, respectively. With additional information from the soil samples on the contents of nutrients and organic matter these areas can serve as agricultural management zones.

*Key words:* clay soils, geostatistics, kriging, sampling, soil types, spatial variation

## Introduction

In precision farming fertilizers are applied to meet the needs of the crop and, on the other hand, to avoid excessive applications with the associated environmental implications. To fertilize different parts of a field according to the specific needs of the crop, the field can be divided into management zones based on previous yields and the data obtained from soil testing. The zones can be delineated in advance and the locations of the zones stored in the computer of advanced

machinery for fertilizer application, which then adapts the rates of application according to the location in the field. When sufficient numerical data on soil characteristics are available, management zones can be delineated using the methods of geostatistics and mapping.

Geostatistical analysis has become a widely used for predicting and mapping the spatial variation of soil properties provided that there is a large number of soil samples. According to Webster and Oliver (1992) at least 100 data are required to estimate the variogram acceptably. Geostatistics has been applied in agriculture,

especially in soil science. Recently, geostatistical analyses have been used to model the spatial variability of topsoil (Brooker 2001), investigate spatial variation of radon concentration in the soil (Oliver and Khayarat 2001) and predict spatially bulk density and field capacity of Ferral soils (Utset et al. 2000). Bocchi et al. (2000) used factorial kriging to characterize the spatial variation of soil physical, hydrological and chemical properties in a field in northern Italy, and Saldana et al. (1998) examined the variation of soil properties at different scales.

So far, geostatistics has not been widely adopted as an agricultural tool in Finland. Haapala (1995) used experimental variograms to analyse yield variation over short distances and found periodic variograms with wavelengths of 2 m to 2.5 m. Usually, the information on soil characteristics of a field is based on few widely spaced samples. In practical farming one or a few soil samples per field are used as the basis for management of the entire field as a homogeneous unit. However, soil properties, such as the clay content, may vary from under 10% to over 60% within a distance of 100 m (Jokinen 1983). If the spatial variation of soil properties can be modelled and predicted using geostatistics, management zones could be delineated according to soil characteristics instead of using the entire field as a management unit.

In the present study, soil survey data were used to predict the spatial variation in particle size fractions of agricultural soil, in an attempt to delineate management zones. The objectives of this study were to: i) determine how soil survey data can be used in the spatial prediction of clay content, ii) analyse the effect of sampling

density on the errors of the kriged predictions, and iii) delineate management zones for agricultural fields using geostatistical analysis and mapping.

## Material and methods

### Study area

The study area is located on the research farm of MTT Agrifood Research Finland in Jokioinen, southwestern Finland (23°27'E, 60°48'N). The survey area consisted of 37 fields with a total area of 218 ha. The material for this study comprised of 180 plots of soil survey data collected in 1996 with an average sampling density of 0.8 samples per hectare (Fig. 1). The Loimijoki River divides the survey area into the southern and northern parts.

The soil samples were bulked from 5 subsamples taken within a 100 m<sup>2</sup> plot. The soil texture of the 150 plots was analysed using a pipette method (Elonen 1970), in which hydrogen peroxide is used to decompose the aggregates and sodium pyrophosphate as the dispersing agent. The agricultural soil in Jokioinen is typically dominated by clay.

In Finland, the fine-earth fraction (< 2 mm) of soil is traditionally divided into 4 parts (Table 1), which are referred to here as clay, silt, fine sand and sand (Aaltonen et al. 1949). The particle sizes and their approximate equivalents in the USDA system (Soil Survey Staff 1993) are given in Table 1.

Table 1. The particle sizes and their approximate equivalents in the USDA system.

| Particle size, mm | Finnish term (Aaltonen et al. 1949) | Approximate equivalent in the USDA system (Soil Survey Staff 1993) |
|-------------------|-------------------------------------|--|
| < 0.002           | clay                                | clay   |
| 0.002–0.02        | silt                                | fine silt  |
| 0.02–0.2          | fine sand                           | coarse silt + very fine sand + fine sand                           |
| 0.2–2             | sand                                | medium + coarse + very coarse sand                                 |

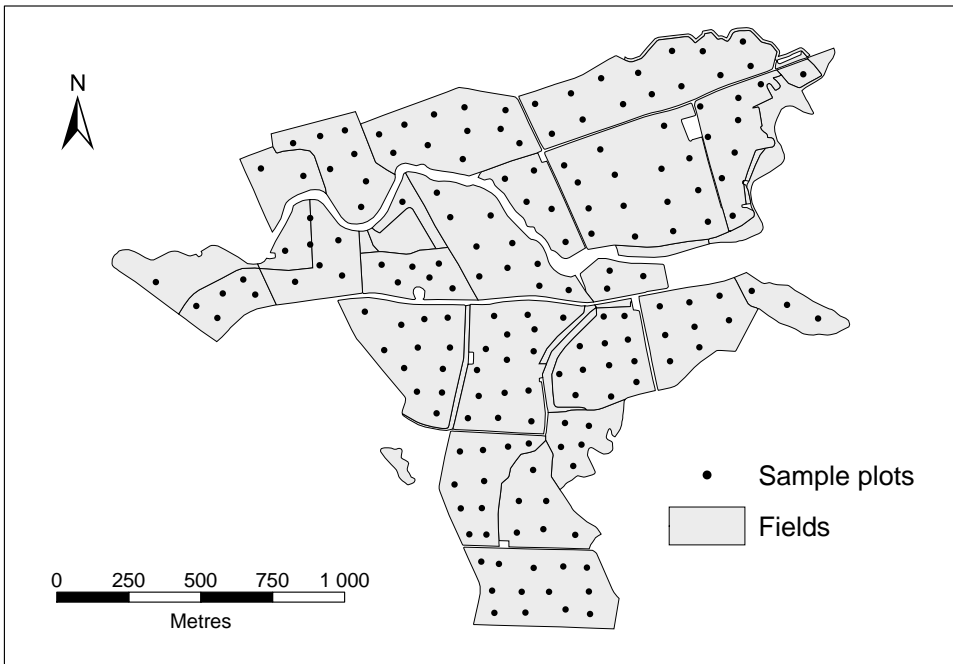


Fig. 1. Location of the fields and sample plots in the survey area.

## Geostatistical methods

The data were analysed using geostatistical methods to describe and interpret the spatial variation of soil texture fractions. Since the variogram is sensitive to departures from normality an exploratory data analysis using graphical plots, descriptive statistics and analysis of the distribution was done before modelling the spatial dependence. Furthermore, possible anisotropy was explored by computing the variograms in three directions, i.e. 0°, 60° and 120°, and plotting the semivariances in the same graph.

The exponential (Eq. 1) and spherical (Eq. 2) models with a nugget were fitted to determine the best fitting function based on the residual sum of squares. Variograms were computed and non-linear models fitted using VARIOGRAM- and NLIN-procedures of SAS for Windows release 8.2 (SAS 1999). The equation for the exponential model is

$$(1) \quad \gamma(h) = c_0 + c \left\{ 1 - \exp\left(-\frac{h}{r}\right) \right\}$$

where  $c_0$  is the nugget,  $c$  is sill variance and  $r$  is a distance parameter that defines the spatial extent of the model (Webster and Oliver 2001). This function approaches its sill asymptotically and so does not have a finite range. For practical purposes an effective range,  $a = 3r$ , is assigned to this model, which is usually the distance at which  $\gamma$  equals 95% of the sill variance (Webster and Oliver 2001). The spherical model has the function

$$(2) \quad \gamma(h) = \begin{cases} c_0 + c \left\{ \frac{3h}{2a} - \frac{1}{2} \left( \frac{h}{a} \right)^3 \right\} & \text{for } h \leq a \\ c_0 + c & \text{for } h > a \end{cases}$$

where  $c_0$  is the nugget,  $c$  is the sill as in equation 1 and  $a$  is range.

Kriging is only one technique among many for interpolating a variable from sample points,

but its advantage is that the estimates for a point or block are obtained with minimum variance (kriging variance) (Lark 2000). Therefore, the estimated values of the parameters of equations (1) and (2) were used to predict values at the nodes of a 20-m grid by ordinary punctual kriging, using the SAS software and the KRIGE2D procedure (SAS 1999). Cross validation was used to compare the goodness of the models in terms of predictions. Maps of the kriged predictions and standard errors were produced, using ArcMap GIS software from ESRI (Minami et al. 1999).

### Sampling density

The kriging variances depend on the configuration of the sampling points in relation to the target point or block and on the variogram, they do not depend on the observed values at those points. Therefore, if the variogram is known the kriging errors can be determined for any sampling configuration (Webster and Oliver 2001). Thus, the effect of sampling density was evaluated using the estimated variogram parameters of the spherical model (Eq. 2). Sample density was increased by adding new points to the map until the required sample size was achieved. Randomly selected new points were accepted if the minimum distance between any two points was at least 50 m. When the sampling density was 3.0 points per hectare, the minimum distance between any two points was 40 m. The randomization method was selected because it simulates quite well the choosing of points in practice. Randomization was repeated 3 times for each density.

## Results

### Exploratory data analysis

The particle size distribution of the soil samples from the survey area are dominated by the clay fraction, i.e. the average clay content was 52% (Table 2). The proportions of silt, fine sand and sand were 23%, 17% and 8%, respectively. The statistical distributions of clay and silt contents were near normal, but the fine sand and sand fractions were positively skewed. The latter reflects a few very sandy soils among the sample. After the transformation of fine sand and sand by the arcsine square-root, these values were normally distributed. No trend was observed based on graphical exploration of the clay data.

### Variograms and variogram models

The lag distance used in the variogram was 80 m, i.e. this was the shortest distance that could be used without decreasing the number of the observations and the reliability of the variogram. The spatial variation in the data is isotropic, because no clear differences could be found in the variograms computed in different directions. The parameters of the fitted, authorised variogram models are given in Table 3. Both models have a clear nugget effect, which was expected because of the coarse sampling intensity used in the soil survey. The nugget variance encompasses spatially dependent variation over distances less than the shortest lag, measurement error and any

Table 2. Summary statistics of soil texture components in the survey area.

| Soil fraction, % | Mean | Min | Max | Variance | Standard deviation | Skewness |
|------------------|------|-----|-----|----------|--------------------|----------|
| Clay             | 52   | 11  | 78  | 159.74   | 12.64              | -0.6899  |
| Silt             | 23   | 6   | 36  | 57.84    | 7.61               | -0.3161  |
| Fine sand        | 17   | 4   | 57  | 137.64   | 11.73              | 1.6228   |
| Sand             | 8    | 0   | 43  | 39.76    | 6.31               | 2.6997   |

Table 3. Parameters of the fitted variogram models.

| Variogram model | Nugget | Sill   | Range  |
|-----------------|--------|--------|--------|
| Exponential     | 41.87  | 139.50 | 297.10 |
| Spherical       | 59.79  | 107.80 | 659.60 |

purely random variation (Oliver and Khayrat 2001). Both models are bounded and appear quite similar (Fig. 2); however, there is a difference in the goodness of fit of the first three lags. This is important, because kriging is local and near points carry more weight than more distant points. The lack-of-fit statistic ( $MS_{\text{error}} 43.96$ ) confirms the visual appraisal that the exponential model was poorer. The spherical model had the smaller error ( $MS_{\text{error}} 33.72$ ) and it was chosen for predicting the clay fraction and for determining the optimal sample density. The range of spatial dependence was 660 m for the spherical model (Table 3). The effective range, i.e. 95% of the sill variance attained, for the exponential model was 891.3 m.

### Predictions and standard errors

Cross validation showed that the spherical model was better than the exponential model in terms of mean error (0.05 vs. 0.09), mean squared error (102.7 vs. 104.9) and mean squared deviation ratio (1.09 vs. 1.11). Graphical plots of the cross validated residuals, true values and locations did not reveal any shortcomings, e.g. areas, where the models were not compatible with the data. The kriged estimates of the clay percentage using the parameters of the spherical and exponential variogram models show the same pattern of variation in clay content for the study area (Fig. 3). The spherical model produced slightly smoother variation than the exponential function. The latter produced a map with a little more detail, i.e. some variation at shorter distances, because the estimated nugget effect was smaller than for the spherical model. The results suggest that the medium-range pattern had been

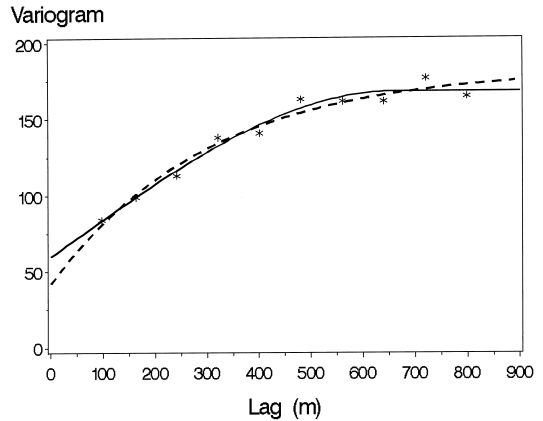


Fig. 2. The spherical (solid line) and exponential (dash line) variogram models (solid line) and the sample variogram (stars) for clay fraction data.

resolved by the sampling, but the short-range variation has not.

According to the Finnish classification, the predicted values were divided into 3 classes based on the clay percentage. Areas with clay content below 30%, between 30% and 60% and over 60% belong to non-clay, clay and heavy clay zones, respectively. Most of the fields of the survey area (Fig. 3) contain 2 or even 3 of the classes defined above, i.e. the soil of a field can vary from non-clay to heavy clay types. This classified surface can be used to delineate the preliminary management zones, for example by digitising or vectorising the boundaries between the classes.

In general, the kriging errors were consistently between 8% and 10%, and the median of error was 9.3% and 9.0% for the spherical and exponential models, respectively (Fig. 4). The largest errors were in the northwestern and eastern edges of the survey area where they exceeded 12%, but generally they were smaller than 10% in over 95% of the study area. Only 5% of all errors were smaller than 9.0% and 8.4%, according to the spherical and exponential model, respectively (Table 4). Hence, the patterns in the spherical and exponential model error maps were very similar, but the errors in the exponential model were slightly smaller (Fig. 4). Again, the

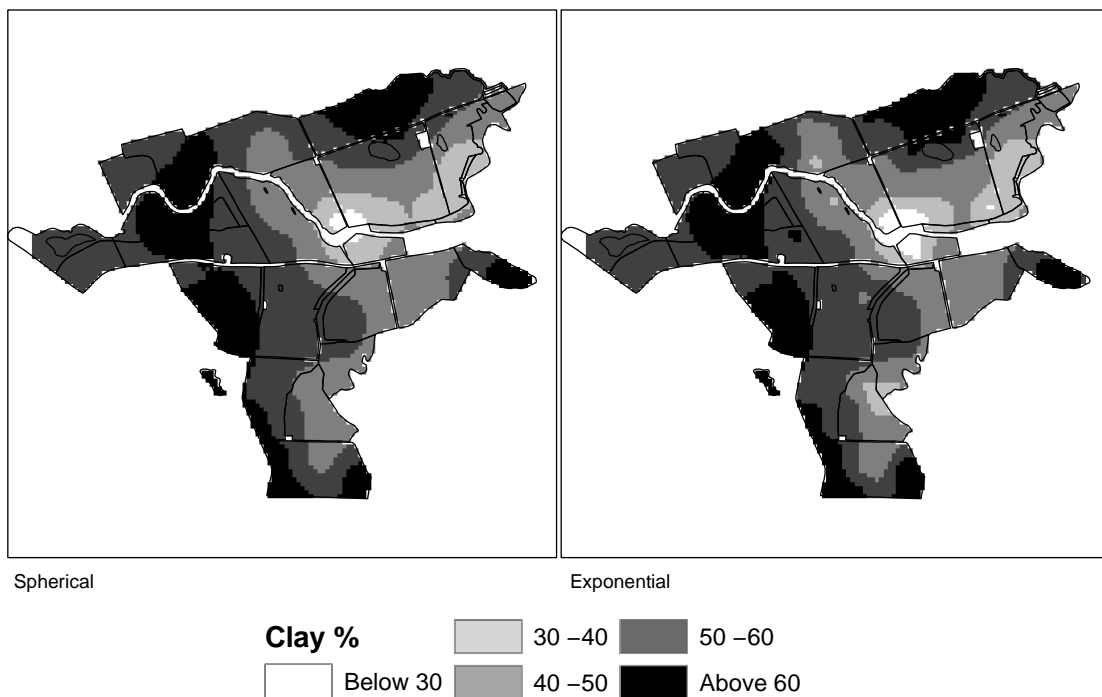


Fig. 3. Kriging estimates of the clay percentage when parameters of spherical and exponential variogram models were used in prediction.

main reason for this is the nugget effect, which is smaller in the exponential model than in spherical model. However, from the practical point of view the difference between error levels is so small that it is of no major consequence.

As the number of samples increased, the final the mean standard error was 8.7 and 8.0 for the spherical and exponential models, respectively.

### Effect of sampling density

The standard error of the kriged estimates decreased only slightly when the density of the sampling was increased (Table 4). However, the range in errors became smaller and the effect was strongest at the edges of the map and where there were no samples originally. In general, the exponential model has smaller errors because the estimated nugget effect was smaller than that for the spherical model. The mean standard error was 9.3 for the spherical model (Table 4). The difference between models increased when the den-

### Discussion

This study shows how management zones can be delineated based on geostatistical analysis. In mineral soils the clay content may be the single most important soil characteristic determining those soil properties important in management. The clay content is inversely related to the content of coarser textural fractions and to a large extent determines the soil type in fine- and medium-textured mineral soils that dominate the fields studied. The clay content influences many physical properties such as water-holding capac-

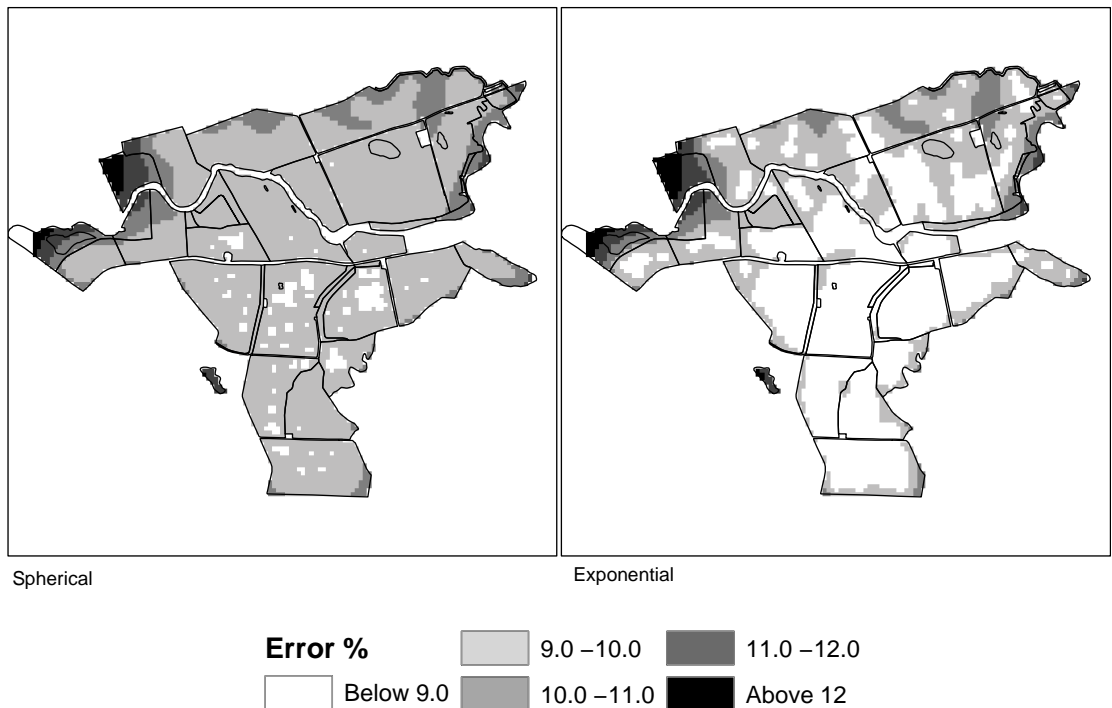


Fig. 4. Kriging standard errors in clay percentage when parameters of spherical and exponential variogram models were used in prediction.

Table 4. Distribution of standard error of kriging estimates for 6 different densities. The original density was 0.7 points per hectare. Columns 5%, 25%, 50%, 75% and 95% include 5%, 25%, 50%, 75% and 95% percentiles of error distribution, respectively.

| Model and density     | 5%  | 25%  | 50%  | 75%  | 95%  |
|-----------------------|-----|------|------|------|------|
| Spherical             |     |      |      |      |      |
| 0.25 ha <sup>-1</sup> | 9.6 | 10.0 | 10.3 | 10.9 | 11.9 |
| 0.50 ha <sup>-1</sup> | 9.1 | 9.4  | 9.7  | 9.9  | 10.6 |
| 0.70 ha <sup>-1</sup> | 9.0 | 9.2  | 9.3  | 9.6  | 9.9  |
| 1.50 ha <sup>-1</sup> | 8.8 | 8.9  | 9.0  | 9.1  | 9.4  |
| 3.00 ha <sup>-1</sup> | 8.6 | 8.6  | 8.7  | 8.8  | 9.0  |

ity. It also influences the concentration of cationic nutrients and cation exchange capacity. Clay is the principal variable used to interpret the results of soil testing. The interpretations are different on clayey, silty and sandy soil (Viljavuuspalvelu 2000) resulting in different recommendations of P, K and Mg in clayey and non-clayey

soils. The pH buffering and liming requirements correlate with clay content. Therefore, different liming rates are recommended in Finland for soils with clay contents of over 60%, 30–60%, 15–30% and under 15% (Viljavuuspalvelu 2000). The zones delineated in this study on the basis of clay content can thus be used, when amended

with chemical soil test data, as management zones for fertilizer applications and liming.

The textural variation seen in several experimental areas suggests that sub-division of fields into management zones is reasonable. The large variation in clay content and several other soil characteristics within a field in the study by Jokinen (1983) resulted in large standard deviations of the mean values. Therefore to determine the mean value accurately, a large sample size would be required, but even then, because of the spatial variation, the mean for the entire field would be unrepresentative. Identification of zones on the basis of clay content might enable targeted sampling within each zone for chemical analyses, which might reduce the sample sizes and the cost of analyses substantially.

Interpolation is appropriate only if the characteristic studied varies continuously and the sample data are spatially dependent or correlated (Oliver and Khayrat 2001). The variogram for clay showed that with 150 samples and an average separation between samples of 145 m the spatial variation of the clay fraction could be modelled satisfactorily. This supports the finding of Webster and Oliver (1992) that at least 100 samples are required to model the variogram.

The nugget effect present, i.e. 55% of the sill variance, showed that there was unresolved variation at distances less than the sampling interval. The short-range variation in clay percentage could be identified with a more intensive sampling, for example Saldana et al. (1998) used a 10-m sampling interval. In the present study the nugget variance was so large that increasing the sampling interval from 145 m to 33 m by random simulation method did not decrease greatly the standard error of the predicted clay content. This conclusion is based on assumption that the estimated variogram is adequate for all densities. Because of lack of the observations at small distances, the estimated nugget variance is probably too large. Therefore, the effect of increasing the sampling density was underestimated. However, it is noteworthy that at the edges of the research area the standard error decreased more with increased sampling density

than in the central areas. In large field areas, more accurate estimates for clay content could be obtained if farms bordering each other could cooperate by combining the data for the geostatistical delineation of management zones.

The kriging variance can be used to estimate the reliability of the predictions, bearing in mind that it depends on how accurately the variation is presented by the chosen spatial model (Webster and Oliver 2001). There were differences between the nugget variances of the examined models, but there is no evidence of which estimate was the best. If the nugget variance is overestimated then the punctual kriging variances will also be overestimated, in which case the kriged estimates would be more reliable than they appear to be (Webster and Oliver 2001). Thus, it is possible that the kriging errors were overestimated by the spherical model in the present study.

Fertilizer recommendations in Finland depend on: 1) the crop, 2) expected yield (nutrient uptake), 3) soil texture and humus content, and 4) the nutrient content of the soil. Soil nutrients (except N) are analysed accurately in the laboratory. These chemical data are interpreted according to soil texture, which is currently determined by finger assessment in routine soil testing. It was recently shown (Peltovuori 1999) that estimates of soil texture and humus content, obtained from various Finnish laboratories, contained errors that occasionally resulted in deviations of  $\pm 10 \text{ kg ha}^{-1}$  from the correct recommendations for phosphorus fertilisation. Fine-tuning of K and also N fertilizer applications are equally dependent on soil texture. Incorrect estimates of soil texture and humus content can undermine the accurate results of chemical analyses on soil nutrients. This shortcoming also prevents precision farming from being practiced to the full extent of its capabilities.

Reliable data on soil texture and humus content are required by precision farming to delineate management zones and for accurate fertilizer application. Obtaining these data is expensive because it seems that finger assessment should be replaced with more accurate laboratory de-



termination. This aim can be achieved only by less expensive methods for these determinations, particularly for texture. However, texture is a permanent soil characteristic and needs to be determined only once, and organic matter content changes slowly. Investment in these analyses could be counterbalanced by reducing the number of samples taken in each management

zone for analysis in repeated routine soil tests. Responding to soil variation at any scale requires suitable systems for processing the data and generating from it information to assist in making decisions (Lark and Bolam 1997). Geostatistical methods with GIS provide tools for identifying the management zones of large field areas.

## References

- Aaltonen, V., Arnio, B., Hyypä, E., Kaitera, P., Keso, L., Kivinen, E., Kokkonen, P., Kotilainen, M., Sauramo, M., Tuorila, P. & Vuorinen, J. 1949. Maaperäsanaston ja maalajien luokituksen tarkistus v. 1949. Summary: A critical review of soil terminology and soil classification in Finland in the year 1949. *Journal of the Scientific Agricultural Society of Finland* 21: 37–66.
- Bocchi, S., Castrignano, A., Fornaro, F. & Maggiore, T. 2000. Application of factorial kriging for mapping soil variation at field scale. *European Journal of Agronomy* 13: 295–308.
- Brooker, P. 2001. Modelling spatial variability using soil profiles in the Riverland of South Australia. *Environment International* 27: 121–126.
- Elonen, P. 1970. Particle-size analysis of soil. *Acta Agraria Fennica* 122: 1–122.
- Haapala, H. 1995. Position dependent control (PDC) of plant production. *Agricultural Science in Finland* 4: 239–350.
- Jokinen, R. 1983. Variability of topsoil properties at the southern coast of Finland and the number of soil samples needed for the estimation of soil properties. *Journal of the Scientific Agricultural Society of Finland* 55: 109–117.
- Lark, R. & Bolam, H. 1977. Uncertainty in prediction and interpretation of spatially variable data on soils. *Geoderma* 77: 263–282.
- Lark, R. 2000. Designing sampling grids from imprecise information on soil variability, an approach based on the fuzzy kriging variance. *Geoderma* 98: 35–59.
- Minami, M., Sakala, J. & Wrightsell, J. 1999. *Using Arc-Map*. Environmental Systems Research Institute, Inc. 560 p.
- Oliver, M. & Khayrat, A. 2001. A geostatistical investigation of the spatial variation of radon in soil. *Computers & Geosciences* 27: 939–957.
- Peltovuori, T. 1999. Precision of commercial soil testing practise for phosphorus fertilizer recommendations in Finland. *Agricultural and Food Science in Finland* 8: 299–308.
- Saldana, A., Stein, A. & Zinck, J. 1998. Spatial variability of soil properties at different scales within three terraces of the Henares River (Spain). *Catena* 33: 139–153.
- SAS Institute Inc. 1999. *SAS/STAT Users Guide. Version 8*. Cary, NC: SAS Institute Inc.
- Soil Survey Staff 1993. Soil survey manual. *USDA Handbook* 18. 2nd Ed. Government Printing Office, Washington, DC. 437 p.
- Utset, A., Lopez, T. & Diaz, M. 2000. A comparison of soil maps, kriging and a combined method for spatially predicting bulk density and field capacity of ferralsols in the Havana-Matanzas Plain. *Geoderma* 96: 199–213.
- Viljavuuspalvelu 2000. *Viljavuustutkimuksen tulkinta peltoviljelyssä*. (Interpretation of soil test results in arable cropping). Viljavuuspalvelu Oy. 32 p. (in Finnish).
- Webster, R. & Oliver, M. 1992. Sample adequately to estimate variograms of soil properties. *Journal of Soil Science* 43: 177–192.
- Webster, R. & Oliver, M. 2001. *Geostatistics for environmental scientists*. John Wiley & sons, Ltd. 271 p.

## SELOSTUS

### Viljelymaiden savespitoisuuden alueellistaminen geostatistiikan ja pistemäisen tiedon avulla

Ari Talkkari, Lauri Jauhiainen ja Markku Yli-Halla  
*MTT (Maa- ja elintarviketalouden tutkimuskeskus)*

Täsmäviljelyssä pellot jaetaan käsittelyvyöhykkeisiin maaperän ominaisuuksien alueellisen vaihtelun mukaan. Tämän tutkimuksen tavoitteena oli analysoida viljelymaan savespitoisuuden alueellista vaihtelua viljavuusanalyysiaineiston perusteella sekä rajata käsittelyvyöhykkeitä geostatistiikan ja paikkatietotekniikan menetelmiä käyttäen.

Tässä tutkimuksessa 218 peltohehtaarin alue jaettiin vyöhykkeisiin savespitoisuuden perusteella, koska se on keskeinen maaperän ominaisuus viljelytoimenpiteiden kannalta. Savespitoisuuden alueellistamiseen oli käytettävissä 150 näytepisteen tiedot. Näistä laadittiin empiirinen variogrammi, johon sovitettiin kaksi erilaista mallia. Malleja käytettiin kriging-interpoloinnissa, jossa savespitoisuudelle laskettiin jatkuva pinta 20 m:n resoluutiolla. Menetelmällä pystyttiin mallintamaan ja ennustamaan savespitoisuuden keskipitkän matkan vaihtelu, mutta lyhyen matkan vaihtelua ei havaittu, koska lähtöaineis-

ton pisteverkko oli melko harva (0,7 kpl/ha). Lisäksi tutkittiin näytteenottoiheyden vaikutusta ennustevirheeseen simuloimalla aineistoon uusia havaintopisteitä useilla eri tiheyksillä. Ennustevirhe pieneni näytteenottoiheyden kasvaessa 0,7 pisteestä 3 pisteeseen hehtaarilla vain 0,5–1 %-yksikköä. Tämä johdettiin aineiston selittämättömän vaihtelun suuresta osuudesta (n. 36 % kokonaisvaihtelusta).

Tutkittu peltoalue jaettiin geostatistisen analyysin perusteella kolmenlaisiin alueisiin, joiden savespitoisuudet olivat 1) alle 30 %, 2) 30–60 % ja 3) yli 60 %. Täsmäviljelyn käsittelyvyöhykkeiden määrittämisessä tarvitaan savespitoisuuden lisäksi tietoa muista maalajitteista, ravinteista ja orgaanisen aineksen määrästä. Tässä esitetyllä menetelmällä käsittelyvyöhykkeet voidaan tuottaa numeerisesti, jolloin niitä pystytään käyttämään tietokonepohjaisissa täsmäviljelyjärjestelmissä.