ACCADEMIA DEI GEORGOFILI

Ce. S. I. A. - Centro di Studio per l'applicazione dell'Informatica in Agricoltura
C.N.R. - I.A.T.A.

7th ICCTA - International Congress for Computer Technology in Agriculture

"COMPUTER TECHNOLOGY IN AGRICULTURAL MANAGEMENT AND RISK PREVENTION"

Florence, 15th - 18th November 1998

C. Conese e M.A. Falchi

A. N. A. - Accademia Nazionale di Agricoltura
DLG - Deutsche Landwirtschafts - Geselischaft
RASE - Royal Agricultural Society of England
SAF - Société des Agriculteurs de France

"ACCOUNTING FOR LOCAL UNCERTAINTY IN AGRICULTURAL MANAGEMENT DECISION MAKING"

Gabriele Buttafuoco[1], Annamaria Castrignanò[2] and Matteo Stelluti[2]
[1] CNR - Istituto di Ecologia e Idrologia Forestale
Via Cavour - 87030 Rende (CS) - Italy
phone +39-0984-466036, fax +39-0984-466052
e-mail: buttafuo@area.cs.cnr.it

[2] MiPA - Istituto Sperimentale Agronomico
Via Celso Ulpiani, 5 - 70125 Bari
phone +39-080-5475024, fax +39-080-5475023
e-mail: agrobari@interbusiness.it

Abstract

Many soil surveys lead to important decision making in agriculture, such as delineation of areas targeted for fertilization or some remedial treatment. Such decisions are often based on critical values of the concentrations of nutrient or salt in the soil. If the estimates are less or more than specified thresholds, farmers are advised to act. But such estimates are usually affected by large uncertainty, arising from sampling, modelling and interpolation, which must be quantified to allow an evaluation of the risk involved in any decision. Geostatistics allows to assess such uncertainty through the determination of a conditional cumulative distribution function (ccdf) of the unknown attribute value. This paper considers the problem of modelling uncertainty about the value of an attribute at any unvisited location. The uncertainty is modelled through the ccdf conditional to the local information and gives the probability that the unknown is not greater than any given threshold. The paper describes a non-parametric approach to estimate the uncertainty, called "indicator kriging" (Journel, 1983), based on the interpretation of the conditional probability as the conditional expectation of an indicator random variable.

A soil survey data set of a 18000 ha-area in southern Italy was used as a support for presenting a potential application of modern Geostatistics to agricultural management decision making. Accounting of "soft" information as provided by a geological and soil maps is shown to reduce the uncertainty.

1. INTRODUCTION

In order to produce economic yields of agricultural crops soil needs to have sufficient contents of the major plant nutrients. A common approach is to delineate areas where nutrient deficiencies limit yields and attempt to remedy the shortcoming of the soil by applying fertilizer. Farmers generally apply fertilizers on the basis of critical values of the nutrient concentrations: when the nutrient is less than a given threshold, the farmer is advised to apply fertilizer. The choice of acting or not acting is then based on nutrient

concentration estimates from soil samples. Unfortunately, such estimates are always subject to error, coming from different sources, such as inherent soil variability, sampling, laboratory analysis and interpolation technique. Moreover, the modern trend to reduce and optimise the use of agro-chemical in the production of arable crops urges farmer to be fully conscious of his choices in order to reduce environmental impact and input costs. Therefore, he needs to know the probability that the actual values of nutrient concentration fall short of the critical values and the risks he runs in taking a given decision.

Traditional techniques do not provide any measure of the reliability of the estimates and thus any assessment of the decision-making risks is not possible. The main advantage of geostatistical techniques, essentially ordinary kriging, is to produce an estimation variance related to each estimate. The main disadvantage is, unless spatial error distribution is Gaussian, estimation variance does not provide probability intervals of error distribution required for risk assessment. Most interpolation algorithms, including kriging, that allow to characterize uncertainty, are parametric, in the sense that a model of error distribution is assumed. Such a model is very often normal or at least symmetric, which is congenial to characterize the distribution of measurement errors in the highly controlled environment of a laboratory but it is generally not well suited to spatial interpolation errors.

The newly developed non-parametric geostatistical technique, called indicator kriging (Journel, 1983) puts as priority, not the derivation of the optimal estimator, but the modelling of uncertainty, that takes the form of a probability distribution.

The major advantage of the indicator approach is the ability to take into account qualitative data in addition to the measurements of the variable, e.g. geological and soil maps. Both "hard" (measurements) and "soft" (qualitative) information are coded in the same way, as local probability values and thus can be processed together, by cokriging, independently of their origin (Castrignanò et al., 1997).

In this paper is presented an application of non-parametric geostatistics to agricultural fertilization using phosphorus concentration as hard information and geological and soil maps as soft information.

2.0 MATERIALS AND METHODS

2.1 INDICATOR APPROACH

Consider the following information available over the study area:

- values of the continuous variable [phosphorus expressed in parts per million (ppm; S.I. units = mg kg^{-1}], Z [1], at n locations $\mathbf{u}_\alpha$, $z(\mathbf{u}_\alpha)$, α=1, 2, ..., n;
- rock type, r($\mathbf{u}$) at all locations $\mathbf{u}$ within the area, with r_1, r_2, r_3, r_4, being the set of rock types as read from a geological map;
- soil type s($\mathbf{u}_\alpha$) at all locations within the area, with s_1, s_2, s_3, s_4, being the set of soil type as read from a soil map.

The z-values are hard in the sense that they are direct measurements of phosphorus content. On the contrary, each rock or soil type provides only indirect (soft) information about the value of the variable Z. Under a prior decision of stationarity (statistical

[1] The capital letter is used to indicate the random variable; the small letter for a variable measurement.

homogeneity) the soft information consists of a distribution of *z*-values given the rock type or soil type prevailing. Using both hard and soft data the approach is aimed at assessing the probability that the value of *z* at any unsampled site $\mathbf{u}_0$ is not greater than a given threshold z_k. Indicating with *F* the conditional cumulative distribution function (ccdf) of the variable *Z*, it results:

$$F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3)) = \text{Prob}\{Z(\mathbf{u}_0) \leq z_k \mid (n_1 + n_2 + n_3)\} \quad [1]$$

where the notation $\mid (n_1 + n_2 + n_3)$ expresses the conditioning to n_1 hard data $\{z(\mathbf{u}_\alpha); \alpha=1, 2, ..., n_1\}$, n_2 soft data $\{r(\mathbf{u}_\alpha); \alpha=1, 2, ..., n_2\}$ and n_3 soft data $\{s(\mathbf{u}_\alpha); \alpha=1, 2, ..., n_3\}$ retained in the neighbourhood of $\mathbf{u}_0$.

In this approach the probability distribution is regarded as the conditional expectation of the indicator random variable $I(\mathbf{u}_0; z_k)$, given the information set $(n_1 + n_2 + n_3)$:

$$F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3)) = E\{I(\mathbf{u}_0; z_k) \mid (n_1 + n_2 + n_3)\} \quad [2]$$

with

$$I(u_0; z_k) = \begin{cases} 1 & \text{if } Z(u_0) \leq z_k \\ 0 & \text{otherwise} \end{cases} \quad [3]$$

According to the projection theorem (Luenberger, 1969), the least square estimate of the indicator $I(\mathbf{u_0}; z_k)$ is also the least square estimate of its conditional expectation. Thus the ccdf $F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$ can be estimated by (co)kriging (Journel and Huijbregts, 1978) the indicator $I(\mathbf{u_0}; z_k)$ using indicator transform of hard and soft data.

2.2 INDICATOR CODING OF INFORMATION

The indicator approach requires a preliminary coding of the hard and soft data into local hard and soft "prior" probabilities:

$$\text{Prob}\{Z(\mathbf{u}) \leq z_k \mid \text{local information at } \mathbf{u}\} \quad [4]$$

The term "local prior" means that the probability in equation [4] originates from the hard or soft information at location **u**, prior to any updating based on neighbouring data. The final target of this approach is the updating this local prior probability in the posterior probability ([2]). Thus, the prior information can take one of following form:

- local hard indicator data $i(\mathbf{u}_0; z_k)$, with binary indicators defined by [3];
- local soft indicator data $y_r(\mathbf{u}_0; r_k)$ and $y_s(\mathbf{u}_0; s_k)$ assuming values within the interval [0,1]. They are derived by calibration curves of deficiencies from those locations where both z-value and rock type or soil type, respectively, are known.

The local hard and soft indicator data, $i(\mathbf{u}_0; z_k)$, $y_r(\mathbf{u}_0; r_k)$ and $y_s(\mathbf{u}_0; s_k)$ are interpreted as realisations of three correlated random functions , $I(\mathbf{u}_0; z_k)$, $Y_r(\mathbf{u}_0; r_k)$ and $Y_s(\mathbf{u}_0; s_k)$ that can be combined by using cokring where $I(\mathbf{u}_0; z_k)$ is the primary variable and $Y_r(\mathbf{u}_0; r_k)$ and $Y_s(\mathbf{u}_0; s_k)$ the secondary variables.

2.3 RISK ASSESSMENT

Availability of either model semivariogram of soil attribute Z or of ccdf model

$F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$ for each location $\mathbf{u}_0$ within the study area allows contouring of isopleth curves of:

1. The optimal estimates $z^*(\mathbf{u}_0)$ by using ordinary kriging. These estimates are derived indipendently of the uncertainty model $F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$.
2. Probabilities that the actual unknown value $Z(\mathbf{u}_0)$ does not exceed a given critical threshold, such as:
$$\text{Prob}\{Z(\mathbf{u}_0) \le z_k \mid (n_1 + n_2 + n_3)\} = F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$$
3. The risk $\alpha(\mathbf{u}_0)$ of declaring a location "well-supplied" by plant nutrient on the basis of the estimate $z^*(\mathbf{u}_0)$ when actually $Z(\mathbf{u}_0) \le z_k$:

$\alpha(\mathbf{u}_0) = \text{Prob}\{Z(\mathbf{u}_0) \le z_k \mid z^*(\mathbf{u_0}) > z_k \text{ and } (n_1 + n_2 + n_3)\} = F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$,
for all $\mathbf{u}_0$ such that $z^*(\mathbf{u}_0) > z_k$.

4. The risk $\beta(\mathbf{u_0})$ of declaring "deficient" a location when in fact it is not: $Z(\mathbf{u}_0) > z_k$:

$\beta(\mathbf{u}_0) = \text{Prob}\{Z(\mathbf{u}_0) > z_k \mid z^*(\mathbf{u_0}) < z_k \text{ and } (n_1 + n_2 + n_3)\} = 1 - F(\mathbf{u}_0; z_k \mid (n_1 + n_2 + n_3))$,
for all $\mathbf{u}_0$ such that $z^*(\mathbf{u}_0) \le z_k$.

2.4 CASE STUDY

To investigate the method's feasibility and potential in agriculture we applied it in a region of southern Italy. The aim of the research was to map the soil fertility in detail.

The study area[2], covering 18000 ha (figure 1), is in an extensive sedimentation watershed of the Crati River in the northern extremity of Calabro-Peloritano arc, is bordered to north by the Pollino mountains, to east by Sila massif, to west and to south by coastal Tyrrhenian Chain.

Figure 1: Localization of the study area.

It has been hypothesised that the formation of the Crati River watershed happened in the Pliocene period at the same time of the emergence Sila massif and of the Pollino mountains (Colella, 1988). According to geological map (CASMEZ, 1967), there are four rock types: Pleistocene sediments, Pliocene sediments, Miocene sediments and Holocene sediments[3].

Following the Keys of Soil Taxonomy (Soil Survey Staff, 1992) there are four soil orders (the order is the highest category in Soil Taxonomy): Alfisols, Entisols, Inceptisols, Mollisols (ARSSA, 1996).

The sampling has concerned particularly the areas of more agricultural interest. The upper 40 cm of soil were sampled by collecting randomly cores at 100 sites (figure 2). The soil samples were analysed for texture, pH and the available major nutrients, potassium

[2]The X and Y values correspond to the UTM system values by adding 600000 m to X and 4300000 m to Y.
[3]The geologic classification is stratigraphic, not lithological. The term *rock type* should be understood as *stratigraphic class.*

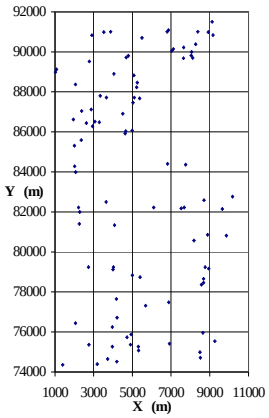

Figure 2: Localization of the100 sample sites.

and phosphorus, among them; here we will present the results of geostatistical analysis only for phosphorus.

3.0 GENERAL PROCEDURE

The histogram of the 100 phosphorus (P) data showed a mean of 2.31 ppm, a strong positive skewness (3.8) with a very large coefficient of variation equals to 91.4 % and a long tail: 25 % of the data range from 2.6 ppm to a maximum of 16.6 ppm. This distribution was clearly not normal.

A nested variogram model was fitted on the natural logarithms of the original variable by weighted least-squares including a nugget effect (0.28) and a spherical structure with range equals to 3000 m and sill of 0.5.

The contents of phosphorus at the nodes of a 100 m x 100 m - grid were estimated by ordinary kriging. The grid values were then contoured using the program Surfer (Version 6.4) (fig. 3).

As it is clear from the figure, most of the area is characterized by phosphorus contents less than mean value, except two large zones to north and in the south middle.

At the time the samples were collected and analysed a critical value of readily soluble phosphorus would have been 2.6 ppm for cereal crops and farmers would have been advised to apply phosphorus fertilizer where there was less than this.

Following indicator approach, soil phosphorus content data constituted hard data and were coded in to indicator data, according to the definition [3] using the critical value of 2.6 ppm as threshold. The soft information was calibrated estimating the marginal distributions of phosphorus content within each rock type and soil type; the cumulative distributions of deficiency are listed in table 1.

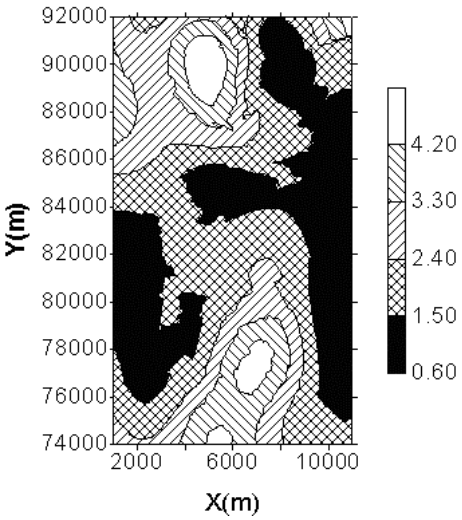

Figure 3: Map of the krigged estimates of available phosphorus in soil.

Table 1 - Cumulative distribution (F) table of P-values less or equal than 2.6 ppm for different rock and soil types.

Rock Type	F	Soil Type	F
Pleistocene sediments	0.67	Alfisols	1.00
Pliocene sediments	0.86	Entisols	0.56
Miocene sediments	-	Inceptisols	0.75
Holocene sediments	0.60	Mollisols	0.89

As it results from the table, deficiencies lie mainly in Pliocene sediments, which is also the most representative rock type of the study area, and in Alfisols and Mollisols. However, these two kinds of soil are the least present in the area. In general, all of the rock and soil types are characterized by large proportion of phosphorus deficient samples.

To assess the accuracy of either rock map information or soil map information in predicting phosphorus content, the coefficient $B(z_k)$ was computed as the difference between the two conditional expectations (Goovaerts and Journel, 1995):

$$B(z_k) = m^{(1)}(z_k) - m^{(0)}(z_k) \in [-1, 1] \qquad [5]$$

where the quantity $m^{(1)}(z_k)$ is estimated by the arithmetic average of the soft indicator data for each type of information (rock type and soil type) at location where hard information is equal to 1 and, conversely, $m^{(0)}(z_k)$ where hard information is equal to 0.

$B(z_k)$ then measures the ability of the soft information to separate the two cases: soil phosphorus content less and greater than the threshold value of 2.6 ppm. $B(z_k)$ is then an accuracy index for the soft information with best being 1. In our case B was equal to 0.10 for rock type and 0.01 for soil type. The rather low values, especially for soil type, shows the small portion of the phosphorus content variance explained by rock type and soil type. For the selected threshold all the direct variograms of hard and soft indicator data (rock and soil) and their corresponding cross-variograms were computed and fitted to a linear coregionalization model (Goovaerts, 1997), including three spatial structures:

1. a nugget effect;
2. a spherical structure with range equals to 2000 m;
3. a linear structure up to 3000 m distance.

The semivariogram parameters, forming the coregionalization matrices, are reported in table 2, for each spatial scale.

Table 2 - Coregionalization matrices corresponding to the three spatial structures of the Linear Coregionalization Model.

			Hard data	**soft data**	
			P	**rock**	**soil**
Nugget Effect	**Hard data**	**P**	1.18E-01	-7.87E-03	1.01E-02
	Soft data	**rock**		3.83E-03	-6.05E-03
		soil			9.62E-03
			Hard data	**soft data**	
			P	**rock**	**soil**
spherical model	**Hard data**	**P**	7.67E-02	2.60E-02	-1.38E-02
	Soft data	**rock**		1.96E-02	-1.10E-04
		soil			5.95E-02
			Hard data	**soft data**	
			P	**rock**	**soil**
linear model	**Hard data**	**P**	1.86E-09	5.43E-09	1.97E-09
	Soft data	**rock**		1.59E-08	5.76E-09
		soil			2.09E-09

The conditional probabilities that the true phosphorus concentrations were less than 2.6 ppm were estimated by using ordinary cokriging of the hard and soft variables and then contoured. Figure 4 shows the results.

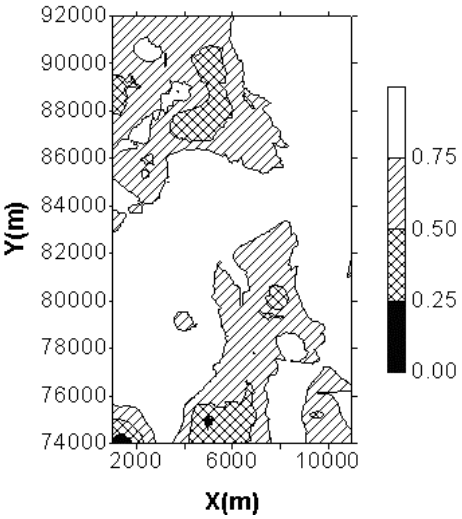

Figure 4: Conditional probability map of the available phosphorus in soil is less or equal to 2.6 ppm.

The hatched and white areas are where the probability that the estimates are less than the critical threshold is greater than 50 %. Clearly the farm manager would be well advised to fertilize in the hatched and mainly in the white regions which cover most of study area. Even outside the hatched and white regions there is a not negligible probability (less then 50 %) that the true concentration falls short of the critical value. The farmer might be willing to risk a diminished yield for lack of phosphorus in such regions or, on the contrary, willing to make a fertilization, being aware that a nutrient such as phosphorus is not very mobile and then that which is not taken up by the crop in any one year remains in the soil for succeeding crops. This is a personal choice, but farmer can well be assisted in his decision-making by the knowledge of the estimated risk α and risk β. The risk α of a false positive, i.e. the risk of declaring well-supplied a location when it is not, is contoured in figure 5. The darker zones (in the north-west corner and in the south middle) are where the probability of overestimation is greater than 50 %. The impact of the risk α can be easily evaluated because it is linked to the yield loss for not having fertilized.

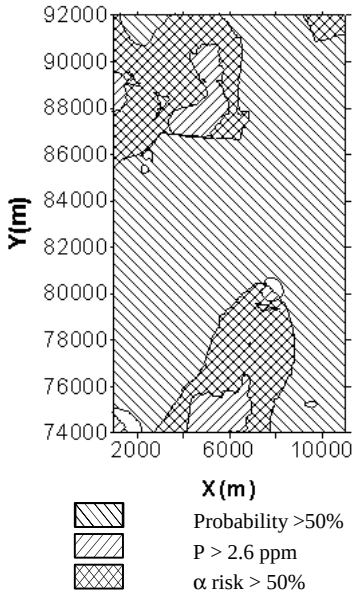

Figure 5: Contoured map of the risk α of a false positive; i.e. declaring wrongly that a location is well - P supplied.

The risk β of a false negative, i.e. the risk of fertilizing unduly, is contoured in figure 6. The zones where the probability of $P \leq 2.6$ ppm is greater than 50 % are very small and well delimited (darker zones). In this case evaluating the impact of the risk β is very likely linked to the cost of clearing unduly only, but does not involve such non-monetary notions as health hazards or environmental quality.

Deciding on the critical value and on the balance of the two risks α and β is a clearly political decision which falls well beyond the realm of geostatistics. Nevertheless geostatistics can assist in decision-making management by ranking the areas targeted for fertilization using a tested function of the assessed impact. Moreover, the joint availability of maps such as those of fig. 3 (phosphorus estimate), figure 4 (probability of deficiency), fig. 5 (risk α) and fig. 6 (risk β) allows also to make decisions concerning additional sampling.

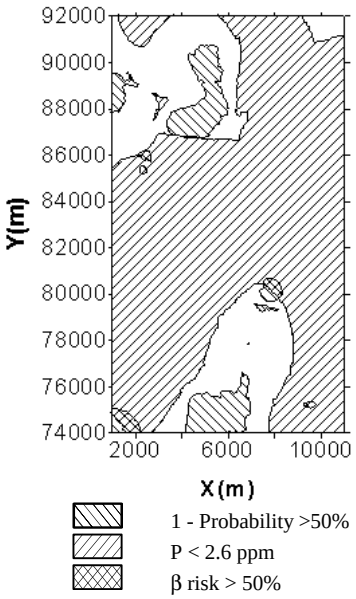

Figure 6: Contoured map of the risk β of a false negative; i.e. declaring wrongly that a location is P deficient.

Additional sampling should be considered in zones within the study area with high misclassification risks α and β, rather than areas with low estimated values (≤ 2.6 ppm). First candidates for additional sampling might not be zones declared highly phosphorus deficient on the basis of the estimation but rather zones where the risk α is high. It needs to note, however, that the risk α depends on both the uncertainty estimation (indicator approach) and on the algorithm used to determine phosphorus content (ordinary kriging).

4. CONCLUSIONS

In this paper we have described an approach where the available information, both hard data and soft information under the form of prior probability distribution, is coded as a series of indicator variables. The non-parametric method has yielded, at each location, a probability distribution of no exceedance of a critical threshold for soil phosphorus content. These probability distributions together with the kriging estimates of phosphorus have allowed to estimate the risk α of a false positive and the risk β of a false negative.

The approach allowed also an assessment of the need additional sampling and the ranking of zones as candidates for fertilization. It then enables advisers and farmers to be more conscious of their responsibilities and should be evaluated more widely in agriculture.

5. REFERENCES

ARSSA, 1996. *Carta dei suoli della Media Valle del Crati (Scala 1:50.000). Calabria.*

CASMEZ, 1967. *Carta geologica della Calabria*. Poligrafica e cartevalori, Ercolano - Napoli.

Castrignanò A., Stelluti M., De Giorgio D., 1997. *How to improve modelling the spatial variation of durum wheat yield by using soil tillage information as soft information.* Proceeding of the 14th ISTRO Conference on : Agroecological and Economical aspects of Soil Tillage, Lublin - Poland.

Colella A. *International workshop on: Fan deltas*, 1988, Calabria - Italia.

Surfer Version 6.4. *Surface Mapping System*, Golden Software Inc., 1993-96.

Goovaerts P. *Geostatistics for Natural Resources Evaluation*, 1997. Oxford University Press, New York, p. 483.

Goovaerts P., Journel A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. *European Journal of Soil Science.* 46, 397-414.

Journel A.G., Huijbregts C.J., 1978. *Mining geostatistics*, Academic P., New York, p. 600.

Journel, A.G, 1983. Non-parametric estimation of spatial distributions. *Mathematical Geology*, 15, 445-468.

Luenberger D. , 1992. *Optimization by vector space methods.* John Wiley, N. Y., 1969.

Soil Survey Staff., 1996. *Keys to Soil Taxonomy*, Pocahontas Press Inc., Blacksburg - Virginia.