Listy Biometryczne - Biometrical Letters
Vol. 36(1999), No. 2, 145-157

Application of geostatistical methods to analysis of the data from a pea breeding trial[1]

Janusz Gołaszewski

Department of Plant Breeding and Seed Production,
Olsztyn University of Agriculture and Technology,
Plac Łódzki 3, 10-724 Olsztyn, Poland

SUMMARY

In the field experiments the effects of spatial variability inflate the estimate of experimental error. As a result these effects can mask real treatment effects. The main problem considered in the paper is an improvement of the statistical analysis by using additional information about soil fertility trends and possible competitive interference between neighbouring plots. The basis of the consideration were the data from the field experiment with pea arranged as the balanced lattice design with 25 treatments. Among the statistical methods used were: ANOVA for completely randomized design, randomized block design and balanced incomplete block design and ANCOVA. Kriging and Papadakis' nearest neighbour analysis were used to calculate the concomitant variables connected with the soil characteristics (pH and P, K, Mg) and interplot competition. To compare different statistical methods the relative efficiency was established. Kriging applied for yield analysis and nearest neighbour analysis applied for plant height analysis significantly reduced the experimental error. It was stated that the methods can be good supplemental tools in improving the evaluation of treatments in breeding trials with pea.

KEY WORDS: pea, block design, incomplete block design, ANOVA, ANCOVA, nearest neighbour analysis, geostatistics, kriging.

1. Introduction and objectives

Plant breeders, agronomists and soil scientists conduct thousands of field experiments annually to determine which variety, fertilizer rates or management techniques will optimise crop yield or other agriculturally important plant characteristics. Very often

[1] The study was supported by the Scientific Research Committee grant PO6A 00711 and statutory funds of Olsztyn University of Agriculture and Technology 01020.801.

treatment effects in these experiments are not significant despite distinctive differences between variants or levels of the treatment.

A major source of non-significance may be spatial correlation between closely spaced plots. This correlation is due to soil fertility trends and possible competitive interference between neighbouring plots. As a result, the reasons can mask real treatment effects.

The effects of spatial variability, especially those atributed to soil heterogeneity, inflate the estimate of experimental error. To reduce the adverse effects of spatial variability on the estimation of the experimental error one can select the appropriate experimental design (blocking, randomization) or improve statistical technique of data elaboration by much sophisticated methods such as the nearest neighbour analysis or geostatistical approach.

The objectives of the study were:

– to evaluate the extent and pattern of spatial variability of soil pH and available macronutrients within the experimental field,

– to discuss some methods of statistical analysis which account for spatially variable soil properties and interplot interference between pea forms of a different growth type,

– to compare efficiency of different statistical methods of data evaluation.

2. Material and methods

The basis of the consideration will be the data from a breeding experiment with pea. The experimental material consisted of 25 pea forms of different growth types, including 15 original cultivars and 10 multipodded strains bred at the Department of Plant Breeding and Seed Production of Olsztyn University of Agriculture and Technology.

The balanced lattice design was arranged in three experimental strips (Fig. 1). Each of the six replications (two on each strip) consisted of five incomplete blocks and each of the incomplete blocks contained five plots. Each plot of size 4.5 m^2 had 5 rows, 3 m long, spaced by 0.3 m.

Soil properties related to its fertility (pH and available macronutrients P, K, Mg) were measured before laying out the experiment. Single mesh of the sampling grid was 4 by 8 meters. There were totally 72 sampling points for the whole experimental area under two experiments (pea and faba bean). Besides, mean plant height calculated from a sample of 25 plants/plot and yield/plot were measured.

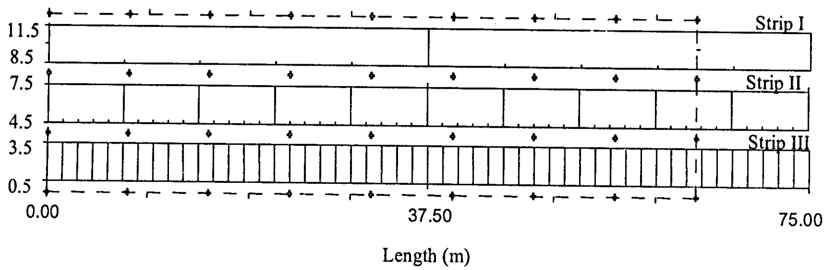


Figure 1. A part of sampling grid of soil properties within pea experiment and layout of the experiement

For the three points across a diagonal of each plot the values of soil pH and available macronutrients P, K and Mg were predicted by kriging. These predictions were then averaged to one observation per plot.

3. Statistical methods

3.1. Kriging

Kriging is a method of spatial prediction that can be used for soil and agricultural properties. It is a form of weighted local averaging. It is optimal in the sense that it provides unbiased and minimum variance estimates of values at unrecorded places. The variance of these estimates can also be computed. It is worth noting that there are several other interpolation methods such as linear interpolation, inverse distance, least squares polynomials, etc., but they are often theoretically unsatisfactory. They may give biased interpolation, they provide no estimate of the error of interpolation, nor do they attempt to minimalize that error.

Kriging is based on the theory of regionalized variables developed by Matheron (1963, 1971) and Krige (1966). One who wishes to study geostatistical methods in depth can use, for example, the books by Journel and Huijbregts (1978) (mining) and Webster and Oliver (1990) (pedology).

The first stage in kriging is the measurement of spatial variation in a property of interest. This measure is called a semivariance. Estimates of semivariances are then used to determine the weights applied to the data when computing the averages for predicted points and are presented in the kriging equations (i.e. Gołaszewski, 1997).

Consider a transect along which observations have been made at regular intervals to give values $z(i)$, $i = 1, 2, .., N$. Then, the relation between pairs of points, h interval apart, can be expressed as the variance of the differences between all such pairs.

So, the per-observation variance is half this which gives

$$\gamma(h) = \frac{1}{2}\mathrm{var}[z(i) - z(i+h)]. \tag{1}$$

The distance h is called the *lag*, and it is any integral multiple of the sampling interval. The scheme (Fig. 2) shows the lagged comparisons along a single transect when all sampling points are present (a) and when there are missing points (b).

For example, the estimate of semivariance for a single transect when $h = 1$ is

$$\hat{\gamma}(1) = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} [z(i) - z(i+1)]^2. \tag{2}$$

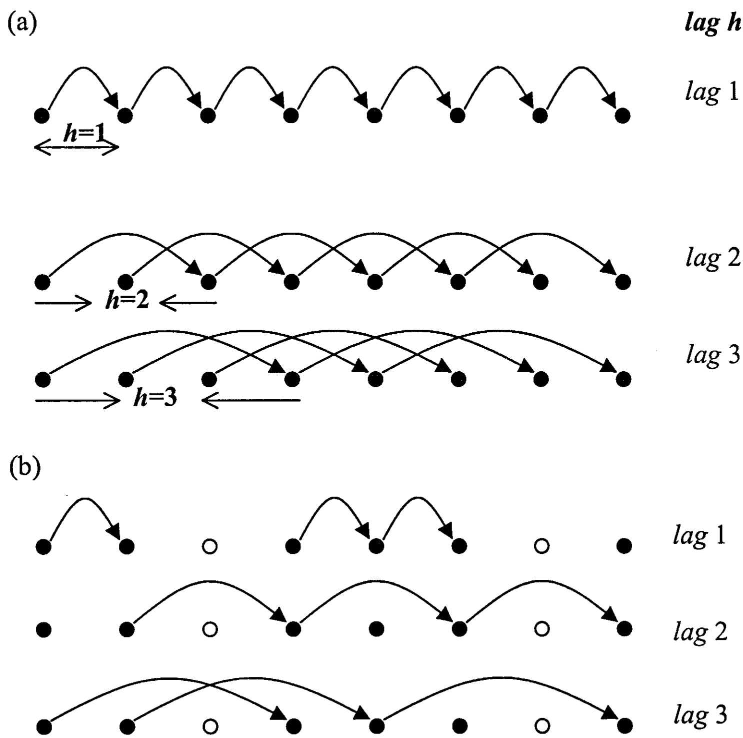


Figure 2. The lagged comparisons for calculating semivariances when (a) all data are present and (b) there are missing data

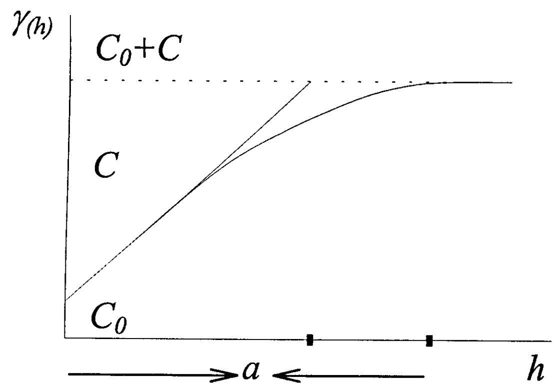
The general form of this equation is

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(i) - z(i+h)]^2 \quad \text{for } i = 1, 2, ..., N(h), \tag{3}$$

where $N(h)$ is the number of observation pairs $\{z(i), z(i+h)\}$ with a distance h. These equations refer to the single transect but the generalization of that formula to the two-dimensional area is quite straightforward.

The semivariance measures the similarity, on average, between points a given distance h apart. The more alike are the points, the smaller is $\gamma(h)$, and *vice versa.*

As above, γ depends on h, and the function relating the two is known as *the semivariogram.* Soil varies continuously in space, and so semivariograms of soil properties and other agriculturally important traits depending on soil fertility (like yield) are continuous functions. They are estimates and as such are subject to error. They can be joined by straight lines or curves to give intermediate values, but their distribution is inevitably irregular. Nevertheless, in most cases it is possible to fit simple function to them. Generally, as can be noted from many papers related to the topic, the two models of semivariograms: linear and spherical are appropriate for the vast majority of agricultural studies. The model of spherical semivariogram and relating equations are presented in Figure 3.



$$\gamma(h) = \begin{cases} C_0 + C\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}\right) \\ \qquad \text{for } 0 < h \leq a \\ C_0 + C \qquad \text{for } h > a \end{cases}$$

Figure 3. Theoretical model of spherical variogram

The semivariogram has certain important characteristics:

(i) it shows the nature of the geographic variation in the property of interest,

(ii) it is needed to provide kriged estimates at previously unrecorded points.

In most instances $\gamma(h)$ increases with increasing h to a maximum, approximately the variance of the data.

The distance a is known as *the range* and it is assumed that points closer than the range are spatially dependent; points further apart bear no relation to one another.

The intercept C_0, when $h = 0$, is known as *nugget variance* and this phenomenon is known as the *nugget effect.* The term derives from gold mining in which the inclusion of a gold nugget in a narrow core sample is possible by chance. Practically, the nugget effect embraces fluctuation of the property that occurs over distances much shorter than the sampling interval and the measurement errors, and limits the precision of interpolation.

The component C represents the *range of variance due to spatial dependence in the data.* The sum of the nugget variance C_0 and the component C is known as *a sill*, the value at which the variance is stabilising.

3.2. Other statistical methods

Among the analytical methods used, beside descriptive statistics, the analyses of variance and covariance were applied. ANOVA and ANCOVA are well known so here they are presented briefly, from the point of view of their implication for spatial variability and a possibility of reducing the experimental error which leads to increased experiment precision.

For a completely randomized design (CRD) all spatial variability fully weights down the experimental error. A randomized complete block design (RBD) enables to control the total spatial variability connected with different fertility of blocks. But it does not take into account the gradient or the periodicity of soil fertility and works well when the number of treatments per block is not too large. For numerous treatments the reduction of experimental error can be accomplished by the joining of experimental units in incomplete blocks containing only a portion of the treatments; here the balanced incomplete block design (BIBD) was applied. In that case blocking of plots in groups smaller than a complete replication eliminates soil heterogeneity to a greater extent than RBD.

Further reduction of experimental error and an increase in experiment precision can be accomplished by the use of accessory observations in the analysis of covariance. In the presented study, two different sets of data were used as the concomitant variables.

The first set was calculated according to Papadakis's nearest-neighbour first differences. Essentially, this technique involves subtracting the mean treatment yield from the yield of each plot and subsequently using the average of the residual yields of adjacent plots as the concomitant variable in the analysis of covariance. The iterative approach suggested by Bartlett (1978) was applied. The iteration was continued until the nearest-neighbour local trends for each treatment averaged to zero.

Wilkinson et al. (1983) discussed some limitations of the iterated nearest-neighbour analysis. These include loss of efficiency due to yield correction with the treatments means, and upward bias in the treatment F-ratio. However, these limitations are usually not significant unless there are substantial non-linear trend effects in the experiment.

The second set of data, as mentioned earlier, was calculated by kriging and consisted of covariate or covariates describing soil fertility. The concomitant variables in ANCOVA were: plant height and yield calculated according to NNA and pH, as well as P_2O_5, K_2O, Mg content and all the four soil properties as single concomitant variables calculated by kriging.

To compare the two analytical techniques the relative efficiency was established as follows (Steel and Torrie, 1980):

– Randomized Block Design (RBD) to Completely Randomized Design (CRD)

$$RE = \frac{(r-1)MSR + r(t-1)MSE}{(rt-1)MSE}, \tag{4}$$

(r – number of replications, t – number of treatments, MSR – replication mean square, MSE – error mean square),

– Balanced Incomplete Block Design (BIBD) to RBD

$$RE = \frac{100[SSB(adj.) + SSE(intrabl.)]}{k(k^2-1)MSE(ef.)}, \tag{5}$$

(k – number of treatments per block, $SSB(adj.)$ – adjusted block sum of squares, $SSE(intrabl.)$ – intrablock error sum of squares, $MSE(ef.)$ – effective error mean square),

– Nearest Neigbour Analysis (NNA) and Kriging to CRD and RBD

$$RE = \frac{100[MSE.Y]}{MSE.Y(adj.)\left[1 + \frac{MST.X}{SSE.X}\right]}, \tag{6}$$

($MSE.Y$ – mean square error of Y, $MSE.Y(adj.)$ – adjusted error mean square of Y, $MST.X$ – treatment mean square of X, $SSE.X$ – error sum of squares of X).

4. Results

Raw data from sampling points connected with the measurements of soil fertility accross the whole area under the two experiments are presented in Figure 4. The first four transects refer to the experimental strips of the experiment with pea. All the analysed soil properties displayed distinctive variation across experimental field. The distributions of these characteristics describe soil fertility and their potential impact

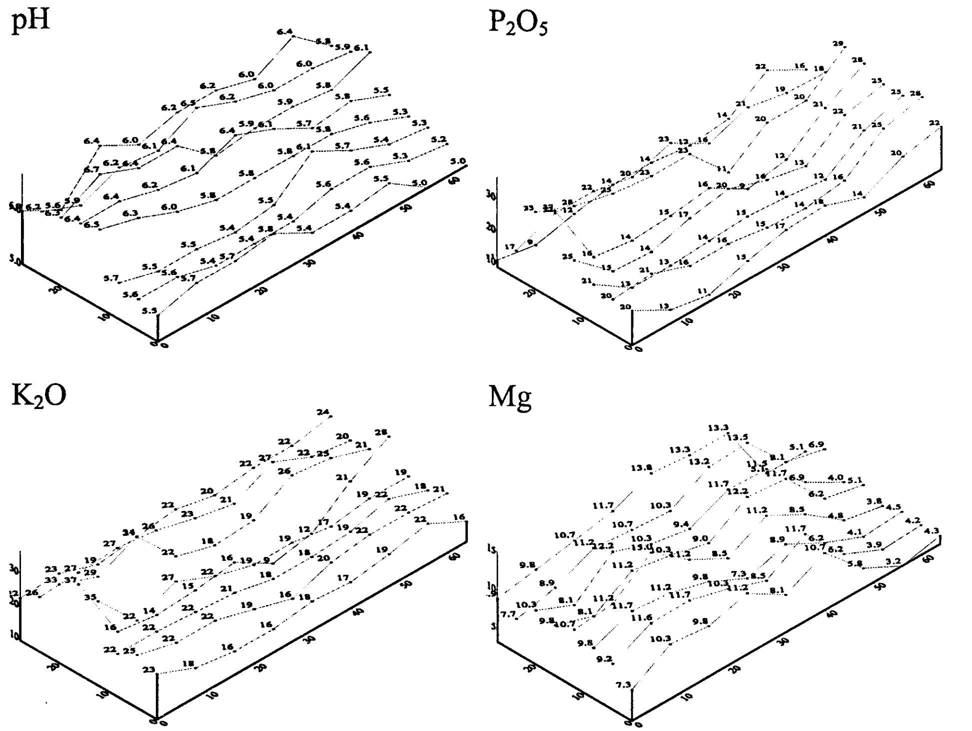


Figure 4. Raw data from sampling points within the experimental field for pH and available macronutrients P, K and Mg (mg/100 g soil)

on plot yields. The final 15 m of the experimental strips with pea were set on the soil of a higher acidity and lower magnesium content than on the other parts of the strips. At the same time the phosphorus content on that area reached higher values. Such results are in accordance with physical properties of the soil (not presented here) showing that the last part of the transects was set on the lighter soil.

Similar values for means and medians point to symmetric distribution of the data for the soil properties (Tab. 1). The lack of symmetry could point to possible trends or outliers. Of the analysed properties, the highest variability was found in the phosphorus and magnesium content, the lowest – in pH.

Average semivariograms for the soil properties presented in Figure 5 reveal different patterns of spatial dependence. The maximum distance assumed was 40 m, because estimates of semivariances based on less than 25 comparisons are unstable and thus they were omitted when fitting the models. The models of semivariograms fitted were spherical for the pH and phosphorus content and linear for the magnesium

Table 1. Descriptive statistics for pH and element concentrations (mg/100 g soil) within the experimental field

Specification	Min.	Max.	Mean	Median	SD (CV%)
pH	5.0	6.7	5.8	5.8	0.39 (7)
P_2O_5	8.8	36.6	18.2	17.4	5.63 (31)
K_2O	9.3	36.6	21.4	21.2	4.95 (23)
Mg	3.2	15.0	9.1	9.8	2.88 (32)

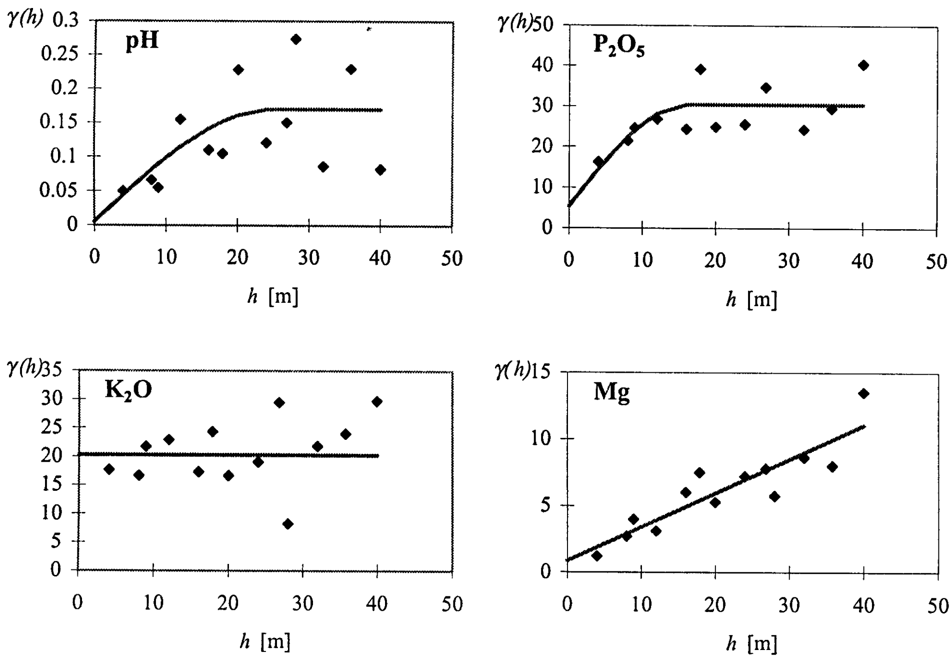


Figure 5. The estimates of semivariograms $[(\text{mg/100g soil})^2]$ and the best-fitting semivariogram function for soil characteristics

content. No spatial dependence was obtained for the potassium content. Parameters of the models, specified in Table 2, show that the range of spatial dependence was 25 m for pH and 16.1 m for phosphorus. The percentage of the nugget effect in the total variability constituted merely 3% for pH and 16% for phosphorus content. The parameters from the semivariogram models were used to predict the values outside the sampling points for each plot.

Table 2. Nugget (C_0), spatially dependent component (C) and range (a) for average semivariograms of pH and element concentrations [(mg/100 g soil)^2]

Specification (Model)	C_0	C (or slope for linear model)	a [m]
pH (spherical)	0.005	0.165	25.0
P_2O_5 (spherical)	5.400	24.97	16.1
K_2O (pure nugget)	-	-	-
Mg (linear)	0.849	0.255	-

The new maps for the data after kriging for the soil properties are presented in Figure 6. The earlier suggestions about the distribution of the characteristics did not change but their spatial patterns are much more visually readable.

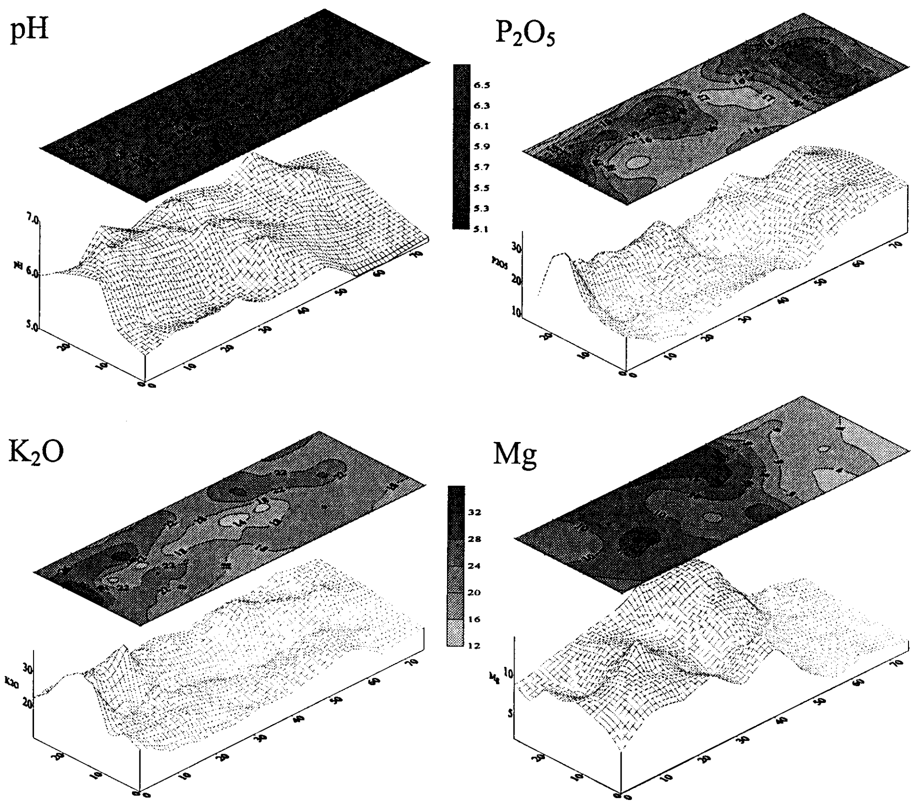


Figure 6. Spatial variability of the experimental field (data after kriging) for pH and available macronutrients P, K and Mg (mg/100 g soil)

Table 3 contains the mean squares for error from different ANOVAs and ANCOVAs as well as the relative efficiency of the methods to CRD and RBD.

For plant height a strong reduction of MSE in relation to the standard methods (CRD and RBD) was observed for BIBD, followed by NNA and kriging with four soil properties as a covariate. Efficiency of BIBD was 134% to CRD and 120% to RBD. Similar results were obtained for NNA, notwithstanding the method was less effective than BIBD in comparison with RBD (108%). The results confirm the intuitive guess that for morphological traits, especially when we study such morphologically differentiated pea forms, the NNA which accounts for possible competition effects of neighbouring plots can be a good alternative to BIBD.

For yield, the strongest MSE reduction was observed for BIBD to CRD – about 10%, but compared to RBD, the relative efficiency was only 101%. However, practical importance to analyse the yield seems to lie in the analysis of covariance with the four properties as a covariate obtained by kriging. The relative efficiency was 102% and

Table 3. Mean square error (MSE) and relative efficiency (RE) for different statistical methods of pea data evaluation

Trait	MSE		RE (%)	
Method	CRD	RBD	CRD	RBD
Plant height	231		100	
RBD		184	122	100
BIBD	153		134[a]	120
NNA[b]	174*	168*	132	108
Kriging(pH)	226	179*	102	102
Kriging(P_2O_5)	233	175*	99	104
Kriging(K_2O)	222*	182	103	100
Kriging(Mg)	229	182	101	100
Kriging (pH, K_2O, P_2O_5, Mg)	215*	176*	107[a]	104[a]
Seed yield	66852		100	
RBD		60556	109	100
BIBD	59953		110[a]	101
NNA[b]	66687	60772	100	98
Kriging(pH)	66020	61065	101	98
Kriging(P_2O_5)	66975	60453	99	100
Kriging(K_2O)	67025	57205	99	105
Kriging(Mg)	67077	59894	99	100
Kriging (pH, K_2O, P_2O_5, Mg)	65775*	55277*	102[a]	106[a]

* – significance of covariate(s) at 0.05

[a] – percentage of MSE reduction

[b] – 2[nd] iteration (stabilizing of MSE)

106% to CRD and RBD, respectively. It can be assumed that the four soil properties determined greatly the soil fertility and productivity. None of the single properties had a significant effect on pea yield estimation.

5. Conclusions

1. Soil heterogeneity caused by different physical and chemical soil properties across the experimental field can be a major environmental factor of high variability of pea traits and may mask real treatment effects in breeding experiments with pea.
2. In a pea experiment when no interplot sowing is applied one ought to account for some effects of spatial correlation and interplot competition of neighbour observations.
3. In the evaluation of treatment effects for pea morphological traits the nearest neighbour analysis can be applied as a supplementary or alternative method to the standard ones.
4. Spatial approach to results from field breeding experiments makes a good promise for using geostatistical methods in order to detect real treatment effects (further studies needed).

References

Bartlett, M.S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion). *J.R.Statist.Soc. B* **40,** 147-174.

Cochran, W.G., Cox, G.M. (1950). *Experimental design.* New York.

Gołaszewski, J. (1997). Zastosowanie metod geostatystyki w doowiadczalnictwie polowym. *Listy Biometryczne - Biometrical Letters* **34**, 63-85.

Journel, A.G., Huijbregts, C.J. (1978). *Mining geostatistics. Academic Press*, London.

Krige, D.G. (1966). Two dimensional weighted moving average trend surfaces for ore-evaluation. *Journal of South African Institution of Mining and Metallurgy* **66**, 13-38.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246-1266.

Matheron, G. (1971). The theory of regionalized variables and its applications. *Cahiers du Centre de Morphologie Mathématique*, *Fointainbleau*, 5.

Papadakis, J.S. (1937). Methode statistique pour des experiences sur champs. *Bull. Inst. Amel. Plantes a Salonique*, 23.

Steel, R.G.D., Torrie, J.H. (1980). *Principles and procedures of statistics. A biometrical approach.* McGraw-Hill Book Company, New York.

Webster, R., Oliver, M.A. (1990). Statistical methods in soil and land resource surveys. Oxford University Press.

Wilkinson, G.N., Eckert, S.R., Hancock, T.W., Mayo, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion). *J. R. Stat. Soc. Ser. B* **45**, 151-211.

Received 22 September 1998; revised 8 March 1999

Zastosowanie metod geostatystycznych do analizy wyników doświadczenia hodowlanego z grochem

STRESZCZENIE

Efekty zmienności przestrzennej mogą zwiększać błąd doświadczalny i w konsekwencji maskować faktyczne efekty obiektowe. W pracy rozważa się możliwości zmniejszenia wielkości błędu doświadczalnego dzięki wykorzystaniu dodatkowych informacji związanych z trendami żyzności gleby oraz konkurencyjnym oddziaływaniem między sąsiednimi poletkami.

Podstawą rozważań były wyniki doświadczenia polowego z grochem założonego metodą kraty kwadratowej zbalansowanej z 25 obiektami. W opisie wyników zastosowano różne metody statystycznej analizy danych, w tym analizę wariancji (ANOVA) układów całkowicie losowego (CRD), losowanych bloków (RBD), bloków niekompletnych (BIBD) oraz analizę kowariancji (ANCOVA). Jako zmienne towarzyszące w analizie kowariancji wykorzystano kwasowość gleby (pH) i zasobność gleby w P, K i Mg oraz efekty oddziaływań sąsiedzkich wyznaczone odpowiednio za pomocą krigingu oraz metody najbliższego sąsiada (NNA). W celu porównania metod wyznaczono ich efektywności względem klasycznych metod analizy dla układu całkowicie losowego i losowanych bloków.

Stwierdzono, że kriging zastosowany w ocenie plonu oraz metoda najbliższego sąsiada zastosowana w ocenie wysokości roślin istotnie zredukowały wielkość błędu doświadczalnego. Zatem, obie te metody moga stanowić dobre uzupełniające narzędzie statystyczne w ocenie obiektów w doświadczeniach hodowlanych z grochem.

SŁOWA KLUCZOWE: groch, układ blokowy, układ bloków niekompletnych, ANOVA, ANCOVA, analiza sąsiedztwa, geostatystyka, kriging.