Australian Journal of Crop Science AJCS
AJCS 9(5):424-430 (2015) ISSN:1835-2707

Geostatistical approach for testing wheat lines using a non-replicated design

Lindolfo Storck[1*], Giovani Benin[1], Alberto Cargnelutti Filho[2], Cristiano Lemes da Silva[3], Samuel Cristian Dalló[1], Thiago Duarte[1] and Luiz Henrique Sassi[1]

[1]Department of Agronomy, Federal University of Technology - Paraná (UTFPR), PR 469, km 01, 85501-970, Pato Branco, PR, Brazil
[2]Department of Crop, Federal University of Santa Maria (UFSM), Santa Maria, RS 97105-900, Brazil
[3]Interdepartmental Genetics, 2004 Throckmorton Plant Sci. Center, Kansas State University, Manhattan, KS 66506, USA

***Corresponding author: lindolfostorck@gmail.com**

Abstract

Seed availability is a limiting factor in early generations of wheat breeding programs, creating difficulties for the performance of replicated trials and leading to selection errors attributed to environmental effects. Thus, the objectives of this study were to determine the percentage and spatial distribution of check plots in the experimental area and to propose criterion based on a geostatistical analysis for selecting promising non-replicated wheat lines based on their grain yield. In this study, grain yield data (GYo) were obtained from 300 plots, arranged in a 15 row × 20 column matrix, from a uniformity trial (one genotype). Multiple scenarios were generated by varying the percentage of check plots, which were randomly determined with 1000 re-samplings to estimate grain yield (GYe) for non-sampled plots using an ordinary kriging method. The efficiency of the estimation method was assessed by calculating the Pearson's correlation coefficient and the mean square error between the GYo and GYe values for each re-sampling. Various spatial distributions of the check plots were evaluated using distinct models for semivariogram fitting. The correlation between the observed and the ordinary kriging-estimated values in the test area plots demonstrates that this approach can be used to identify superior lines for allocation with non-check plots. The estimated results (generated from the check plots) can be used as a reference point for the observed values of a given line. A systematic distribution of check plots in which the entries are alternated with and without checks in the sequence of rows or columns was the best geostatistical approach.

Keywords: Evaluation method; *Triticum aestivum*; semivariogram; spatial dependence; spatial fitting.
Abbreviations: GYe_grain yield estimated; GYo_grain yield data; HW_half-width of the confidence interval; LL_lower limit; MSE_ mean square error; Nc_number of check plots; OKr_ordinary kriging; r_Pearson's linear correlation coefficient; SD_spatial dependence; UL_upper limit.

Introduction

The advances achieved through breeding combined with management techniques have enabled an increase in wheat yield from 700 to 2,300 kg ha^{-1} from 1940 to the present in Brazil (Conab, 2014). However, the genetic gains are decreasing, mainly because of the narrow genetic base, prompting the search for new approaches to identify superior genotypes. In this context, it has been a challenge for breeders to continually provide wheat cultivars with high grain yield potentials.

In the early generations of wheat breeding programs, the limited availability of seed is one of the major factors that limits the use of designs with replicated plots. At present, inbred lines are cultivated in head-rows or small plots without replication. Hence, selection is accomplished by comparing the performance of these lines with standard or control cultivars that are randomly distributed throughout the experimental area. The authors hypothesize that this approach might not be the most efficient selection methodology. Therefore, the most efficient selection techniques should be investigated to identify and realize small gains in grain yield for lines that are still in the pre-trial phase. Spatial statistical analysis based on experiments with replicates has been applied in plant breeding to improve genotype selection for a long time. Briggs and Shebeski (1968) reported a significant correlation (0.88) between check plots spaced at 2.7 m apart and that this association decreased as the distance between plots increased. Other studies also have indicated that the efficiency of genotype selection can be improved by employing spatial methods of statistical analysis (May and Kozub, 1995; Kehel et al., 2010). Another spatial approach for analyzing replicated experiments is the Papadakis method. This design uses the mean of the estimated experimental error, which is calculated using adjacent plots as covariates, to decrease the variance of the experimental error. Its use has been effective for improving the values of indicators of experimental precision in soybeans (Storck et al., 2008), wheat trials (Benin et al., 2013) and other crops. An alternative method is the use of a moving average (Stam, 1984; Weber and Stam, 1988; Edmé et al., 2007), where the covariate is estimated as the mean of the values of the neighboring plots. This approach has been shown to be efficient for the selection of wheat (Townley-Smith and Hurd, 1973) and soybean lines (Diers et al., 1991). Applying spatial analysis to an augmented block design, Müller et al. (2010) also demonstrated that the linear variance model (autoregressive variance-covariance structure for rows and

columns) and the spherical model (with nugget) were the most promising. Alternative methodologies have been studied in an attempt to reduce the effect of spatial variation on plots. A modified augmented design that uses check plots as a fertility index is convenient for the measurement of the environmental heterogeneity and increases the efficiency of the selection process (Snijders, 2002). Similar results have been reported by Morejón and Díaz (2013) who combined multivariate analysis methods with a Latin square design to achieve a higher accuracy for the selection of rice lines.

Geostatistical resources, including semivariograms and the subsequent estimates of values for plots using the kriging method, may be employed to use a portion of the plots as a control and the remaining plots for non-replicated lines (Samra et al., 1990). In this study, the authors analyzed the data from 276 plots that were arranged in a grid of six rows and 51 columns. In columns 1, 11, 21, 31, 41 and 51, a standard genotype (control) was cultivated, while unreplicated lines were cultivated in the remaining plots (90%). The authors used the residual values to estimate an index value for each plot using the kriging method. The data were expressed as a percentage of the index value to fit the observed data as a function of environmental variability. Notably, neither previous studies nor applications that include this approach are known.In contrast, several studies have tested the use of the ordinary kriging (OKr) method to evaluate the spatial variability of soil properties in order to estimate unobserved points (interpolation) and to draw spatial variability maps (Camargo et al., 2008; Oliveira Júnior et al., 2011; Santos et al., 2012; Silva Júnior et al., 2012a,b). This method is a great local interpolator and is calculated using the structural properties of semivariograms (nugget, threshold and range), with the minimum variance of its estimate and with no trend (Isaaks and Srivastava, 1989; Pebesma, 2004). As an application of this methodology, based on a 275-point grid, Angelico (2006) observed that a co-kriging method could be employed to estimate the pH and Mn with a high accuracy using the OM content as a covariate. The use of geostatistical methods, i.e., the kriging method, to analyze the spatial distribution of the check plots within the non-replicated lines enables the establishment of better analytical approaches. Thus, the analysis of a standard cultivar, that is sensitive to variations in soil heterogeneity, allows for the estimation of the values of neighboring plots that are occupied by different lines. The efficiency of using geostatistics to estimate the values of plots occupied by lines (without replicates) from the values of plots occupied by the same standard cultivar (control), using semivariogram parameters, is not well known. In addition, the appropriate proportion and positions that are allocated the check plots to infer the expected production of plots with lines is also not known. Therefore, the objectives of the present study were to determine the percentage and spatial distribution of check plots in the experimental area and propose criterion based on a geostatistical analysis for selecting promising non-replicated wheat lines based on their grain yield.

Results and Discussion

Description of data

The grain yield ranged from 2.286 to 6.852 t ha^{-1} with a mean of 4.933 t ha^{-1}, a standard deviation of 0.817 t ha^{-1}, and an acceptable coefficient of variation (16.5%). A normal distribution of values was confirmed using the Kolmogorov-Smirnov test (p=0.81). These inferences indicate that an appropriate standard yield and the normality of the data support the validity of this study. The yield contours for the 10 yield classes generated using the OKr process (Fig 1D) also demonstrate that this experimental area is suitable for the calibration study, because there are heterogeneity between plots and spatial dependence.

Determination of the percentage of check plots

The mean, the lower and upper limits of the "bootstrap" confidence interval (1-p= 0.95) of the Pearson's correlation (r, Fig 2) and the mean square error (MSE, Fig 3) between the observed (GYo) and estimated (GYe) values in the plots with trial lines were correlated to the percentage (p) of check plots for each scenario. With an increase in the percentage of check plots, there was a significant second-degree polynomial function ($p<0.05$) in terms of the correlation coefficient ($r = 0.5461 + 0.00219p - 0.0000091p^2$) and a significant second-degree polynomial function in terms of the MSE ($MSE = 0.4899 – 0.002167p + 0.0000096p^2$), with maximum r values and minimum MSE values for p>100%. However, the width of the "bootstrap" confidence interval (LL-UL) for the correlation coefficient (r) and MSE was minimal when p=51% and p=50% for r and MSE, respectively. The increase in the width of the estimates for r and MSE from the midpoint (p= 50%) was attributed to a decrease in the sample size, the number of check plots with trial lines used to estimate the r, and the MSE statistics, which was predictable based on statistical theory. The ideal proportion of checks/entries can vary according to several factors, such as the availability of seeds and spatial heterogeneity, among others. Müller et al. (2010) recommended the proportions of 1/8 and 1/5, and the latter was also recommended by Martin et al. (2006). In the current study, it was not possible to use smaller values for p (p< 30%) because the sampling procedure for a smaller number of check plots may have selected for plots that were spatially grouped and, consequently, to flaws when evaluating larger areas; a smaller number of check plots may have also inhibited the re-sampling process because of the indeterminacy of estimating the semivariogram parameters. Using a proportion of 10% check plots and a similar methodology, Samra et al. (1990) also reported a low efficiency for the selection of wheat lines. For greater values of p (p> 80%), the applicability of the process is no longer feasible because of the costs of evaluating the controls. Hence, breeders could assess the real gain of using the selection method proposed in this study. Further studies should select inbred lines using traditional methods in addition to the methods proposed in this study in order to enable proper comparisons. Analyzing the outcomes of the 11 simulated scenarios, we verify that using a percentage of 50% check plots was sufficient to estimate the values of the non-check plots. This determination was based on the mean and confidence interval of the correlation coefficient and the mean square error for the remaining plots in which the trial inbred lines were placed. If a lower p is used, the efficiency width increases and can result in a lower (or higher) accuracy when selecting lines because of the increased unpredictability of the efficiency.

Approach for line selection

Table 1 provides the Pearson's correlations observed between the real values and those estimated for the non-check plots using OKr. The parameters (nugget, threshold and reach effects) of the semivariogram were fitted using different models (exponential, Gaussian and spherical) and different spatial distributions of the check plots (random, strips and

Table 1. Percentage (n%) of check plots, the estimated Pearson's linear correlation coefficient (r) between the observed grain yield (t ha^{-1}) and the yield estimated by applying ordinary kriging with different semivariogram models (Exponential, Gaussian and Spherical), estimates of the nugget (semivariance), threshold (semivariance) and range (m) effects, and estimates of the spatial dependence (SD) for the evaluated check plot spatial distributions within the wheat line trial area.

Type / n%	Statistic	Exponential	Gaussian	Spherical
Random / 50	r	0.676*	0.693*	0.702*
	Nugget	0.416	0.432	0.404
	Threshold	0.495	0.485	0.476
	Range	6.98	3.51	2.73
	SD[+]	Weak	Weak	Weak
Random 1 to 10 / 50	r	0.509*	0.499*	0.509*
	Nugget	0.888	0.301	0.669
	Threshold	20.04	666.2	4.207
	Range	1831	243.1	139.6
	SD	Strong	Strong	Strong
Random 11 to 20 / 50	r	0.732*	0.729*	0.731*
	Nugget	1.027	1.191	1.007
	Threshold	92.11	3600	15.19
	Range	3192	234.8	205.1
	SD	Strong	Strong	Strong
Strips / 46	r	0.449*	0.415*	0.410*
	Nugget	0	0.182	0.086
	Threshold	0.445	0.424	0.422
	Range	3.40	2.83	2.85
	SD	Strong	Moderate	Strong
Strips 1 to 10 / 52	r	0.288*	0.320*	0.328*
	Nugget	0	0.045	0.043
	Threshold	0.344	0.311	0.311
	Range	3.12	1.58	2.24
	SD	Strong	Strong	Strong
Strips 11 to 20 / 52	r	0.459*	0.554*	0.443*
	Nugget	0	0.130	0.001
	Threshold	1.045	0.709	0.677
	Range	7.84	2.44	2.98
	SD	Strong	Strong	Strong
Systematic / 50	r	0.691*	0.694*	0.689*
	Nugget	0	0.051	0.330
	Threshold	0.373	0.865	0.575
	Range	2.40	1.15	2.62
	SD	Strong	Strong	Moderate
Systematic 1 to 10 / 50	r	0.484*	0.484*	0.526*
	Nugget	1	1	1
	Threshold	1.201	1.201	1.201
	Range	9.00	5.10	3.00
	SD	Weak	Weak	Weak
Systematic 11 to 20 / 50	r	0.684*	0.689*	0.718*
	Nugget	1	1	1
	Threshold	1.424	1.424	1.424
	Range	9.00	5.10	3.00
	SD	Moderate	Moderate	Moderate

* Significant according to the t-test ($p < 0.01$); [+]SD (< 25% = Strong; 25-75% = moderate; > 75% = weak)

systematic) in the trial area and in the two portions of the trial area. Using all the plots (N = 300) in the trial area and a systematic spatial distribution of 50% check plots, the correlation coefficient between the estimated values (GYe), based on OKr and exponential models, and the observed values (GYo) was significant (r= 0.691, p< 0.01). These results also agree with those reported by Müller (2010), who reported that a systematic arrangement is more effective. This result is attributed to the strong spatial dependence between neighboring plots in the trial area (SD = nugget effect / threshold effect, as a percentage, which was less than 25%) (Cambardella et al., 1994). The precision of the moving average estimator for genotypic values is considered satisfactory for a neighborhood when the radius is slightly greater than twice the interplant distance of the circular area (Stam, 1984). This value is equivalent to the range (2.4 distances between plots) obtained in this study, when using the exponential model for a systematic spatial distribution of check plots. Estimating the first order spatial autocorrelation (ρ) (Paranaíba et al., 2009) for the 15 plots in each of the 20 columns of the trial area, the mean estimate was $\rho = 0.32$ and the "bootstrap" confidence interval (1-p = 0.95; 3,000 re-samplings) had limits of $\rho = 0.22$ and $\rho = 0.42$. These results also indicate a well-defined spatial dependency for the trial area and can account for the significant correlation between the observed values and the values estimated using the OKr process. Significant correlation between the observed and the estimated values for the test area plots was verified, demonstrating that this information can be used to identify superior lines. In practice, if different wheat lines are allocated

A	C1	C2	C3	C4	C5	C6	...	C20
R1		0	0		0	0	...	
R2	0	0		0	0		...	
R3			0			0	...	0
R4				0	0		...	
R5	0	0			0	0	...	0
R6	0		0				...	0
...	...	...	...	...	...	...	...	...
R15		0			0	0	...	0

B	C1	C2	C3	C4	C5	C6	...	C20
R1	0	0	0	0	0	0	...	0
R2	0						...	0
R3	0						...	0
R4	0	0	0	0	0	0	...	0
R5	0						...	0
R6	0						...	0
...	...	...	...	...	...	...	...	...
R15	0	0	0	0	0	0	...	0

C	C1	C2	C3	C4	C5	C6	...	C20
R1	0		0		0		...	
R2		0		0		0	...	0
R3	0		0		0		...	
R4		0		0		0	...	0
R5	0		0		0		...	
R6		0		0		0	...	0
...	...	...	...	...	...	...	...	...
R15	0		0		0		...	

D

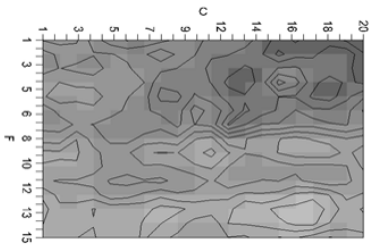

Fig 1. Model of a random spatial distribution (A, 50% check plots), model of a spatial distribution in strips (B, 46% check plots), and model of a systematic spatial distribution (C, 50% check plots) of check plots ("0") in the test area to evaluate grain lines and grain yield results (D) using an ordinary kriging method.

into plots in which the check lines are not allocated in the initial selection process (progeny), or a different line is allocated to each non-check plot, it is possible to use the GYe results (generated from the check plots) as a reference point for the observed values (GYo) of a line. In the process of estimating the plot values (GYe), the variance and half-width of the confidence interval (HW, 1-p = 0.95) of each plot also are estimated (see the R commands in Supplementary data). We could establish the following criterion: if "GYo > GYe + HW" or if "GYo < GYe - HW (for undesirable characteristics)", thus the line should be selected as more favorable than expected based on its estimated value (per point and interval) compared to the neighboring check plots, as determined using the range of the semivariogram. This criterion can be altered depending on the desired selection index when selecting one group of lines, or depending on whether the elevated values for a certain characteristic are favorable or unfavorable. The efficiency of the selection process relies on the correlation between GYo and GYe. In the current study, this efficiency was quantified by calculating the Pearson's correlation (r = 0.691). The efficiency of the selection process may change depending on the genotype used as the check line (adaptability and sensitivity of the genotype in response to variations in soil heterogeneity), the history of the trial area, the model used to fit the semivariogram (exponential, or Gaussian or spherical), the number of plots, and the format of the area, among others. The R programming codes used for selection based on the criteria GYo > GYe + HW, or GYo < GYe - HW for unfavorable characteristics, can be modified and are given in the appendix (R code in Supplementary data). The commands can be rearranged for any type of check plot spatial distribution and for any format and size of data matrix (rows and columns). In a study that considered a spatial distribution with 10% check plots and 90% non-replicated lines (Samra et al., 1990), the authors reported that an index value could be obtained only with check plots. Even after noting the efficiency of such a method compared to other methods, this strategy requires additional verification of its applicability for breeding. In fact, the method presented in our study appears to validate that proposed by Samra and to be applicable to breeding evaluations based on the reliability given by already established geostatistical methodology. According to the yield data generated for 89 points in a 22-ha area of wheat, Roman et al. (2008) reported a moderate spatial dependence. Estimates of the values for non-sampled sites were determined using an OKr method, and the observed and estimated results are presented as a scatterplot. In this scatterplot, the width of the variation of the estimated values is approximately one quarter of the width of the variation of the observed values, implying a low correlation or low efficiency. In contrast, a positive and relatively high correlation coefficient (r= 0.691) was observed between GYo and GYe in the present study. The efficiency of the procedure when using a random spatial distribution of check plots depends on the sampled positions in each case. The results of the study for the random spatial distribution (Table 1, Fig 1A) indicate that the efficiency is similar to that obtained for the systematic spatial distribution, although there are differences in the SD and in the estimates for the semivariogram parameters (nugget, threshold and range effects) between the models for semivariogram fitting. Therefore, this type of spatial distribution can be employed to validate studies of various scenarios for determining an appropriate percentage of check plots. Despite the better appearance in the field and the more practical arrangement for controlling planting and harvesting, the spatial distribution of check plots in the strips (Fig 1B) was associated with a lower efficiency (lower value for r), even with a strong (exponential and spherical model) or moderate (Gaussian model) spatial dependence between the check plots. The magnitude of the correlation between the observed and estimated values may differ based on changes in the position or the direction of the strips; hence, this type of spatial distribution can be more vulnerable to variation compared to a systematic approach. When the trial area (300 plots) is divided into two portions of 150 plots each to simulate a smaller area, the results indicate a difference in efficiency between the two portions (Table 1). The differences in the semivariogram parameters and in the efficiency (r) occur for all three types of spatial distributions (random, strips and systematic) for the check plots. These differences are believed to be attributed to the smaller numbers of non-check plots in each individual portion. The widths of the variation (UL - LL) in r (Fig 2) and the variation in the MSE (Fig 3) increased as the number of non-check plots decreased. Müller et al. (2010) also verified that in several cases, greater block sizes result in lower MSEs. Therefore, the impact of a reduction in area is similar, and the

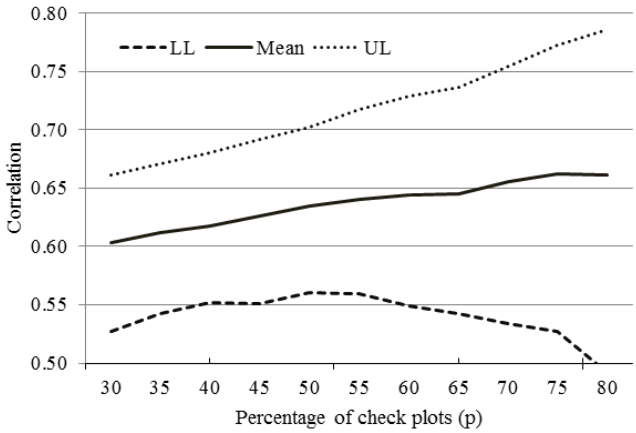

Fig 2. Mean of the estimate of the Pearson's correlation coefficient between the observed and estimated values using an exponential model and the ordinary kriging method as a function of the percentage (p) of check plots, including the lower limit (LL) and upper limit (UL) of the "bootstrap" confidence interval (1-p= 0.95; 1,000 re-sampling).

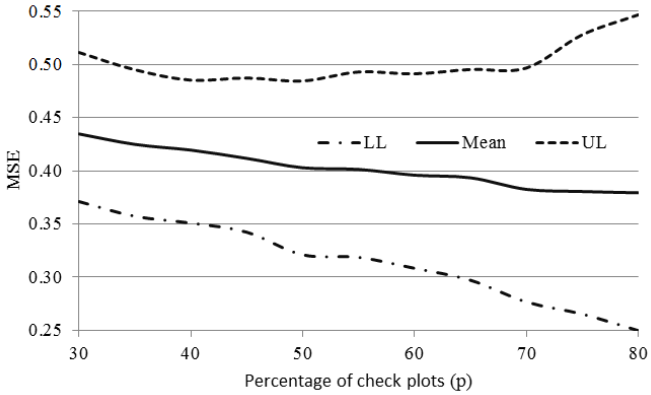

Fig 3. Mean of the mean square error (MSE) between the observed and estimated values using an exponential model and the ordinary kriging method as a function of the percentage (p) of check plots, including the lower limits (LL) and upper limits (UL) of the "bootstrap" confidence interval (1-p= 0.95; 1,000 re-sampling.

most extreme r or MSE values may occur in association with the smallest numbers of plots (smaller sample sizes). Therefore, the efficiency of the selection process may depend upon the positioning of the group of check plots within the larger area. However, it cannot be concluded that the semivariogram model performed better than the other models because the exponential, Gaussian and spherical models produced similar results (similar correlation coefficients) from the data set used in this study.

Materials and Methods

Plant material

The wheat experiment, cultivar Marfim, was sown on June 20, 2013 in the experimental area of the Federal Technological University of Paraná, which is located at 26º10'S, 52º41' W and is 730 m above sea level. The seeding density was 350 viable seeds per square meter. The base fertilization consisted of 30 kg N ha^{-1}, 60 kg P_2O_5 ha^{-1}, 60 kg K_2O ha^{-1} and an additional 70 kg N ha^{-1} in the form of urea (45% N), which was applied at the beginning of tilling (Z2.2 on the Zadoks scale). The controls for weeds, insects and diseases were performed based on the technical recommendations for the wheat crop. At wheat maturity, the grain yield data were obtained from 300 plots that were arranged in the field in a 15 row × 20 column matrix (Fig 1). Each plot consisted of five rows that were 1.0-m long and spaced 0.20 m apart (1 m^2 area). The grain yield (GYo) was determined by weighing the grains of each plot and adjusting the weight to 13% of moisture content (wet basis) and then converting to kg ha^{-1}.

Geostatistical analysis

The grain yield data were georeferenced by the row number (1 to 15) and column number (1 to 20), as a matrix 15 x 20.
The experimental semivariogram is estimated by $\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}\{Z(x_i) - Z(x_i + h)\}^2$ for the h class distance and N(h) represents the number of the regionalized variable pairs separated by an h distance.
Features available in the R software (R Development Core Team, 2013) were used to calculate the semivariogram, and the functions available in the "gstat" package (Pebesma, 2004) to estimate the parameters of the exponential model $\gamma(h) = C_o + C_1\left[1 - e^{-3(h/a)}\right]$, for all h; the Gaussian model $\gamma(h) = C_o + C_1\left[1 - e^{-3(h/a)-3(a/h)^2}\right]$, for all h; and the spherical model $\gamma(h) = C_o + C_1\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}(h/a)^3\right]$, for $0 \leq h \leq a$, and $\gamma(h;\theta) = C_o + C_1$ for a<h. In "gstat" package, a iterative process is used until convergence to obtain estimates for the

nugget (C_o), threshold (C_o+C_1) and range effects (a) (Isaaks and Srivastava, 1989).
Using an ordinary kriging (OKr), a non-observed x_o point can be predicted. We calculate the n by n distances matrix (D) and the estimated value (GYe) for this point as:

$GYe = M + (C_o+C_1)\, r_o' \, S\, (GYo-M)$, where:
$M = (1' \, S\, 1)\, (1' \, S\, GYo)$, $1' = [1 \;\; 1 \;\; 1 \ldots n]$
$r_o = \exp(-(D_o/a)^2)$, D_o is n-vector of distances from x_o point;
$S = C_o\, I_n + (C_o+C_1)\, R$, n x n variances-covariances matrix;
$R = \exp(-(D/a)^2)$, is the matrix with distances between all points and with 1´s in the diagonal;
GYo = the n observed values.
The estimated variance of GYe is: $Var(GYe) = (C_o+C_1) - (C_o+C_1)\, r_o' \, S\, (C_o+C_1)\, r_o$
and the half-width of the confidence interval (HW) can be calculated using the standard normal distribution.

Determining percentage of check plots

The percentage (p) of check plots ranged from p= 30% to p= 80% at 5% intervals (a total of 11 scenarios was considered). The number of check plots (Nc) for each scenario was calculated using the equation Nc = pN/100; where, N = 300 plots in the trial area (15 rows × 20 columns). For each scenario, repeated 1,000 times, the GYo data for the Nc check plots were resampled (without replacement). A semivariogram was generated for each set of Nc check plots, and the exponential model was adjusted to obtain estimates for the nugget, threshold and range effects. Features available in the R software (R Development Core Team, 2013) were used to calculate the semivariogram, and the functions available in the "gstat" package (Pebesma, 2004) of the R software were used to estimate the parameters of the exponential model of the semivariogram. Subsequently, the estimates per point and per interval were generated for the N-Nc non-check plots (plots with inbred lines), a process known as ordinary kriging (OKr) (Isaaks and Srivastava, 1989), using commands in the R software environment (R commands in Supplementary data). The GYo was designated as the value for the wheat grain yield observed in one non-check plot, and the GYe was the respective estimated value for the yield for this plot, based on OKr. To evaluate the efficiency of the estimation process for the non-check plots for each resample of 11 scenarios, the Pearson's linear correlation coefficient (r) between the GYo and GYe was calculated, with (N-Nc)-2 degrees of freedom. The mean square error (MSE) was also calculated as $MSE = (1/n)\, \sum_i^n (GYo - GYe)^2$. The *r* and MSE values are the measurements for the validation of the estimation process, which can only be performed in cases where the observed data (GYo) from the non-check portion used to generate the estimates (GYe) are available. In each scenario, the mean and 0.025 (LL, lower limit) and 0.975 (UL, upper limit) quantiles were calculated based on 1,000 estimates of r and MSE, which were estimated using the "bootstrap" (p=0.05) (Ferreira, 2009) interval for the mean, r and MSE. These estimates (mean, LL and UL) for r and the MSE are presented graphically with the percentage (p) of the check plots, in an attempt to identify, using the lower width of the confidence interval, an appropriate range of values for p. All of the analyses were performed in the R software environment (R Development Core Team, 2013) and Microsoft Office Excel.

Procedure for selecting lines

Tests to evaluate the normality of the distribution of values (Kolmogorov-Smirnov) were performed. Based on the previously defined suitable percentage (p) of check plots, three types of spatial distributions for the plots in the trial area were simulated. A random distribution was one distribution evaluated to determine the percentage of check plots as described. Fig 1A shows, in part, a potential scenario for a design with a random distribution of 50% check plots. The trial area was divided into two portions: one portion included columns C1 to C10 and the other portion included columns C11 to C20 (Fig 1A), although these two groups did not necessarily contain the same percentage of check plots. This division was established to evaluate the effects of using a smaller trial area. The second type of spatial distribution of check plots evaluated was check plots in strips, including strips of check plots on the borders and inner strips (Fig 1B), for a total of 46% check plots. For this type of spatial distribution, the trial area was also divided into two portions. The first portion, columns 1-9 was arranged the same as in Fig 1B, while column 10 was the same as column 20. The spatial distribution of the second portion (columns 11-20) of the check plots was the same as the first portion. This portion (smaller area) contains a higher percentage of check plots (p= 52%). The third type of spatial distribution of check plots was a systematic distribution consisting of a control plot followed by a non-check plot, followed by a control plot, and repeated (Fig 1C), resulting in 50% control plots. For this type of spatial distribution, the trial area was also divided into two portions, applying the same systematic spatial distribution criteria to the distribution of check plots on each side, with a total of 50% check plots. For each type of spatial distribution of check plots and for each trial area size, semivariograms were generated, and exponential, Gaussian and spherical models were adjusted (Isaaks and Srivastava, 1989; Pebesma, 2004). Using the nugget (Co) and threshold (C_o+C_1) effects, a spatial dependence (SD) coefficient was estimated: $SD = 100Co/(C_o+C_1)$. The spatial dependence is strong when SD is < 25%, moderate when 25% < SD < 75%, and weak when SD is > 75% (Cambardella et al., 1994). The Pearson's correlation coefficient (r) between the observed (GYo) and estimated (GYe) values was also calculated for the non-check plots. Correlation values can be used to predict the efficiency of the estimation or validation method, given that it is not possible to estimate this value in real cases of line evaluations. The R software was used for this analysis, including the functions available in the "gstat" package (Pebesma, 2004). The commands used to select lines are available as supplementary information (R commands in Supplementary data 1). In this application, the user prepares a text file of data where the first column indicates the plot row number; the second column indicates the plot column number of sort; the third column contains the values for the grain yield; and the last column identifies the classification of the plot (zero for the check plots and a number of the line being evaluated for the remaining plots). At the end of processing, the same input data were reported (row, column, GYo and classification), in addition to the estimates for the values (GYe) and the half-width of the confidence interval (HW, p= 0.05) for all of the plots and the suggestions for selection. Lines were selected in which GYo > GYe + HW for favorable characteristics, or GYo < GYe - HW for unfavorable characteristics.

Conclusions

The efficiency of selection can be maximized by designating 50% of the experimental area for check plots. A systematic spatial distribution in which plots are alternated with and without checks in the sequence of plots is the best approach for a non-replicated design. The correlation between the observed and the ordinary kriging estimated values in the test area plots demonstrates that this information can be used to identify superior lines allocated to the parcels with non-check plots. It is possible to use the estimated results (generated from the check plots) as a reference point for the observed values of a line.

Acknowledgements

We acknowledge the National Council for Scientific and Technological Development (CNPq) for a research productivity grant and the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES).

References

Angelico JC (2006) Desempenho da co-krigagem na determinação da variabilidade de atributos do solo. Rev Bras Cienc Solo. 30:931-936.

Benin G, Storck L, Marchioro VS, Franco FA, Schuster I, Trevizan DM (2013). Improving the precision of genotype selection in wheat performance trials. Crop Breed Appl Biotech. 13:234-240.

Briggs KG, Shebeski LH (1968) Implications concerning the frequency of control plots in wheat breeding nurseries. Can J Plant Sci. 48:149-153.

Camargo LA, Marques Júnior J, Pereira GT, Horvat RA (2008) Variabilidade espacial de atributos mineralógicos de um latossolo sob diferentes formas de relevo. I-Mineralogia da fração argila. Rev Bras Cienc Solo. 32:2269-2277.

Cambardella CA, Moorman TB, Novak JM, Parkin TB, Karlen DL, Turco RF, Konopka AE (1994) Field-scale variability of soil properties in Central Iowa. Soil Sci Soc Am J. 58:1501-1511.

CONAB-Companhia nacional de abastecimento (2014) Acompanhamento da safra brasileira: grãos. Conab, Brasília, Brazil.

Diers BW, Voss BK, Fehr W (1991) Moving-mean analysis of field tests for iron efficiency of soybean. Crop Sci. 31:54-56.

Edmé SJ, Tai PYP, Miller JD (2007) Relative efficiency of spatial analyses for non-replicated early-stage sugarcane field experiments. J Am Soc Sugar Cane Technol. 27:89-104.

Ferreira DF (2009) Estatística básica. 2rd edn. UFLA, Lavras, Brazil.

Isaaks EH, Srivastava RM (1989) Applied geostatistics. University Press, New York.

Kehel Z, Habash DZ, Gezan SA (2010) Estimation of spatial trend and automatic model selection in augmented designs. Agron J. 102:1542-1552.

Martin RJ, Eccleston JA, Chauhan N, Chan BSP (2006) Some results on the design of field experiments for comparing unreplicated treatments. J Agr Biol Envir St. 11:1-17.

May KW, Kozub GC (1995) Success of a selection program for increasing grain yield of two-row barley lines and evaluation of the modified augmented design (type 2). Can J Plant Sci. 75:795-799.

Morejón R, Díaz SH (2013) Combinación de las técnicas estadísticas multivariadas y el diseño aumentado modificado (DAM) en la selección de líneas de prueba en el programa de mejoramiento genético del arroz (Oryza sativa L.). Cultivos Tropicales. 34:65-70.

Müller BU, Schützenmeister A, Piepho H-P (2010) Arrangement of check plots in augmented block designs when spatial analysis is used. Plant Breeding. 129:581-589.

Oliveira Júnior JC, Souza LCP, Melo VF, Rocha HO (2011) Variabilidade espacial de atributos mineralógicos de solos da formação guabirotuba, Curitiba (PR). Rev Bras Cienc Solo. 35:1481-1490.

Paranaíba PF, Ferreira DF, Morais AR (2009) Tamanho ótimo de parcelas experimentais: Proposição de métodos de estimação. Rev Bras Biometria. 27:255-268.

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci. 30:83-691.

R Development Core Team (2013) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org

Roman M, Uribe-Opazo MA, Nóbrega LHP, Johann JA (2008) Variabilidade espacial do número médio de perfilhos e rendimento da cultura de trigo. Bragantia. 67:361-370.

Samra JS, Anlauf R, Weber WE (1990) Spatial dependence of growth attributes and local control in wheat and oat breeding experiments. Crop Sci. 30:1200-1205.

Santos D, Souza EG, Nóbrega LHP, Bazzi CL, Gonçalves Jr AC (2012) Variabilidade espacial de atributos físicos de um Latossolo Vermelho após cultivo de soja. Rev Bras Eng Agric Amb. 16:843–848.

Silva Júnior JF, Siqueira DS, Marques Junior J, Pereira GT (2012b) Classificação numérica e modelo digital de elevação na caracterização espacial de atributos dos solos. Rev Bras Eng Agric Amb. 16:415- 424.

Silva Júnior JF, Marques Junior J, Camargo LA, Teixeira DDT, Panosso AR, Pereira GT (2012a) Simulação geoestatística na caracterização espacial de óxidos de ferro em diferentes pedoformas. Rev Bras Cienc Solo. 36:1690-1703.

Snijders CHA (2002) Field evaluation of type 2 modified augmented designs for non-replicated yield trials in the early stages of a wheat breeding program. Tagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs. 53:59-64.

Stam P (1984) Estimation of genotypic values without replication in field trials. Euphytica. 33:841-852.

Storck L, Steckling C, Roversi T, Lopes SJ (2008) Utilização do método de Papadakis na melhoria da qualidade experimental de ensaios com soja. Pesqui Agropecu Bras. 43:581-587.

Townley-Smith TF, Hurd EA (1973) Use of moving means in wheat yield trials. Can J Plant Sci. 53:447-450.

Weber WE, Stam P (1988) On the optimum grid size in field experiments. Euphytica. 39:237-247.