Experimental Designs and Estimation Methods for On-Farm Research: A Simulation Study of Corn Yields at Field Scale

Carlos Agustín Alesso, Pablo Ariel Cipriotti, Germán Alberto Bollero, Nicolas Federico Martin*

ABSTRACT
On-farm experimentation using Precision Agriculture technology enables farmers to make decisions based on data from their fields. Results from on-farm experiments depend on the experimental design and statistical analyses performed. Detailed information about the accuracy of the treatment effect estimates, and Type I error rates of hypothesis testing under different spatial structure scenarios attained by alternative experimental designs and analysis is required to improve on-farm research experiments. Three thousand yield data sets were drawn from 15 random fields simulated by unconditional Gaussian geostatistical simulation technique and were modeled by applying 10 experimental designs and three estimation methods with experimental units ranging from 138 to 9969 m². No effect of spatial structure, experimental design, and estimation methods was observed on overall mean yield and treatment bias. Unaddressed changes of nugget/sill ratio and range of variogram had a significant effect on estimator efficiency and accuracy with Type I error rates above the nominal rate, which increased with higher spatial autocorrelation. Spatial methods were robust to changes in spatial structure regardless of the design. Randomization of treatment increased the uncertainty of model estimators. In general, the accuracy of treatment effect estimates increased with the number of replications of smaller size. The opposite trend was observed between those estimates and the size of the plots. Analyses showed that the best designs for testing the overall treatment effect in two-treatment experiments would be split-planter, strip-plots, and chessboard because of their size and number of experimental units.

Core Ideas
• Spatial autocorrelation increases grand mean estimator variance in any design or method.
• Spatial autocorrelation reduces treatment effect estimator efficiency if not modeled.
• Spatial autocorrelation increases Type I error if not modeled.
• Designs with small experimental units (strip plots or chessboard) performed better.

Published in Agron. J. 111:2724–2735 (2019)
doi:10.2134/agronj2019.03.0142
© 2019 The author(s).
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

FARMERS HAVE been conducting “trial and error” experiments for centuries to gain knowledge about their production systems (Marchant et al., 2019). In modern agriculture, on-farm experimentation has played a significant role in research and extension activities by allowing the development and generalization of agronomic recommendations, from genotype selection at the regional scales to site-specific rates of fertilizer (Panten et al., 2010). Agricultural experiments are typically performed at research centers to achieve a high control of experimental conditions and to improve the precision of estimated treatment effects. Recommendations made from these experiments involve extrapolations, which might result in additional uncertainty, particularly those related to the variability of soil, weather, farming systems, and farmer skills (Marchant et al., 2019). Conversely, on-farm experiments are claimed to be more representative of real conditions (although higher heterogeneity within experimental units is expected; Piepho et al., 2011).

In general, farmers and crop advisors run experiments on their farms to compare different agronomic practices or to validate recommendations from research centers (Panten et al., 2010).

During the last decade, many farm-scale experiments were powered by the development of precision agriculture technologies. For example, variable rate technologies allow farmers to modify some inputs (e.g., seeds, fertilizers, herbicides, etc.) within fields and combine-mounted yield monitors help to quantify crop responses (Griffin, 2009). This has reduced the time and effort required for applying treatments and measuring yield responses. However, some negative externalities of on-farm experimentation might be still present (e.g., yield losses resulting from extremely low or high inputs rates, nutrients leaching from excessive application of fertilizer, etc.). Also, farmers’ experiments tend to follow systematic and straightforward designs, which are often incompatible with the classical statistical methods for small-plot agronomic experiments (Hicks et al., 1997). The lack of a sound experimental design could lead to making

C.A. Alesso, Facultad de Ciencias Agrarias, Univ. del Litoral (PCA-UNIL), Instituto de Ciencias Agropecuarias del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (ICyT/Litoral - CONICET), Kreder 2805, S3080HOF, Esperanza, Argentina; P.A. Cipriotti, Facultad de Agronomía, Univ. de Buenos Aires (FAUBA), Instituto de Investigaciones Fisiológicas y Ecológicas, Consejo Nacional de Investigaciones Científicas y Técnicas (IFVEA-CONICET), Av. San Martín 4453, C1417DSE, Buenos Aires, Argentina; G.A. Bollero, N.F. Martin, Dep. of Crop Sciences, Univ. of Illinois at Urbana-Champaign, Urbana, IL 61801. Received 1 Mar. 2019. Accepted 25 Aug. 2019.
*Corresponding author (nfmartin@illinois.edu)

Abbreviations: ANOVA, analysis of variance; BLUE, best linear unbiased estimators; GLS, generalized least-squares; OLS, ordinary least-squares; REML, restricted maximum likelihood; RMSE, root mean squared error; VRT, variable rate technology.
decisions based on misleading conclusions or could unnecessarily increase the risk of negative impacts associated with the experiment such as yield losses, nutrient leaching, or pesticide loads caused by extreme treatments (Griffin et al., 2014).

On-farm experiments can be grouped based on their primary objective: Experiments aimed to estimate the global effect of cultural practices and experiments aimed to explore the spatial variability of crop responses within fields (Pringle et al., 2004a). Whereas georeferenced data is not required for the estimation of a global effect of treatments, spatial mapping of response patterns within the field is only possible if spatial information is available (Piepho et al., 2011). The spatial information about crop response enables testing for the null hypothesis of precision agriculture, i.e., spatial variability of crop response is not enough to warrant site-specific management of inputs (Whelan and McBratney, 2000). Alternatively, this information can be used to estimate spatially variable crop response functions required for prescription maps (Pringle et al., 2004a). In this case, the interaction between treatments and site conditions is explored with the final goal of site-specific crop management, either continuous or by management zones (Piepho et al., 2011).

Basic principles of experimental design (i.e., randomization, replication, and local control) ensure valid statistical tests, avoiding sources of potential bias and controlling nuisance factors that can be confounded with treatments effects (Fisher, 1937; Whelan and McBratney, 2000; Pringle et al., 2004a; Panten et al., 2010; Piepho et al., 2011, 2013). However, intrinsic features of production systems, as well as the footprint of the machinery and sensors used to apply treatments or retrieve crop information, have to be taken into account at the design and analysis stage of on-farm experiments (Willers et al., 2008; Whelan et al., 2012). According to Willers et al. (2008), in on-farm experiments involving precision agriculture machinery, multiple experimental units can be recognized as a result of combinations of site-specific characteristics, the type of machinery used, and the kind of treatments applied. In these experiments, the assignment of the site-specific treatments is commonly driven by the crop and field characteristics of different geographical regions and not by a randomization process (Milliken et al., 2010). As data gathered by precision agriculture machinery are the result of treatments systematically assigned to large plots or strips with several spatially repeated observations, spatial correlation is often an issue for classical statistical analyses (Pringle et al., 2004a). Thus, although some farm operations can be simplified, such violation of classic principles of experimentation challenges the design and analysis of experiments using the classical statistical procedures (Piepho et al., 2011).

Advances in precision agriculture technologies and spatial statistical methods bring new opportunities for the design and analysis of on-farm experiments (Griffin et al., 2004). According to Whelan et al. (2012), on-farm experimental designs must fit within technical, agronomic, and economic restrictions, but at the same time, they have to be statistically robust to provide reliable effect estimates and gain knowledge about the process under study.

The first on-farm experiments in precision agriculture settings were developed by applying classical principles of designs of experiments, which resulted in very complex designs and spatial arrangements of experimental units like a checkerboard or checkerboard (Pringle et al., 2004b), sinusoidal patterns (Pringle et al., 2004b), and whole-field blocks (Panten et al., 2010). Advanced statistical modeling techniques were proposed to account for multiple experimental unit sizes and random effects associated with farm operations (Milliken et al., 2010). Although these approaches are statistically robust and optimized for information gain, their design and analysis require a great deal of agronomic and statistical knowledge. Besides, high yield losses could occur due to some treatment combinations, controls, or extreme rates (Whelan et al., 2012; Griffin et al., 2014). Recently, simpler approaches have been proposed such as unreplicated (Lawes and Bramley, 2012) or replicated strip trials (Panten and Bramley, 2012), or small strip trials (Whelan et al., 2012). These designs might be easily more accepted by farmers, and their implementation is not dependent on technology (Pringle et al., 2004a). However, spatial autocorrelation of observations overestimates the number of independent observations needed for assessing treatment effects and the orientation of strips without knowledge of spatial variability patterns can result in lower efficiency (Whelan et al., 2012).

Although several experimental designs and statistical methods have been proposed for on-farm experiments with precision agriculture technologies, information on the comparison of these designs regarding statistical properties are limited. For example, Griffin et al. (2004) compared four systematic large block designs and two estimation procedures under different autocorrelation scenarios and concluded that statistical analysis could provide unbiased and efficient parameter estimates regardless of the variability and number of replications. Piepho et al. (2013) compared systematic and randomized designs using a uniformity trial data (n = 1082 plots) and simulated data from a separable double autoregressive spatial model. In the first case, they concluded that even if a spatial analysis is applied, systematic designs can result in an over-representation of small p-values of F tests. Their simulation study showed that spatial analysis could improve precision compared with classical analysis of variance (ANOVA), provided the model is correctly specified.

In recent work, Marchant et al. (2019) assessed the precision of UK farmers’ on-farm experiments using yield monitor data and showed that even simpler experimental designs could provide reliable estimates of treatment effects if the spatial variation of the response is adequately accounted in statistical analyses.

Although these previous works have addressed some aspects of design and analysis of on-farm experiments, a comprehensive comparison of systematic and randomized experimental designs and statistical analysis in terms of estimator properties, accuracy, and Type I error rates of hypothesis testing under the same spatial structure scenarios is necessary for on-farm experiment optimization. Thus, the objective of this study is to assess the effect of changes on spatial structure features (nugget/sill ratio and range), experimental layouts, randomization, and estimation methods on the bias, efficiency, and accuracy of estimators and the Type I error rate of hypothesis testing of treatment effect by using geostatistical simulation techniques.

**MATERIALS AND METHODS**

**Hypothetical Experimental Conditions**

A simulation study was performed to compare the statistical properties of parameter estimators obtained by systematic and
randomized on-farm experimental designs and estimation procedures. For this purpose, an on-farm experiment aimed to estimate the effect of two treatments (A and B) applied during planting operation under Argentinean Pampas conditions was hypothesized. Although, other aims of on-farm experiments have been proposed, e.g., the estimation of spatial variability of treatment effects (Pringle et al., 2004a; Panten et al., 2010), in this exercise only the estimation of the overall treatment effect was considered.

The following technologies and machinery configurations were assumed to be available for running the experiment and collecting data: (i) A tractor powered with automatic guidance; (ii) a 16-row 52 cm spacing planter with variable rate technology (VRT); (iii) and a combine harvester equipped with an 8-row header, yield monitor system and automatic guidance. Although somewhat smaller than the average, these machinery dimensions can still be found in several farms of Argentina (e.g., Agrometal, 2017; Maizco, 2019). Based on combine specifications, the minimum width of the experimental units was the combine cut swath or half of the planter width, which is equal to 4.16 m.

Yield data of corn (*Zea mays* L.) from a 100 × 200 m long experimental field were simulated. These dimensions were chosen to be approximately equivalent to 12 and 24 times the header width, respectively. This experimental field was assumed to be a subarea (about 2 ha) of a representative larger field from Argentinean Pampas with Tropic Argisol as a significant soil type and a flat landscape with a slope less than 1%. The yield within the experimental field was assumed to have a constant mean and variance values of 8 Mg ha⁻¹ and 0.64 (Mg ha⁻¹)², respectively. This mean yield is representative for corn fields of the Argentinian Pampas in regular seasons (INTA Staff, 2017). Yield variance represents a coefficient of variation of about 10%, which might be low for an entire field, especially if the field is not uniform or flat. However, in this study, only a 2-ha experimental area of a flat field was considered thus variability lower than the entire field can be expected. Yield data were gridded to a 100 × 200 m grid with a spatial resolution of 4.16 m. A total of 1152 data points were simulated for each spatial scenario described in the following section. The size of the experimental field was chosen to keep a balance between the minimum size of the field for a realistic representation of different spatial patterns, allocation of experimental unit sizes and a number of data points, and the time required for the simulations of random fields and fitting of linear models to large yield data sets. In real scenarios, a coarser resolution could be expected due to larger machinery footprints (header or planter width) and the smoothing process along harvest direction (Lark et al., 1997; Marchant et al., 2019).

**Simulated Yield Responses: Random Field Scenarios**

Spatial variability of corn yields within the experimental area was simulated by unconditional Gaussian geostatistical simulation procedure (Webster and Oliver, 2007). Each yield map was assumed to be a realization of a first-order stationary random field with the general linear model from Eq. [1]:

\[
z(s) = \mu + \varepsilon(s)
\]

where: \(z(s)\) is the yield at the location \(s = (x,y)\), where \(x\) and \(y\) are the spatial coordinates, \(\mu\) the overall mean of the process (8 Mg ha⁻¹) and \(\varepsilon(s)\) random term having a Normal distribution with expectation equal to 0 and a spatial variance-covariance matrix as a function of distances \(b\). For these simulations an isotropic spherical covariance model including nugget effect was used (Eq. [2]), which, along with the exponential model, is one of the models used in literature to represent both soil and yield properties (Webster and Oliver, 2007; Thöle et al., 2013, Richter et al., 2015).

\[
C_{\text{nug}+\text{sph}}(b) = \begin{ cases}
\sigma_{\text{nug}}^2 + \sigma_{\text{sph}}^2 \frac{3b}{2a} - \frac{1}{2a} \frac{b^3}{a^3}, & 0 \leq b \leq a \\
0, & b > a
\end{cases}
\]

where \(C_{\text{nug}+\text{sph}}(b)\) is the spatial covariance function; \(a\) is the range of spatial dependence, \(\sigma_{\text{nug}}^2\) is the non-spatially structured variance or nugget, and \(\sigma_{\text{sph}}^2\) is the spatially structured variance or partial sill. The variance of the process was assumed to be \(\sigma_{\text{nug}}^2 + \sigma_{\text{sph}}^2 = 0.64 (\text{Mg ha}^{-1})^2\).

Fifteen spatial variability scenarios were created by varying two critical features of the spatial model as follows: (i) five levels (0, 20, 40, 60, and 80%) of nugget/sill ratio (nugget effect) and (ii) three ranges (20, 40, and 60 m) of spatial autocorrelation. The greater nugget/sill ratio or the shorter the range, the lesser the degree of spatial structure. The maximum length of the range considered was defined by the size and shape of the experimental area. Longer ranges of autocorrelation were reported for field scale studies (e.g., Jaynes and Colvin, 1997, Martín et al., 2005, Urriaciet et al., 2011). As a result, the random fields used for geostatistical simulation covered a wide range of scenarios from strong spatial structure of yields (nugget/sill = 0% and range = 60 m) to a weak and local spatial structure of yields (nugget/sill = 80% and range = 20 m). Whereas some of the combinations could be hardly observed, most of the real spatial structures observed in yield maps fall within this gradient. Two hundred realizations of yield responses from each random field scenario were drawn by geostatistical simulation techniques assuming no treatments effect, i.e., \(\tau = 0\), where \(\tau\) is effect of treatments A and B, respectively. In total, 30,000 yield data sets were simulated. A sample of one realization from each random field scenario is shown in Fig. 1.

**Experimental Designs**

According to the hypothetical machinery configuration given above, two treatments were assumed to be applied during planting operation at a minimum spatial resolution of 4.16 m (one-half of the planter width) and 16.64 m (2 × the planter width) perpendicular and along planting direction, respectively. As a result, the smallest experimental unit size was about 138 m². Whereas the first restriction was related to the dimensions of the header, the second arose from the time spent by the actuators for changing the rate. Five spatial layouts with a different number, size, and arrangement of experimental units were applied in a systematic (designs D1 through D5) or random fashion (designs D6 through D10) to each yield data set. The combination of spatial layouts and randomization procedures resulted in 10 designs which are displayed in Fig. 2 and summarized in Table 1.

Systematic strip plot designs (D1 to D4) are appealing in on-farm experiments as they are easy to implement without interfering planting operations, even if VRT is not available (Piepho et al., 2011). Conversely, chessboard design (D5) is practically not feasible without VRT. Also, without VRT, the
randomization process can turn easy to implement spatial layouts into a cumbersome task. However, for comparisons purposes, all randomized versions of systematic designs were considered.

**Data Analysis**

Each yield data set containing 1152 pixels (each pixel is 17.3 m²) was analyzed by fitting the linear model from Eq. [3] to data.

\[
y_{ij(k)} = \beta_0 + \beta_i + \epsilon_{ij(k)}
\]

where \( y_{ij(k)} \) is the yield datum corresponding to the \( k \)-th pixel within the \( j \)-th experimental unit assigned to the \( i \)-th treatment; \( \beta_0 = \mu \) is the overall mean; \( \beta_i = \tau_i \) is the effect of the \( i \)-th treatment, and \( \epsilon_{ij(k)} \) the experimental error associated to the \( k \)-th pixel within the \( j \)-th experimental unit assigned to the \( i \)-th treatment. Errors were assumed to be random variables normally distributed with expectation equal to 0 and variance-covariance matrix which can be described by an identity matrix or a function of distances depending on the estimation procedure OLS or GLS.

For each one of the 30,000 simulated yield data sets, model parameters were estimated by the ordinary least-squares (OLS) and generalized least-squares (GLS) procedures (Griffin et al., 2004; Milliken et al., 2010). In all cases, homogeneity of variances was assumed. For the GLS method, two approaches were used for modeling spatial structure: (i) assuming covariance parameters as known (GLS1) and (ii) estimating covariance parameters from data (GLS2). In total, 90,000 models were fitted (30,000 data sets by three estimation methods) in an automated fashion, and information about their coefficient estimates and related test statistics were retrieved.

The OLS procedure is based on the classical assumption of uncorrelated errors, which is equivalent to assuming a pure nugget effect for the spatial structure of the variance-covariance matrix of errors (Pinheiro and Bates, 2000). If this assumption among others is met, OLS estimators are regarded as the best linear unbiased estimators (BLUE). In contrast, if observations are autocorrelated, OLS estimators remain unbiased, but they are no longer efficient as their variance is not necessarily the smallest among other potential estimators (Cressie, 1993). Under this scenario, the GLS procedure results in BLUE estimators by taking spatial autocorrelation into account. This is achieved by modeling the spatial structure of errors as a function of separation distance by restricted maximum likelihood (REML) (Milliken et al., 2010).

Bias and efficiency of the estimators were assessed by computing the mean and variance of the sampling distribution of model coefficients. Bias was approximated by computing the difference between the expected value, i.e., the mean of the sampling
distribution of the estimator, and the population parameter, zero, assuming no treatment effect (Eq. [4]).

\[
\text{Bias} (\hat{\beta}, \beta) = E(\hat{\beta}) - \beta
\]  

where \( \hat{\beta} \) is the estimator and \( \beta \) is the parameter. Efficiency was approximated by the variance of the sampling distribution of the estimator (Eq. [5]):

\[
\text{Var}(\hat{\beta}) = E\left( \left( \hat{\beta} - \beta \right)^2 \right)
\]  

Overall performance of estimators was assessed by computing the root mean squared error (RMSE), which integrates information about bias and variance. Thus, lower values of RMSE are associated with lower values of bias and variance (Eq. [6]):

\[
\text{RMSE}(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta}) + \text{Bias}(\hat{\beta}, \beta)^2}
\]  

The significance of treatment effects for each combination of spatial structure scenario, spatial layout, randomization, and statistical method, was assessed through ANOVA. Then, simulated Type I error rates were obtained by computing \( p \)-values associated with the \( F \) or \( t \) test statistic for treatment effect hypothesis testing.

Finally, the effect of spatial structure features (nugget/sill ratio and range), designs (layout and randomization) and estimation methods on bias and variance of estimators, accuracy and Type I error rates were summarized by fitting linear regression models. The significance test of those effects was assessed through ANOVA Type III Sum of Squares after backward step-wise procedure. Multiple comparisons among slopes of the effect of nugget/sill ratio and range, or between mean levels of designs and estimation methods were performed using Tukey’s procedure for family-wise error rate control.

### Table 1. Summary of experimental designs applied to simulated yield data.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Random.</th>
<th>Reps</th>
<th>Pseudo reps</th>
<th>EU area (m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Systematic one planter width strip plots</td>
<td>No</td>
<td>6</td>
<td>96</td>
<td>1161</td>
</tr>
<tr>
<td>D2</td>
<td>Systematic half planter width strip plots</td>
<td>No</td>
<td>12</td>
<td>48</td>
<td>831</td>
</tr>
<tr>
<td>D3</td>
<td>Systematic two planter width strip plots</td>
<td>No</td>
<td>3</td>
<td>192</td>
<td>3323</td>
</tr>
<tr>
<td>D4</td>
<td>Systematic half field width plots</td>
<td>No</td>
<td>1</td>
<td>576</td>
<td>9969</td>
</tr>
<tr>
<td>D5</td>
<td>Systematic chessboard</td>
<td>No</td>
<td>72</td>
<td>8</td>
<td>138</td>
</tr>
<tr>
<td>D6</td>
<td>Randomized one planter width strip plots</td>
<td>Yes</td>
<td>6</td>
<td>96</td>
<td>1161</td>
</tr>
<tr>
<td>D7</td>
<td>Randomized half planter width strip plots</td>
<td>Yes</td>
<td>12</td>
<td>48</td>
<td>831</td>
</tr>
<tr>
<td>D8</td>
<td>Randomized two planter width strip plots</td>
<td>Yes</td>
<td>3</td>
<td>192</td>
<td>3323</td>
</tr>
<tr>
<td>D9</td>
<td>Randomized half field width plots</td>
<td>Yes</td>
<td>1</td>
<td>576</td>
<td>9969</td>
</tr>
<tr>
<td>D10</td>
<td>Randomized chessboard</td>
<td>Yes</td>
<td>72</td>
<td>8</td>
<td>138</td>
</tr>
</tbody>
</table>

† Code, identification code of the experimental design; Random., randomization; Reps, number of replications; Pseudo reps, number of pseudo-replications per within experimental unit; EU area, area in m² of each experimental unit.
Data manipulation, visualization, and modeling were performed using the statistical language R (R Core Team, 2019) and functions from packages `gstat` (Pebesma, 2004), `dplyr` (Wickham et al., 2017), `ggplot2` (Wickham, 2009) and `nlme` (Pinheiro et al., 2017).

### Results

#### Bias and Efficiency of Estimators

Sampling distributions of the overall mean ($\hat{\theta}_b$), had mean values closer to the true value (8 Mg ha$^{-1}$). The most extreme biases were obtained by using the OLS estimation procedure. Biases ranged from −0.03 Mg ha$^{-1}$ (D9 in a strong spatial structure scenario with 20% nugget/sill and 40 m range) to 0.02 Mg ha$^{-1}$ (D9 in a moderate spatial structure scenario with 60% nugget/sill and 60 m range). The estimated variances of $\hat{\theta}_b$ ranged from 0.0020 (Mg ha$^{-1}$)$^2$ (D3 estimated by GSL2 under weak spatial structure scenario with 80% nugget/sill and 40 m range) to 0.0956 (Mg ha$^{-1}$)$^2$ (D4 estimated by OLS in strong spatial structure scenario with no nugget and 60 m range). No effect of range ($F_{1,360} = 2.18, p = 0.1408$), nugget/sill ratio ($F_{1,360} = 0.07, p = 0.7895$), design ($F_{9,360} = 0.03, p = 0.9999$), and method ($F_{2,360} = 0.72, p = 0.4865$) and their interactions on the bias of $\hat{\theta}_b$ were found. In contrast, the interaction between the estimation method and nugget/sill ratio ($F_{2,432} = 14.74, p = 6.43 \times 10^{-7}$) and method and range ($F_{2,432} = 17.60, p = 4.49 \times 10^{-8}$) had a significant effect on the efficiency of the estimators of $\hat{\theta}_b$. Also, significant differences were observed between designs ($F_{9,432} = 4.29, p = 2.62 \times 10^{-5}$) (Fig. 3).

The variance of $\hat{\theta}_b$ was positively related to range, and the slope for the OLS method was significantly different from GLS methods ($p < 0.05$). For the OLS method, the variance of the estimator increased by $7.75 \times 10^{-4}$ (Mg ha$^{-1}$)$^2$ m$^{-1}$, whereas the average slope for GLS methods was $5.62 \times 10^{-4}$ (Mg ha$^{-1}$)$^2$ m$^{-1}$. In contrast, the variance of $\hat{\theta}_b$ was negatively related to nugget/sill ratio and slopes differed significantly ($p < 0.05$) between OLS [slope $= -3.48 \times 10^{-4}$ (Mg ha$^{-1}$)$^2$%$^{-1}$] and GLS methods [avg. slope $= -2.33 \times 10^{-4}$ (Mg ha$^{-1}$)$^2$%$^{-1}$]. Independently of the estimation method, the variability of $\hat{\theta}_b$ distributions of randomized designs were on average about 13% more variable than their systematic counterparts ($t_{432} = 3.56, p = 0.0004$). Pairwise comparisons of designs showed that the highest variances of $\hat{\theta}_b$ were associated with D9 and D8, but only D9 had significantly higher variances than the rest of the designs ($p < 0.05$).

Sampling distribution of the treatment effect estimators ($\hat{\beta}_b$) also had mean values close to the true value. Biases values were small and ranged from −0.029 (D4) to 0.037 Mg ha$^{-1}$ (D9), both cases were estimated by OLS in the strongest spatial...
variability scenario (no nugget and 60 m range). The most substantial responses were observed on the efficiency by comparing variances. Estimated variances of \( \hat{\beta} \) ranged from almost \( 1.1 \times 10^{-4} \) (Mg ha\(^{-1}\))\(^2\) (D2 estimated by GLS2 in a no nugget and 20 m range scenario) to 0.12 (Mg ha\(^{-1}\))\(^2\) (D4 estimated by OLS in no nugget and 60 m range scenario). As observed for the bias of \( \hat{\beta} \), bias of \( \hat{\beta} \) was not affected by changes of range \((F_{1,360} = 1.89, p = 0.1698)\), nugget/sill ratio \((F_{1,360} = 1.92, p = 0.1665)\), nor significant differences were observed between designs \((F_{9,360} = 0.44, p = 0.9115)\) and estimation methods \((F_{2,360} = 0.35, p = 0.7032)\). However, the variances of \( \hat{\beta} \) were positively related to changes in the range and negatively to changes on nugget/sill ratio in designs D9, D4, D8, and D6 estimated by the OLS method (Fig. 4). The slopes of the effect of range and nugget/sill ratio on variances of \( \hat{\beta} \) in each design and method combination along with the multiple comparison procedure results are shown in Table 2. The effects of range and nugget/sill ratio were nonsignificant for any design estimated by any of the GLS methods \((p > 0.05)\).

Type I Error Rates and Root Mean Squared Error

The simulated Type I error rate ranged from 0.5 to 78%. The minimum Type I error was observed in several combinations of nugget/sill ratio, range, design, and methods, but the maximum corresponded to D4 estimated by OLS in a strong spatial structure scenario with no nugget effect and range of 40 m. Simulated Type I error rate remained equal or under 5% in 60% of the combinations of scenarios and designs for the GLS1 method. When uncertainty of nugget/sill and range parameters were added (GLS2), that proportion dropped to 44%. Finally, for the OLS, the null hypothesis was correctly rejected in 28% of the cases. Significant interaction between nugget/sill ratio, design and method \((F_{18,350} = 13.46, p = 2.2 \times 10^{-16})\) and range, design, and method \((F_{18,350} = 10.33, p = 2.2 \times 10^{-16})\) were observed. In all cases, no effect of range and nugget/sill ratio was observed for GLS1 and GLS2 method regardless of the design applied \((p > 0.05)\) (Table 3.). Figure 5 shows the effect of nugget/sill ratio and range in the statistical test from models where independent errors were assumed (OLS method).

All randomized designs estimated by the OLS procedure showed a negative relationship between nugget/sill ratio and the simulated Type I error rate. For these designs slopes ranged from \(-0.22 \times 10^{-3}\) (D7) to \(-4.29 \times 10^{-3}\) (D8). Conversely, among systematic designs, only D3 and D4 showed a significant relationship for nugget/sill ratio. Designs with the smaller experimental unit sizes (D1, D2, and D5) were less affected by the lack of independence of observation and their simulated Type I error rates remained below the significance level of 5%. As observed for nugget/sill, all randomized designs resulted in simulated Type I error rates above 5% for all levels of range. However, only D9, D6, and D8 showed a significant positive relationship between range and simulated Type I error rate, with slopes of \(6.5 \times 10^{-3}\), \(3.3 \times 10^{-3}\), and \(3.0 \times 10^{-3}\), respectively. In contrast, except D3, for which the range was negatively related to simulated Type I error rate \((slope = -1.45 \times 10^{-3})\), in all systematic designs the null hypothesis was correctly rejected in 95% of the times.

Table 2. Estimated slope of the effect of range and nugget/sill ratio on the variance of \( \hat{\beta} \).

<table>
<thead>
<tr>
<th>Code</th>
<th>OLS</th>
<th>GLS1†</th>
<th>GLS2†</th>
<th>OLS</th>
<th>GLS1†</th>
<th>GLS2†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>0.93 a</td>
<td>-0.18</td>
<td>-0.12</td>
<td>-0.77 a</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>D2</td>
<td>0.23 a</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.51 a</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>D3</td>
<td>2.43 a</td>
<td>-1.39</td>
<td>-1.35</td>
<td>-6.53 a</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>D4</td>
<td>86.08 b</td>
<td>4.44</td>
<td>4.47</td>
<td>-71.37 b</td>
<td>-2.05</td>
<td>-2.05</td>
</tr>
<tr>
<td>D5</td>
<td>0.01 a</td>
<td>0.06</td>
<td>0.05</td>
<td>-0.43 a</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>D6</td>
<td>22.80 ac</td>
<td>-3.34</td>
<td>-3.25</td>
<td>-15.82 ac</td>
<td>3.29</td>
<td>3.29</td>
</tr>
<tr>
<td>D7</td>
<td>0.15 a</td>
<td>-0.29</td>
<td>-0.27</td>
<td>-0.45 a</td>
<td>3.15</td>
<td>3.16</td>
</tr>
<tr>
<td>D8</td>
<td>52.06 bc</td>
<td>-3.06</td>
<td>-3.03</td>
<td>-36.39 c</td>
<td>3.42</td>
<td>3.34</td>
</tr>
<tr>
<td>D9</td>
<td>164.48 d</td>
<td>9.16</td>
<td>9.11</td>
<td>-72.09 b</td>
<td>5.73</td>
<td>5.90</td>
</tr>
<tr>
<td>D10</td>
<td>-0.41 a</td>
<td>-3.02</td>
<td>-3.22</td>
<td>-6.37 a</td>
<td>3.18</td>
<td>3.01</td>
</tr>
<tr>
<td>SE</td>
<td>9.01</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

† Code, identification code of the experimental design; GLS1 and GLS2, generalized least-squares procedure with known and unknown nugget/sill and range parameters; OLS, ordinary least-squares procedure; SE, standard error of the slope.

‡ Estimated slopes not different from 0 \((p > 0.05)\).

§ Different letters indicate significant differences between designs \((p < 0.05)\).
The accuracy of \( \hat{b}_r \) ranged from 0.04 Mg ha\(^{-1}\) (D3 estimated by GLS2 under weak spatial structure scenario, 80% nugget/sill and 20 m range) to 0.74 Mg ha\(^{-1}\) (D9 estimated by OLS under the scenario with the strongest spatial structure, no nugget effect, and 60 m range). The RMSE of the estimator was affected by the interaction between estimation methods and range (\( F_{2;432} = 19.47, p = 8.03 \times 10^{-9} \)) and methods and nugget/sill ratio (\( F_{2;432} = 16.24, p = 1.58 \times 10^{-7} \)). Also, differences between designs were observed (\( F_{9;432} = 12.37, p = 2.2 \times 10^{-16} \)). In the first case, longer ranges were related with higher RMSE values, and the slope for OLS method was slightly (but significantly) greater than GLS methods, 2.77 \( \times 10^{-3} \) vs. 2.25 \( \times 10^{-3} \) Mg ha\(^{-1}\) m\(^{-1}\) respectively (\( p < 0.05 \)). In contrast, increases of nugget/sill ratio reduced RMSE values of \( \hat{b}_r \) and the effect was slightly more pronounced in OLS (slope = \(-11.6 \times 10^{-4}\) Mg ha\(^{-1}\)%\(^{-1}\) ) compared to GLS (slope = \(-8.84 \times 10^{-4}\) Mg ha\(^{-1}\)%\(^{-1}\) ) (\( p < 0.05 \)). Independent of the estimation method, randomized designs resulted in slightly higher values of RMSE for \( \hat{b}_r \) than those for the systematic designs (\( t_{432} = 7.07, p < 0.0001 \)). Pairwise comparison of designs showed that the highest RMSE (0.139 Mg ha\(^{-1}\) ) of \( \hat{b}_r \) was associated with D9 followed by D8, D6, D10, and D4 with average RMSE of 0.123 Mg ha\(^{-1}\). Provided RMSE is a linear combination of bias and variance of the estimator, and negligible effects were observed on estimator bias, variations on RMSE were attributed to changes on the variance of the estimator.

The accuracy of \( \hat{b}_l \) ranged from 0.01 Mg ha\(^{-1}\) (D2 estimated by GLS2 under strong but small scale spatial structure scenario, no nugget effect, and 20 m range) to 0.31 Mg ha\(^{-1}\) (D4 estimated by OLS under the scenario with the strongest spatial structure, no nugget effect, and 60 m range). The effect of the interaction between nugget/sill ratio, method and design (\( F_{18;360} = 9.29, p = 2.2 \times 10^{-16} \)) and between range, method and design (\( F_{18;360} = 9.70, p = 2.2 \times 10^{-16} \)) on \( \hat{b}_l \) accuracy was significant. The RMSE of \( \hat{b}_l \) was negatively related to nugget/sill ratio in designs D4, D9, D8, D6, D3, and D10 when no spatial correlation was assumed (Fig. 6). In contrast, variation of the range was positively related to RMSE in D9, D4, D8, and D6 when estimated by OLS (Fig. 6). The estimated slopes for the effect of range, nugget/sill ratio and all combinations of designs and methods, along with multiple comparison procedure results are shown in Table 4. In all cases, no effect of range and nugget/sill ratio was observed for GLS1 and GLS2 method regardless of the design applied (\( p > 0.05 \)).

Figure 7 shows the relationship between the size (upper X axis), estimated by the area in m\(^2\), the number of experimental units (lower X axis) and the RMSE associated with systematic and randomized designs estimated by the OLS and GLS procedures. Due to the same experimental area of fixed size was used to test all the experimental designs, the size of experimental units are inversely related to the number of replications per treatment. Thus, the larger plot size of a given experimental design, the lesser number of replications could be assigned to each treatment. Higher overall RMSE were obtained if designs were randomized, and the GLS procedures were less affected.

**Table 3.** Estimated slope of the effect of range and nugget/sill ratio on the Type I error rate.†

<table>
<thead>
<tr>
<th>Code</th>
<th>Slopes of range OLS</th>
<th>GLS1 ‡</th>
<th>GLS2 ‡</th>
<th>Slopes of nugget/sill OLS</th>
<th>GLS1 ‡</th>
<th>GLS2 ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.16 a§ –0.20</td>
<td>–0.05</td>
<td>–0.22  a</td>
<td>–0.43  a</td>
<td>–0.05</td>
<td>–0.22  a</td>
</tr>
<tr>
<td>D2</td>
<td>0.41 a –0.10</td>
<td>0.00</td>
<td>0.58 a</td>
<td>–0.08 –0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>–1.45 a –0.20</td>
<td>–0.23</td>
<td>–3.60  bc</td>
<td>–0.07  0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>6.50 b 0.13</td>
<td>0.15</td>
<td>–3.86  bc</td>
<td>–0.26  0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D5</td>
<td>–0.25 a –0.50</td>
<td>–0.15</td>
<td>0.52 a</td>
<td>–0.18 –0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D6</td>
<td>3.25 c 0.25</td>
<td>0.03</td>
<td>–3.72  bc</td>
<td>–0.22  –0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D7</td>
<td>–0.27 a –0.10</td>
<td>0.00</td>
<td>–1.74  d</td>
<td>–0.22  –0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8</td>
<td>2.96 c 0.05</td>
<td>0.05</td>
<td>–4.29  b</td>
<td>–0.27  0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D9</td>
<td>6.50 b 0.13</td>
<td>0.15</td>
<td>–3.86  bc</td>
<td>–0.26  0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10</td>
<td>0.18 a –0.05</td>
<td>–0.23</td>
<td>–3.01  cd</td>
<td>–0.09  –0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.48</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† Code, identification code of the experimental design; GLS1 and GLS2, generalized least-squares procedure with known and unknown nugget/sill and range parameters; OLS, ordinary least-squares procedure; SE, standard error of the slope.

‡ Estimated slopes not different from 0 (\( p > 0.05 \)).

§ Different letters indicate significant differences between designs (\( p < 0.05 \)).

Fig. 5. Effect of the interaction between designs and (a) proportion of nugget variance and (b) range on the simulated Type I error rate obtained by the OLS estimation method.
by the size of experimental units provided spatial autocorrelation was accounted for. The size of the experimental units was positively related to RMSE, and this relationship was affected by the estimation method and randomization. For randomized designs, a minimum RMSE was observed when the size of the experimental units was about 832 m², which corresponded to the split-planter design (D6). A possible explanation of this feature could be the final shape of plots after randomization. When comparing D6 (minimum) to D5 and D10, the former design preserves narrow strips after randomization, whereas D5’s strips are wider. In addition, D10’s plots are wider and shorter than D6’s strips. Nevertheless, as the area of the experimental field is fixed, the number of experimental units that can be assigned to each treatment is inversely related to the size of the experimental units (Table 1). Consequently, the effect of the size of the experimental units on the RMSE could not be separated from the effect of the number of experimental units.

**DISCUSSION**

The precision of on-farm experiments can vary due to the experimental design, spatial variability within the experimental area, and the inherent unexplained variability among experimental data (Marchant et al., 2019). Our simulation study allowed us to explore the effect of the spatial structure of yield response, assuming no treatment effect, on the properties of estimators and the Type I error rates and to compare different experimental designs and estimation methods applied to the same spatial variability scenarios. These results are conditioned to the assumptions made on the characteristics of the hypothetical field, machinery footprint, and covariance model for the spatial process. However, the simulated conditions might be mapped to real on-farm experiments in the Argentinean Pampas.

All designs and estimation methods tested in this study yielded nearly unbiased estimators of the overall mean yield of \( \beta_0 \) and treatment effect \( \beta_1 \) in all spatial structure scenarios. Bias values for \( \beta_0 \) and \( \beta_1 \) are comparable to those reported by Griffin et al. (2004). For instance, in their study, bias values for the overall mean yield ranged from –0.030 to 0.086 for OLS and from –0.018 to 0.003 (Mg ha\(^{-1}\)) for the GLS method. The standard errors of the simulated treatment effects, i.e., the standard deviation of their sampling distribution, were similar to those reported by other authors for real on-farm experiments data (Marchant et al., 2019) and simulated experiments (Griffin et al., 2004). The lack of bias in our results, even in systematic designs is explained by the fact that all spatial structure scenarios assumed a stationary process with constant mean and variance within the experimental field. As treatment randomization avoids the occurrence of systematic errors, systematic designs that do not account for spatial trend effects could result in biased estimates (Piepho et al., 2013).

Changes on the spatial structure had a marked effect on the variance of both estimators. These effects also differed between experimental designs and methods of estimation. Since both OLS and GLS estimators are unbiased, changes on spatial...
autocorrelation are expected to affect efficiency but not bias (Griffin et al., 2004). In our study, the variability of both estimators increased with spatial structure. Similar results were reported by Griffin et al. (2004) for systematic designs. For example, in their study, the mean squared error associated with the overall mean increased 11 to 31 and 5 to 20 times from the pure nugget effect scenario ($p = 0$) to the high spatial structure scenario ($p = 0.72$) if estimated by OLS or GLS, respectively. The effect of nugget/sill and range on the efficiency of $\beta_1$ estimated by OLS method were larger than for GLS methods, regardless of the design applied (Fig. 3).

In contrast, the response of the variance of $\beta_1$, the spatial structure was also affected by design, but only when coefficients were estimated by the OLS procedure. The GLS method was less sensitive for variations on spatial structure features provided it accounted for the lack of independence between observations. When the OLS estimation procedure is used, the effect of spatial autocorrelation on the efficiency of $\beta_1$ was more pronounced in designs with large experimental units and with those that were randomized (Fig. 4). Randomization resulted in larger variability of the sampling distributions of the estimators in all spatial structure scenarios due to the uncertainty added by the randomization process. Systematic designs are commonly used in these experiments by practical considerations. However, the analysis of these designs might be sensitive to model suitability and model assumptions (Piepho et al., 2013).

The estimation method, and in a lesser extent, the interaction between designs, i.e., combinations of spatial layout and randomization, and spatial structure features, had a large influence on the Type I error rate. Our results showed that the addition of the spatial information, the analysis yielded more reliable estimates of treatment effects, keeping Type I error rates under the nominal rate of 5%, regardless of the spatial layout, whether it was randomized or not, and spatial structure. In contrast, when no spatial analysis was applied to data, the rate of rejected null hypothesis about treatment effect varied due to the degree of spatial autocorrelation and the design applied (Fig. 5). As spatial structure decreased due to combinations of shorter ranges and higher proportions of nugget variance, the Type I error rate approached the declared significance level in systematic designs with large experimental units. Conversely, this effect was observed in all randomized designs which showed higher values of Type I error rates. Robustness of the GLS method compared to OLS methods was also reported by other authors (Griffin et al., 2004; Lawes and Bramley, 2012; Piepho et al., 2013; Marchant et al., 2019). For example, simulation results reported by Griffin et al. (2004) showed that the Type I error rates of hypothesis testing based on GLS method were kept around 6% regardless the degree of spatial autocorrelation, for all the systematic block design tested. Piepho et al. (2013) also reported empirical Type I error rates around the nominal rate of 5% for randomized designs across the spatial structure and covariance model. In contrast, when systematic designs were applied, results were approximately valid only if the covariance structure was known.

Although the GLS method is more appropriate than OLS for analyzing spatially correlated data, its performance is limited by how well the spatial process is approximated from sample data (Piepho et al., 2011). According to Marchant et al. (2019), misspecification of the spatial covariance model can lead to high under estimation of the uncertainty associated to the model parameters, including the treatment effects or treatment means. In our study, we applied the GLS method to yield data under two approaches: (i) We assumed that all parameters of the spatial model were known using the same parameters used for stochastic field simulations (GLS1), and (ii) we let only the range and nugget variance to be estimated from the data (GLS2). However,
our results showed that the uncertainty about the features of the spatial structure had no impact on estimator properties but increased the simulated Type I error rate in some design and scenario combinations. This can be explained by two reasons: first, in the GLS method the covariance model was assumed to be spherical and the variance of the process to be 0.64 \((\text{Mg ha}^{-1})^2\), which certainly removed considerable amount uncertainty; and second, the large number of data points \(n = 1152\) used in the estimation of covariance parameters also reduced the estimation error. Probably, more differences between these methods could be observed if the whole spatial process would be unknown or fewer data were available for estimating the covariance function. Systematic designs with large strip plots are the types of experiments that farmers are familiar with (Griffin et al., 2004; Whelan et al., 2012; Marchant et al., 2019). However, this straightforward approach can lead to incorrect estimations of treatment effects and misleading conclusions, especially under strong spatial autocorrelation scenarios if spatial data are used in analyses (Panten et al., 2010; Piepho et al., 2011, 2013). In this context, spatial statistics methods have helped to get more reliable inferences from these type of experiments (Griffin et al., 2004). In addition to the estimation methods, the spatial layouts, shape, and size of the experimental units could play an essential part in the quality of results. Our results showed that designs with small experimental units or narrower strips, or more replications were, in general, more robust to changes of the spatial variability structure. Also, the accuracy of the estimations was affected by the size and number of experimental units associated with each design (Fig. 7). Marchant et al. (2019) suggested that increases in precision are attainable by having different experimental treatments within a single row and, to a lesser extent, by replicating treatments within the experiment. Conversely, designs in which each treatment is assigned to a large continuous region could be less precise, and treatment effects could be confounded with underlying yield variation.

Yield maps are a noisy but valuable source of information about crop performance (Sudduth and Drummond, 2007; Leroux et al., 2018). To avoid the inclusion of noisy information and get reliable yield estimates, real on-farm experiments might be conducted using larger experimental units than those used in our simulation study. Although this simulation study was based on nominal machinery footprint, the results could be used as a reference for what can be expected in on-farm experiments conducted in real settings. However, potential issues related to the scaling of the spatial structure or unknown covariance model on the performance of the designs under real applications remains to be verified.

CONCLUSION

On-farm experiments are valuable sources of information for helping farmers to make decisions based on data. The experimental designs and the statistical analyses applied to data can have a significant impact on the quality of the results. All designs tested here resulted in unbiased estimators of treatment effect across the estimation methods, even those with large experimental units and no replications. However, their efficiency in estimating the true effect was affected by the spatial structure of responses, the design, and the robustness of the estimation methods. As a consequence, the accuracy of the estimators and the simulated Type I error also were affected. Regardless of the degree of spatial autocorrelation and the designs characteristics, the GLS estimation methods showed superior performance than the OLS method in terms of better efficiency, Type I error rates, and accuracy.

The ease of on-field implementation and simplicity of analysis are two critical aspects sought in on-farm experiments by farmers. Provided a proper spatial analysis is feasible, the selection of design may be driven by the trade-off between simplicity and the desired precision. Based on the general relationship between size/number of experimental units and accuracy, designs can be ordered split-planter, strip-plots, and chessboard. Whereas systematic designs are straightforward, and some designs can be performed even without VRT technology, their analysis needs to take into account the spatial autocorrelation for making approximately valid inferences. Although more challenging to implement, randomized designs are more likely to produce reliable inferences in scenarios with spatial autocorrelation.

These results would help farmers and practitioners of on-farm experimentation by giving insights about the impact of the designs they are employing to conduct their on-farm experiments. At the same time, our methodology provides a way for testing newer designs, to explore the effect of other features of the spatial process that might influence results like weak stationarity, other covariance models, the total amount of variability, the interaction between treatments, and spatial structure, etc. Also, this approach could be applied to evaluate designs for the spatial estimation variation of the treatment effect or estimate the size of the effect that can be detected.

AKNOWLEDGMENTS

We thank to Universidad Nacional del Litoral (UNL) for the financial support of this research by the grant number 50320180200012.II.

REFERENCES


