Ecology, 85(12), 2004, pp. 3202–3214

EFFECTS OF SPATIAL STRUCTURES ON THE RESULTS OF FIELD EXPERIMENTS

PIERRE LEGENDRE,[1,5] MARK R. T. DALE,[2] MARIE-JOSÉE FORTIN,[3] PHILIPPE CASGRAIN,[1] AND JESSICA GUREVITCH[4]

[1]*Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*
[2]*Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9*
[3]*Department of Zoology, University of Toronto, Toronto, Ontario, Canada M5S 3G5*
[4]*Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794-5245 USA*

Abstract. Field experiments have been designed to account for spatial structures since the inception of randomized complete block designs by R. A. Fisher. In recent years, our understanding of spatial structures led to refinements in the design and analysis of field experiments in the face of spatial patterning. In the presence of spatial autocorrelation in the response variable, is it possible to optimize the experimental design to maximize the response to the experimental factors? The questions addressed in this paper are: (1) What is the effect of spatial autocorrelation on type I error of the tests of significance commonly used to analyze the results of field experiments? (2) How effectively can we control for the effect of spatial autocorrelation by the design of the experiment? (3) Which experimental designs lead to tests of significance that have greater power? (4) What is the influence of spatial autocorrelation on power of ANOVA tests of significance? This paper attempts to answer these questions through numerical simulations with known spatial autocorrelation. Response variable were simulated to represent the sum of separate effects: (1) an explanatory environmental variable (which could be used as a covariable in the analysis) with a deterministic structure plus spatial autocorrelation, (2) an effect of the experimental treatments, (3) spatial autocorrelation in the response (e.g., biological) variable, and (4) a random error. The program repeatedly generated and analyzed surfaces with given parameters (1000 replicates). The rejection rate of the null hypothesis of no effect of the treatment onto the response variable provided estimates of type I error and power.

The simulations showed the following: (1) In the presence of spatial autocorrelation, or if repetitive deterministic structures are present in the variables influencing the response, experimental units should not be positioned at random. (2) ANOVA that takes blocking into account is an efficient way of correcting for deterministic structures or spatial autocorrelation. (3) For constant effort, experimental designs that have more, smaller blocks, broadly spread across the experimental area, lead to tests that have more power in the presence of spatial autocorrelation. (4) Short-ranged spatial autocorrelation affects the power of ANOVA tests more than large-ranged spatial autocorrelation.

Key words: analysis of variance; autocorrelation; field experiments; numerical simulation study; spatial structure.

INTRODUCTION

In recent years, our understanding of, and ability to model, spatial structures has opened new possibilities for the design and the analysis of field experiments in the face of spatial patterning caused by autocorrelation in the response variable or by environmental variables that may influence the experiment's response variable. Field experiments add controlled variation to natural variation and spatial structures. Is it possible to optimize the experimental design in order to maximize our ability to detect a response to the experimental factors in the presence of spatial autocorrelation (SA) in the response variable?

Four spatial components are at play in field experiments: (1) the spatial structure of the environmental variables, which may act upon the response variable (e.g., an environmental gradient in soil moisture may cause a differential response); (2) the spatial autocorrelation of the response variable (e.g., the similarity in responses of near neighbors due to genetic resemblance and limited dispersal); (3) the degree of dependence of the response variable on the spatially structured environmental variables (environmental forcing); and (4) the dependence of the response variable on the treatment levels of the spatially structured experimental design, which may be efficiently randomized or not (Dutilleul 1993: Fig. 1).

[5] E-mail: pierre.legendre@umontreal.ca

Field experiments have been designed to account for spatial structure since the inception of randomized complete block designs by Ronald A. Fisher (1926). Fisher was seeking ways of improving field experiments in agriculture. His designs have become standard conceptual instruments for research in many fields, including ecology. A huge number of papers have appeared on the subject (Hahn 1982). Blocking has been introduced to consider the mixed effects of inherent spatial dependence of the environmental variables and the spatial dependence of the response variable to it (Underwood 1997). Other approaches have been proposed to deal with the effects of the spatial autocorrelation of the response variable. Bartlett (1978) perfected a previously proposed method of correction for the effect of spatial autocorrelation due to an autoregressive process in randomized field experiments, adjusting plot values by covariance on neighboring plots before the analysis of variance. Geostatistical approaches have also been proposed to account for spatial autocorrelation (Ver Hoef and Cressie 2001).

There is a great deal of information and discussion in the statistical and ecological literature on the effects of spatial autocorrelation on statistical tests and possible solutions to the problem it presents at the stage of testing (among others, Sokal et al. 1993, van Es and van Es 1993, Hoosbeek et al. 1998, Bonham and Reich 1999, Casler 1999, Dale and Fortin 2002, Keitt et al. 2002). In this paper, we investigate how the impact of the effect of spatial autocorrelation can be reduced at the stage of designing experiments. In a companion paper (Legendre et al. 2002), we investigated the consequences of deterministic spatial structures and spatial autocorrelation for the design and analysis of ecological field surveys. To avoid incorrect conclusions, the study design needs to be adjusted to suit the scale of the heterogeneity of the system being studied (Dutilleul 1998).

The questions addressed in this paper are: (1) What is the effect of spatial autocorrelation on the type I error of the tests of significance commonly used to analyze the results of field experiments? (2) Can we control for the effect of spatial autocorrelation by the design of the experiment? (3) Which designs lead to tests of significance that have the most power? We compared the completely randomized design to several types of blocked experimental designs. (4) What is the influence of spatial autocorrelation on power of the ANOVA tests of significance? To answer these questions, we used numerical simulations to estimate the rate of type I error and power of the tests of significance in analyses of variance associated with different experimental designs and types of spatial structures.

Methods

Data generation model

A response variable (R) measured during a field experiment is considered in our simulations to represent the sum of separate effects: the fixed effect of the treatment (T), the influence of an explanatory environmental variable (E), spatial autocorrelation in the response variable (SA_R), and a spatially unstructured random error component (ε) taking independent values for each observation i:

$$R_i = T_i + f(E_i) + SA_{Ri} + \varepsilon_i. \quad (1)$$

The environmental variable, in turn, may possess a deterministic structure (D) plus a spatially autocorrelated error component (SA_E) and independent error at each point:

$$E_i = D_i + SA_{Ei} + \varepsilon_i. \quad (2)$$

For example, soil texture may vary from the top to the bottom of a hill due to sorting of soil particles during erosion (deterministic structure). In addition, due to their history, local patches of soil may tend to resemble one another more closely than they do patches further away (spatial autocorrelation). As a result, the model for the response variable R comprises some or all of the following elements:

$$R_i = T_i + \beta E_i + SA_{Ri} + \varepsilon_i. \quad (3)$$

The assumptions of this model are the following: (1) All environmental effects can be summarized by a single variable whose effect on R is linear; the effect is thus modeled by multiplying E by a transfer (regression-type) parameter β. (2) The error component ε, which takes independent values (i.e., not spatially autocorrelated) for each observation i, is modeled as a normal error term whose variance (Var_ε) is fixed by a parameter provided for each simulation. A normal error can legitimately be assumed for a natural phenomenon that results from a large number of factors acting independently, whose random effects are cumulative, if the variance of the phenomenon produced by each factor is small (Galton 1898).

Simulation method

For this paper, a simulation run was controlled by a series of parameters as follows: (1) specify the number of simulations to be made and the size of the experimental field, which is given in number of pixels from west to east and from north to south; (2) specify an experimental design (number of treatments, number of blocks, number of replicates per treatment in each block, geographic positions of the experimental units); (3) specify a treatment level for each experimental treatment; (4) specify the characteristics of the environmental component E (the type of deterministic structure D, the parameters of the spherical variogram specifying the autocorrelation function SA_E for the environmental variable, and the slope parameter b through which the environmental component will carry on to the response variable); (5) specify the parameters of the variogram (nugget effect, sill, and range) specifying the autocorrelation function SA_R for the response

variable; and (6) specify the variance of the normal error component ε.

A *simulation* designates the generation and analysis of a single data set. A *simulation run* comprises several simulations (typically 1000 in this study, although a few runs involved 10 000 simulations) in which independent data sets were generated using the same population parameters, and analyzed. At the end of a run, the rate of rejection of the ANOVA null hypothesis at the $\alpha = 0.05$ level was calculated, together with a 95% confidence interval. These statistics were used to assess type I error, evaluate the importance of blocking in the analysis of field experimental results, compare various ways of computing the ANOVA F statistic, and estimate the power of the six experimental designs that we studied in the presence of different amounts of spatial autocorrelation.

Type I error is the rate of rejection of the null hypothesis when the data conform to it. In the analysis of variance context, H_0 is true when the statistical populations from which the data have been drawn at random have equal treatment values, hence there is no treatment effect. A test is said to have a correct rate of type I error if, across repeated simulations, the rate of rejection of H_0 is approximately equal to the significance level α used to make the statistical decision. Two problems may occur if the rejection rate is not approximately equal to the significance level. On the one hand, a test whose rate of rejection of H_0 is larger than the significance level of the test, when H_0 is true, is invalid. An invalid test leads to the wrong statistical decision more often than specified by the significance level of the test. On the other hand, a test whose rejection rate is smaller than the significance level remains valid, but its power is reduced (the test is too conservative).

Simulation setup

For each simulation, our program generated a response variable that represented the sum of separate effects, as explained in the *Data generation model* subsection: (1) an explanatory environmental variable with a deterministic structure plus spatial autocorrelation, (2) an effect of the experimental treatments, (3) spatial autocorrelation in the response (biological) variable, and (4) a random error component. Then it conducted an analysis of variance for the specified experimental design and produced a probability associated with the F statistic. The program generated and analyzed as many replicated data sets as required. Results were accumulated over all simulations of a run. A companion program allowed us to create lists of locations for all kinds of experimental designs; a list of locations was fed into the main simulation program together with the parameters specifying each simulation run.

The final statistic for a simulation run was the number of times the null hypothesis (H_0: no effect of the treatments onto the response variable) was rejected throughout the simulations, using the $\alpha = 0.05$ significance level. A 95% confidence interval over the rejection rate was computed. The program also allowed us to obtain output files containing all values of individual simulated surfaces, which we used to draw maps for illustration of the results.

Spatially autocorrelated surfaces were generated using the conditional simulation method, as implemented in subroutine SGSIM of the geostatistical software library GSLIB (Deutsch and Journel 1992). When requested, SA of equal or unequal intensity was added to the environmental and response surfaces. The autocorrelation structure was determined by spherical variograms with nugget values of 0, sill values of 1, and ranges of {0, 4, 16, 40} pixels in the x and y directions, without anisotropy. SA with range of 0 means that no SA was added to the specified surface.

The following parameters were used in the simulations reported in this paper:

1) Overall field size: The simulations were carried out in a 100×100 pixel field.

2) Types of spatial effects in the underlying environment: Four types of environmental variables were generated in different simulation runs:

- Type 0.—Flat surface, plus spatial autocorrelation if requested, no random normal error;
- Type 1.—Flat surface, random normal error, plus spatial autocorrelation if requested;
- Type 2.—Gradients in the x and y directions (values in the range [0, 1]), random normal error, plus spatial autocorrelation if requested;
- Type 3.—Four waves across the field (values in the range [0, 1]), random normal error, plus spatial autocorrelation if requested.

3) Types of spatial effects in the responses: The effect of the environmental variable was carried over to the response variable using a transfer parameter, $b = 1$, corresponding to the parameter β of the model (Eq. 3), which determines the transfer of the effect from the environmental to the response variable (i.e., the degree to which the response variable reflects the environmental variable). Spatial autocorrelation was added to the response surface, as well as standard normal error $\mathcal{N}(0, 1)$ drawn independently at each point of the surface. Treatment values, described in paragraph 5 below, were added to the response surfaces at the locations of the experimental units.

4) Experimental designs and sample sizes: Six experimental designs were used over the 10 000-pixel field; they involved sample sizes of $n = 36$, 81, or 144 experimental units (or plots) inclusive of all three treatments (Fig. 1; Figs. A1 and A2, Appendix A). These values were squared integers; this was a necessary condition for several of the experimental designs used in this study. Values of n reported in the literature are more commonly at the lower end of the range represented by our values $n = 36$ and $n = 81$; $n = 144$ is

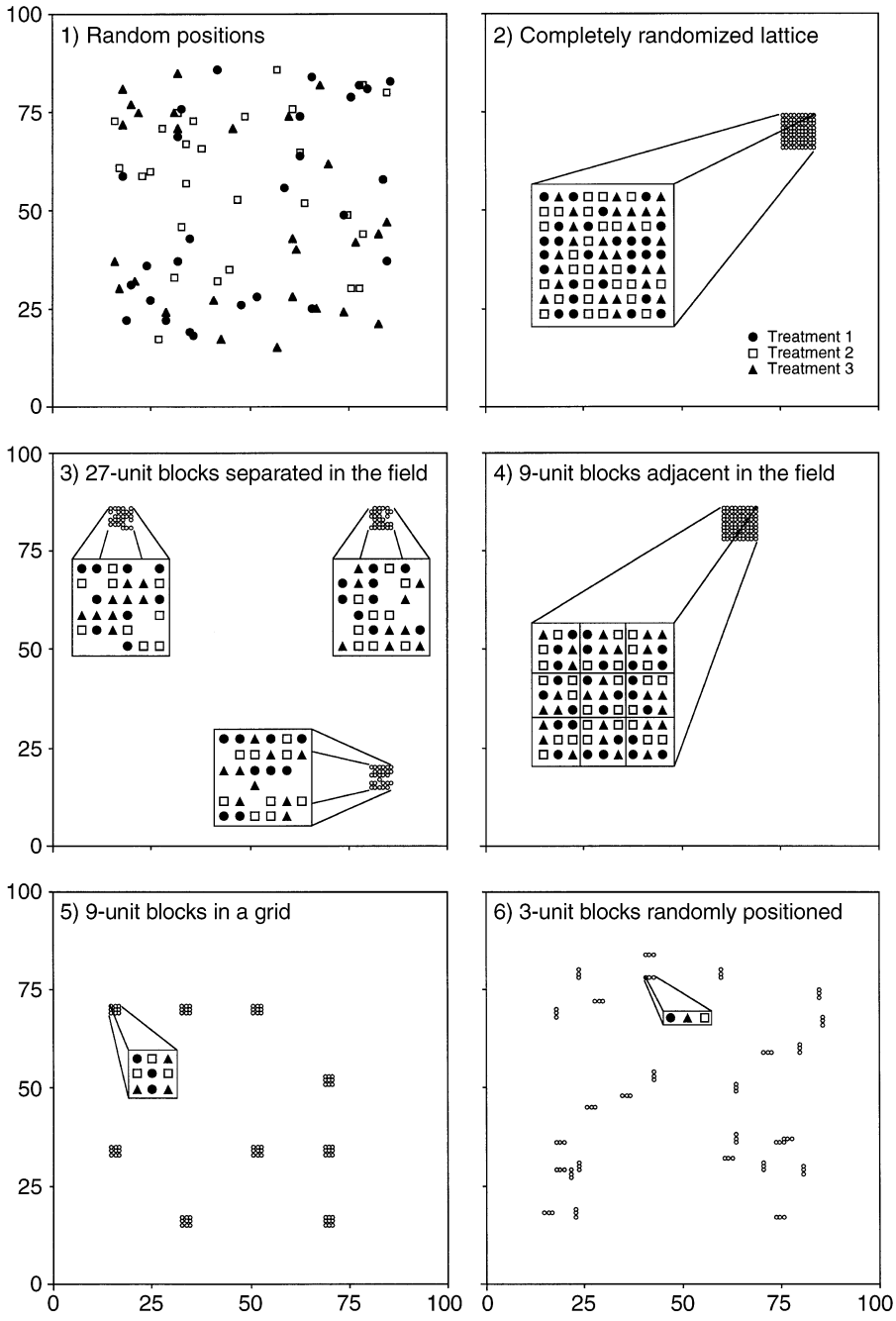

FIG. 1. The six types of experimental designs used in this study are illustrated on maps of the field for $n = 81$ experimental units. The axis units are pixels.

an unusually large experimental effort except for very complicated and extensive studies. The designs involved either random assignment of the treatments to the units (design 1), or randomized complete blocks with (designs 2, 3, 4, 5) or without replication (design 6). The experimental units were located at least 14 pixels away from the borders of the 100 × 100 pixel field to avoid any possible edge effect. The designs were:

Design 1.—Random positions: $n = \{36, 81, 144\}$ randomly located units, with random assignment of treatments to the units;

Design 2.—Completely randomized lattice: a single large lattice of units with random assignment of treatments to the units. The three treatments were each represented by $n = \{12, 27, 48\}$ units. In this design, ANOVA without and with blocking are the same since there is in effect a single large block.

Design 3.—Twenty-seven- or 36-unit blocks separated in the field: {1, 3, 4} big complete blocks, separated in the field, containing n_{tr} = 36, 27, or 36 units respectively, for a total n = 36, 81, or 144. The treatments were equally represented and randomized within each block. Note that for n = 36, design 3 is the same as design 2 (Fig. A2); there are no separate big blocks across the field for this value of n.

Design 4.—Nine-unit blocks adjacent in the field: {4, 9, 16} 3 × 3 complete blocks contiguous in one large square in the field. Each 3 × 3 block contained each treatment three times in a completely randomized design.

Design 5.—Nine-unit blocks in a grid: {4, 9, 16} 3 × 3 complete blocks. The nine sampling units formed a regular grid in each block. The blocks were arranged as a 4 × 4 grid in the field; the grid was complete only for n = 144 units. Each block contained each treatment three times in a completely randomized design.

Design 6.—Three-unit blocks randomly positioned in the field: {12, 27, 48} 1 × 3 blocks, randomly positioned in the field. Each block is a randomized complete block without replication, i.e., containing each treatment once.

The order of the designs reflects an ordering from larger and more compact to smaller and more spread out blocks. The order (1 to 6) was established before the results of the simulations were analyzed.

5) Treatment effects: In the study of type I error, the three treatments had equal values of {1.0, 1.0, 1.0} (effect size = 0). In the study of power, the three treatments had the following values: medium = {1.0, 1.3, 1.6} (effect size = 0.18) and high = {1.0, 1.5, 2.0} (effect size = 0.50). The effect size is the sum of the squared differences of the treatment values from their mean. Treatment values were selected by trial and error to obtain moderate, although measurably different responses in terms of rejection rate of the ANOVA null hypothesis, for the different combinations of types of environmental variables (simulation parameter 2 above), experimental designs (parameter 4 above), and amounts of spatial autocorrelation.

Fig. 2 (top) shows the construction of an environmental surface of type 2. Because of its small variance (s^2 = 0.046), the gradient only accounts for a small fraction of the variance of the environmental surface, in this example, yet its contribution to the response variable is highly significant ($P < 0.0001$); this was also the case over the 100 × 100 pixel surfaces of our simulation study that contained gradients or waves. The gradient is present and highly significant, whether or not that is apparent to the observer. That may correspond more closely to what happens in real field data than if we had generated deterministic surfaces with a pronounced effect. In any case, the emphasis of the present paper is on the effect of spatial autocorrelation on the results of field experiments, not that of the deterministic surfaces.

The construction of a response surface is shown in Fig. 2 (center and bottom). The surface is obtained by adding up, point by point, the values of the environmental surface, the spatially autocorrelated values, the random error, and the treatment values. The treatments had values {1.0, 1.5, 2.0} in that example. The experimental design was four 9-unit blocks in a grid (design 5).

To evaluate the importance of blocking in the analysis of the results, designs 3–6 were repeated without and with taking the blocking into account. Actually, this was achieved by not specifying that the data were blocked.

The simulation effort reported in this paper was the following:

1) For the estimation of type I error: (16 combinations of SA ranges) × (four types of environmental variables) × (five series of simulations for statistic F and three series for statistics $F1$ and $F2$) for n = 36; (16 combinations of SA ranges) × (four types of environmental variables) × (six series of simulations for statistic F and four series for statistics $F1$ and $F2$) for n = 81; (four combinations of SA ranges) × (four types of environmental variables) × (six series of simulations for statistic F and four series for statistics $F1$ and $F2$) for n = 144. Treatment intensities were kept at {1.0, 1.0, 1.0} during type I error simulations. There was a total of 1312 simulation runs. One thousand simulations were carried out in each run, except for two runs that comprised 10 000 simulations.

2) For the power study, the same simulation effort was implemented for two different triplets of treatment effect sizes: {1.0, 1.3, 1.6} and {1.0, 1.5, 2.0}. There was a total of 2624 simulation runs. One thousand simulations were carried out in each run.

The simulation program, written in Fortran, constitutes one of the end products of this work. The source code is available to users who want to develop subroutines to compare different methods of analysis of the data in terms of type I error and power.

Analysis of the simulation results

The ANOVA model for analyzing the results is the following:

$$y_{ijk} = \mu + T_i + B_j + [(TB)_{ij}] + \varepsilon_{ijk} \qquad (4)$$

where y_{ijk} is the value of experimental unit k in the jth block for the experimental unit to which treatment i was applied. The overall mean is μ, T_i is the effect of treatment i and B_j is the effect associated with block j. $(TB)_{ij}$ represents the interaction between blocks and treatments. The brackets indicate that we were not always able to estimate the interaction, either because there was a single block of data (designs 1 and 2) or because there were no replicates of treatments within blocks (design 6). The final term, ε_{ijk}, is the error term of the kth experimental unit in subgroup ij.

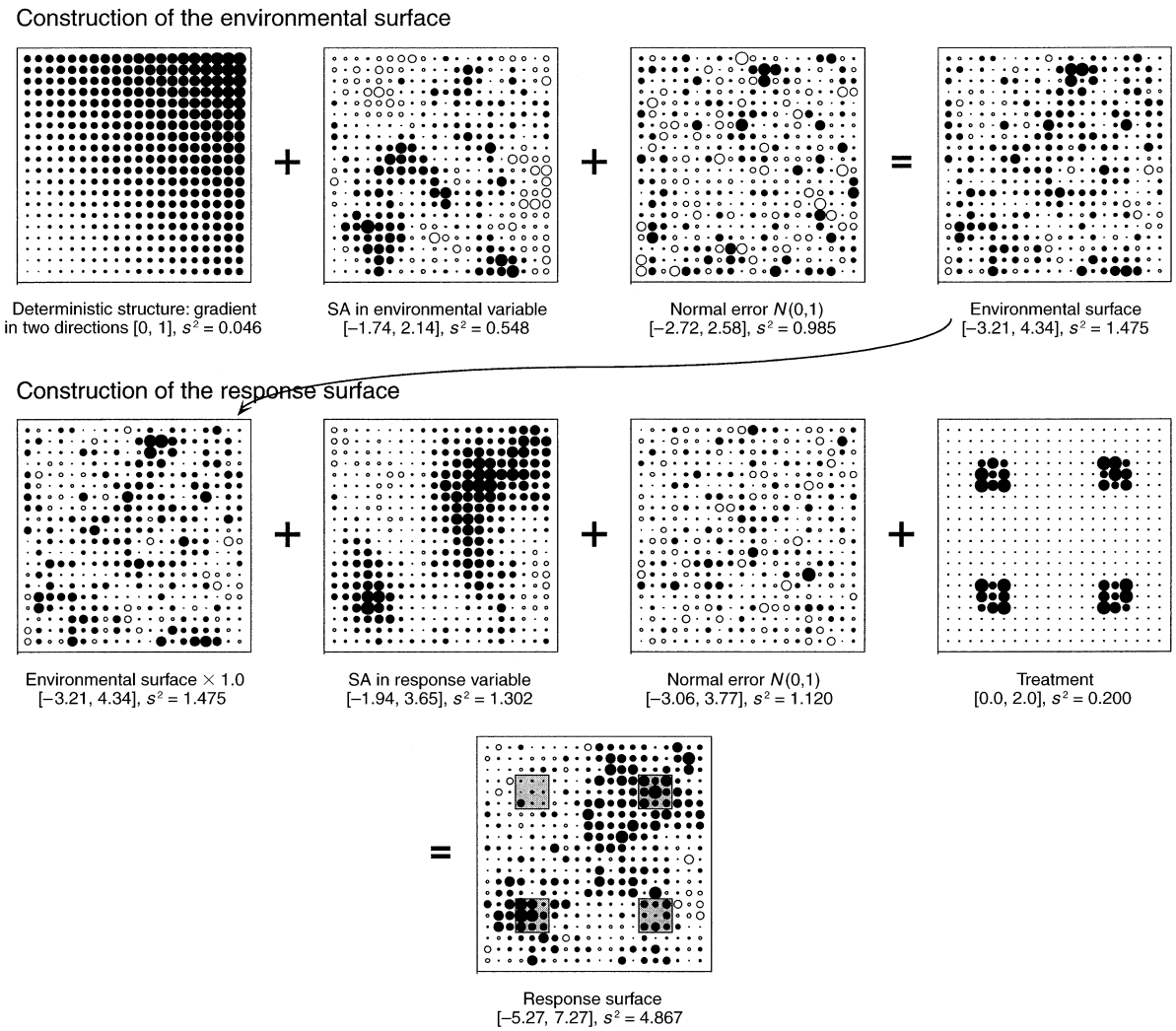

FIG. 2. Construction of the environmental and response surfaces during the simulations. The surface is only 20 × 20 in these illustrations. Bigger open circles represent larger negative values; bigger black circles represent larger positive values. The range of values in each graph is shown in brackets underneath; the variance (s^2) of the values in each surface is also given. Four blocks containing the 36 experimental units are shown in gray in the last panel.

Following Sokal and Rohlf (1995), we considered the blocks to be a random factor. This is the case in most studies, and certainly was in our simulations. If the blocks corresponded to a factor of interest for the study, we would use a crossed two-way model I ANOVA for analysis. Instead, we had a mixed model with a fixed treatment and a random block factor.

When there was more than one block and there was replication of treatments within blocks (experimental designs 3–5), the block × treatment interaction $(TB)_i$ was normally used in the denominator of the F statistic when testing the effect of treatment; we called this statistic $F2$. Alternatively, if the block × treatment interaction is not significant (we can assume that this is the case in our simulations because no interaction was generated in the simulation procedure), one can pool the interaction and residual sums of squares and use the pooled variance estimate as the denominator of the treatment F statistic (called $F1$). Power is increased by having more degrees of freedom attached to the denominator mean square. We verified that a test based on $F1$ has a correct rate of type I error. We also checked if $F1$ always leads to more powerful tests than $F2$, or if the presence of deterministic structures or spatial autocorrelation might alter that. For designs 3–5, our simulation program routinely computed the test statistics in both ways.

There is no replication of treatments within blocks in design 6. In that case, the denominator of statistic $F1$ is simply the residual mean square, and $F2$ cannot be calculated.

A graphical comparison of the results of the ''correct'' analysis that takes blocking into account, and the ''incorrect'' analysis that does not, was first done to

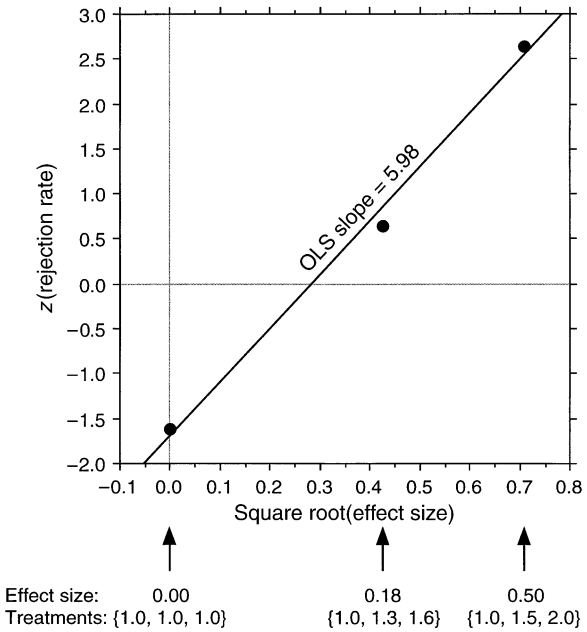

FIG. 3. Calculation of power for experimental design 1 (random positions), no spatial autocorrelation in either variable, environmental variable type 0 (first power value in Table 1, 144 experimental units). The abscissa is the square root of the effect size for the three types of experiments (treatments {1.0, 1.0, 1.0}, {1.0, 1.3, 1.6}, and {1.0, 1.5, 2.0}). The ordinate, *z*(rejection rate), is the probit-transformed rejection rate during the simulations. The slope of the ordinary least-squares (OLS) regression line estimates power.

determine if the blocking structure should be taken into account during the analysis of experimental results when the variables are spatially autocorrelated, as recommended by Dutilleul (1993).

To estimate power, the simulation results were synthesized as follows. For a given set of simulation parameters, and for each triplet of results representing three effect sizes, the rates of rejection of the null hypothesis were subjected to a probit-transformation and regressed on the square roots of the effect sizes. The probit transformation is the back-transformation of the probability under the standard normal error curve into standard normal deviate (*z*) values. The purpose of these transformations is to linearize the relationships. The slope of the ordinary least squares (OLS) regression line was estimated (Fig. 3); larger values of the slope correspond to greater power. The constant 5 may be added to probit-transformed data; this avoids negative values in most cases (Sokal and Rohlf 1995). The slope of the regression line is unaffected by addition of this constant; we did not use it in the construction of Fig. 3.

We expected the experimental designs with smaller and more spread-out blocks to have more power in the presence of spatial autocorrelation. Compared to larger blocks, small blocks will contain less within-block variability due to these factors.

RESULTS

Influence of SA on type I error in simple ANOVA

When the sampling units are distributed at random across the field (experimental design 1), spatial autocorrelation (SA), especially with ranges 16 or 40, influences the results of the tests of significance of simple ANOVA and may render the tests invalid. This is also the case in the presence of a deterministic structure like four waves across the field, even without SA. This effect is stronger in the simulations for $n = 144$ objects (Fig. B1, Appendix B) because the average distance between neighbors is smaller than for $n = 36$ (Fig. B2) or $n = 81$ (Fig. 4).

Experimental design 2 is a lattice of equispaced experimental units in which the treatments are interspersed at random. Spacing of the experimental units is the same for the three values of *n*; it is the size of the experimental square that changes. Simple ANOVA is the only form of analysis that can be used with this design because there is a single block. With $n = 36$ and $n = 81$, the rate of type I error is too conservative in some of the results involving SA or deterministic surfaces; the tests of significance should have reduced power in these situations. For $n = 144$, the rate of type I error is always correct. This design should then be used only for large numbers of experimental units. Such large numbers are rarely used in field experiments for many reasons, including cost and limitations on areas available for experimentation. In any case, this design has low power even for $n = 144$; see the subsection *Which designs lead to tests that have the more power?* below.

With the other designs, in which the experimental units are in blocks distributed across the field, the tests of significance of simple ANOVA remain valid, yet they are highly affected by the presence of spatial autocorrelation. This translates into lower power for these tests (graphs not shown). The tests have correct rates of type I error only when there is no autocorrelation, in neither the environmental nor the response variable.

Lesson learned.—Experimental units should not be positioned at random in the presence of spatial autocorrelation, or if repetitive deterministic structures such as waves are present in the explanatory variables influencing the response. For randomly positioned units, simple ANOVA is the only possible alternative and it may result in incorrect rates of type I error. When the experimental units are distributed into several blocks, the rate of type I error of simple ANOVA is too low in the presence of spatial autocorrelation. This translates into lower power for these tests.

What is the best ANOVA F statistic for blocked designs?

In the presence of spatial autocorrelation, when the experimental units are distributed into several blocks (designs 3–6), simple ANOVA (*F* statistic) has incor-

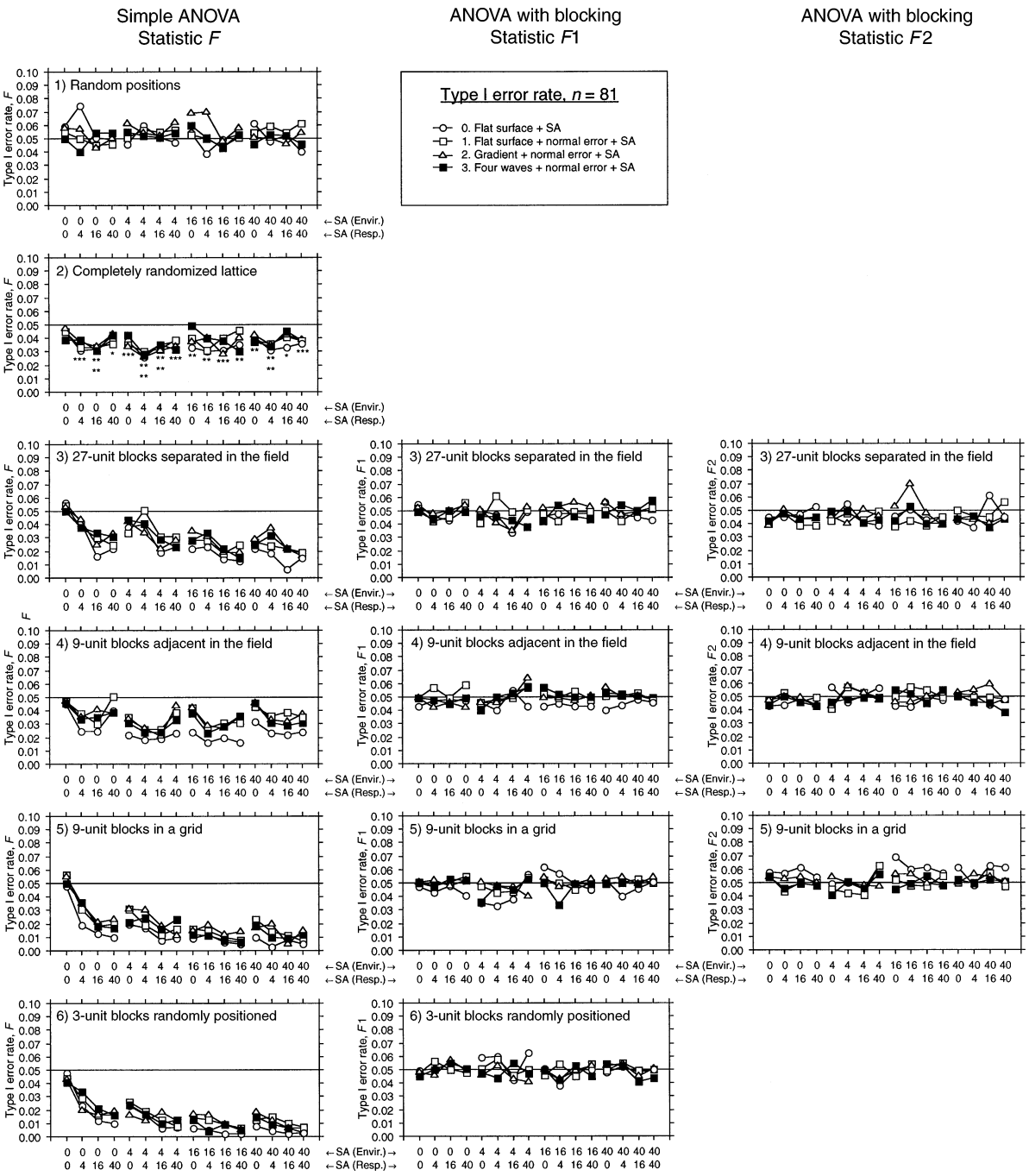

FIG. 4. Rate of type I error in simulations using $n = 81$ experimental units after 1000 simulations. The rows of diagrams contain the results for the six experimental designs, numbered 1–6. The left-hand column shows results for simple ANOVA; the middle and right-hand columns show results for ANOVA taking blocking into account, statistics $F1$ and $F2$. The four types of environmental surfaces are represented by symbols. The abscissa show ranges of values of spatial autocorrelation (SA) in the environmental (Envir.) and response variables (Resp.). Asterisks in panel 2 indicate rejection rates whose confidence intervals do not include the significance level ($\alpha = 0.05$).

rect rates of type I error, as was shown in the previous subsection, whereas ANOVA with blocking always has correct rates of type I error, using either statistic $F1$ or $F2$ (Fig. 4; Figs. B1 and B2, Appendix B; the three forms of F statistics are described in *Methods: Analysis of the simulation results* section). The rate of type I error for design 6, in which there is no replication within the blocks, is as good as that of the other blocked

TABLE 1. Selected estimates of power for designs with 144 experimental units.

Spatial autocorrelation†	Environmental variable‡	Experimental designs			
		(1) Random positions	(2) Completely randomized lattice	(3) 36-U blocks separated in the field (with replication)	
		F	*F*	*F*1	*F*2
0	0	5.98	5.46	6.39	4.52
0	1	3.95	4.01	4.17	3.08
0	2	3.69	4.04	4.17	3.06
0	3	3.77	4.51	3.66	2.75
4	0	3.19	2.95	3.42	2.63
4	1	2.35	2.73	2.97	2.10
4	2	2.51	2.47	3.04	2.19
4	3	2.49	2.82	2.62	1.97
16	0	2.54	4.07	4.60	3.36
16	1	2.07	3.11	3.60	2.55
16	2	2.06	3.13	3.66	2.62
16	3	2.36	3.51	3.17	2.36
40	0	2.82	4.50	5.55	3.87
40	1	2.28	3.46	3.97	2.90
40	2	2.30	3.47	3.92	2.91
40	3	2.32	4.00	3.39	2.61
Mean power		2.918	3.640	3.894	

Notes: Power values are the slopes of the probit-transformed rejection rates as a function of the square root of the effect sizes (Fig. 3); larger slopes mean greater power. Mean powers are only based upon the values of *F* and *F*1. U = experimental unit.

† Spatial autocorrelation is the range of the spherical variogram models for generation of spatial autocorrelation in the environmental and response variables.

‡ Environmental variable types, coded 0–3; see *Methods: Simulation setup*.

designs (designs 3–5) in which there are replicates within blocks. This is the case for all sample sizes investigated in this study.

Even though the tests conducted with statistics *F*1 and *F*2 both have correct type 1 error, they clearly differ in power for designs 3–5. For $n = 144$, the mean ratio of the empirical power estimates (Table 1), power(*F*1)/power(*F*2), is 1.37 for design 3, 1.04 for design 4, and 1.04 for design 5. The ratios increase in size as n decreases: when $n = 81$, the ratios for designs 3, 4, and 5 were 1.55, 1.09, 1.09, respectively (Fig. 5); when $n = 36$, the ratios for designs 4 and 5 were 1.38, 1.35, respectively. The same observations are made in Tables C1 and C2 (Appendix C) for $n = 81$ and $n = 36$. We conclude that statistic *F*1, which uses the pooled (interaction + residual) mean square in the denominator of the treatment *F* statistic, produces tests with higher power than statistic *F*2, which uses the interaction mean square in the denominator.

In real-case studies, one should first test the block × treatment interaction if replicate observations are available per block. This is the case in our designs 3–5. A significant interaction would mean that the effect of treatments differs depending on the blocks; one should not interpret the main effect over all blocks in that case. We did not carry out this preliminary test in the simulation study because no interaction had been generated in the simulation procedure.

Lesson learned.—ANOVA that takes blocking into account is an efficient way of correcting for deterministic structures or spatial autocorrelation in the data. Statistic *F*1, which uses the pooled (interaction + residual) mean square in the denominator of the treatment *F* statistic, produces tests that have a correct rate of type I error, and more power than tests based upon statistic *F*2. It should thus be used to analyze the results of blocked ANOVA if the block × treatment interaction is not significant.

Which designs lead to tests that have the more power?

We will now compare the power estimates obtained using statistic *F* for designs 1 and 2, and statistic *F*1 for designs 3–6, in the presence of different amounts of spatial autocorrelation and different types of environmental variables. Because of its lower power (previous subsection), statistic *F*2 was excluded from the comparisons.

Mean power estimates are given at the bottom-right of Tables 1, C1, and C2 (Appendix C) for $n = 144$, 81, and 36, respectively. The means of the power estimates across all three tables give the following values for designs 1–6: {2.121, 2.604, 2.677, 2.758, 2.761, 2.764}. These values provide an ordering of the experimental designs: power of the ANOVA tests of significance increases from design 1 to design 6. Designs 1 and 2, which do not involve multiple blocks, are clearly the least powerful. Power increases as the number of blocks and their spatial spread increases. Designs 4–6 have nearly equal powers, considering the range

TABLE 1. Extended.

Experimental designs					
(4) 9-U blocks adjacent in the field (with replication)		(5) 9-U blocks arranged in a grid (with replication)		(6) 3-U blocks randomly positioned (unreplicated)	
*F*1	*F*2	*F*1	*F*2	*F*1	Mean power
5.55	5.35	5.90	5.51	6.13	5.902
3.88	3.70	4.01	3.91	3.91	3.988
3.88	3.71	4.03	3.93	3.95	3.960
3.96	3.82	3.86	3.70	3.93	3.948
3.79	3.69	3.78	3.63	3.82	3.492
3.00	2.90	3.04	2.85	3.14	2.872
3.14	3.04	2.97	2.77	3.03	2.860
2.80	2.75	2.93	2.87	3.11	2.795
4.89	4.78	4.98	4.82	5.49	4.428
3.60	3.38	3.59	3.44	3.62	3.265
3.56	3.36	3.82	3.73	3.70	3.322
3.69	3.42	3.53	3.39	3.45	3.285
5.16	5.27	5.39	5.10	6.11	4.922
3.69	3.46	3.97	3.78	3.80	3.528
3.71	3.62	3.97	3.74	3.74	3.518
3.83	3.66	3.69	3.64	3.73	3.493
3.833		3.966		4.041	3.724

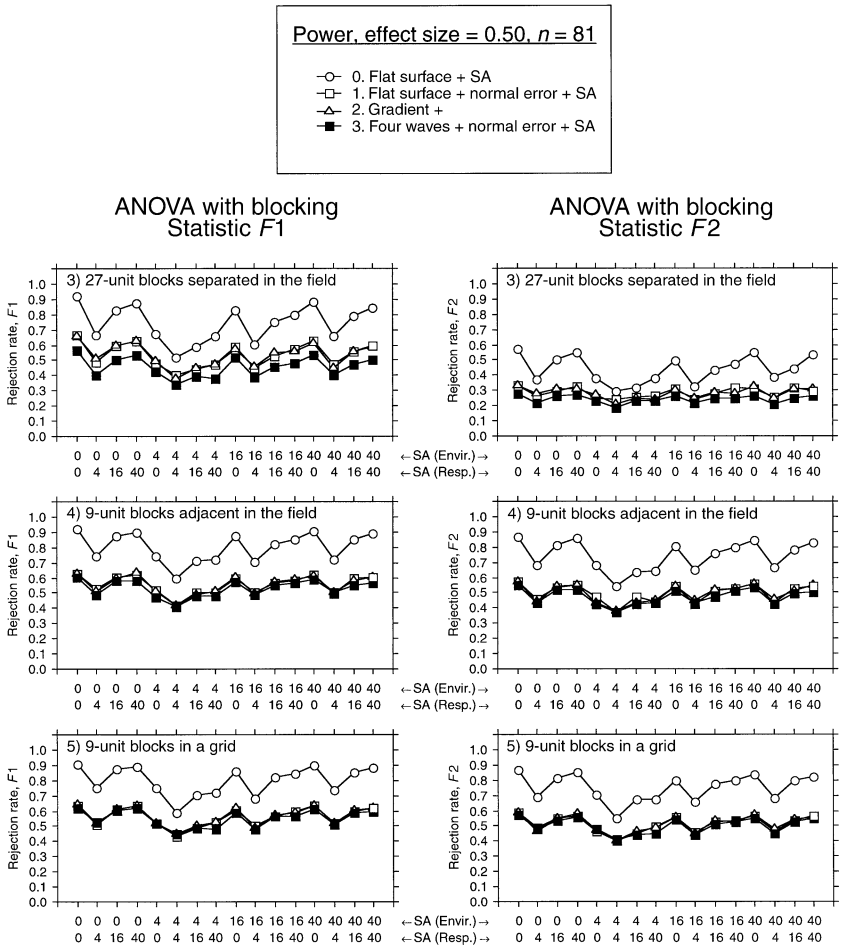

FIG. 5. Power simulation results for $n = 81$ experimental units, high effect (effect size = 0.50): rates of rejection of H_0 over 1000 simulations. The rows of diagrams correspond to designs 3, 4, and 5, where statistics *F*1 and *F*2 were computed.

TABLE 2. Mean power estimates for groups of simulations with different amounts of spatial autocorrelation (stated by the range of the spherical variogram models) and types of environmental variables, as a function of the number of experimental units (n) used in the simulations.

Simulation	Number of experimental units (n)		
	144	81	36
Range of spherical variogram models for generation of spatial autocorrelation			
0	4.450	3.120	1.728
4	3.005	2.111	1.197
16	3.575	2.586	1.441
40	3.865	2.760	1.532
Environmental variable type			
0 = flat surface + SA	4.686	3.401	1.922
1 = flat surface + $\mathcal{N}(0, 1)$ + SA	3.413	2.395	1.316
2 = gradient + $\mathcal{N}(0, 1)$ + SA	3.415	2.436	1.354
3 = waves + $\mathcal{N}(0, 1)$ + SA	3.380	2.346	1.307

Notes: $\mathcal{N}(0, 1)$ indicates random standard normal error; SA indicates spatial autocorrelation.

of situations (deterministic structures and spatial autocorrelation) included in our simulations.

This conclusion only applies to simulations conducted in the presence of spatial autocorrelation. One can compare the individual power estimates to the row means shown in the right-hand column of each table. In the four rows with SA = 0, at the tops of the tables, the power estimates that are larger than the row mean are found in any positions in a row. The coefficient of variation of these rows is also much smaller than in the following rows, which present a gradient of power from left to right. For all practical purposes, the six experimental designs have equal power in the absence of spatial autocorrelation.

Lesson learned.—For constant effort (n), experimental designs that have more, smaller blocks, which are more broadly spread across the experimental area, lead to tests of significance that have more power in the presence of spatial autocorrelation. In the absence of spatial autocorrelation, however, the six experimental designs investigated in this study lead to tests of significance that have equivalent powers.

Influence of spatial autocorrelation on power

Table 2 shows the estimates of mean power for groups of simulations with different amounts of spatial autocorrelation and types of environmental variables. The simulations without SA all have higher power than their counterparts with SA. Short-range SA affects power the most; power increases with the range of SA. This is due to the fact that fine-scale (small range) SA is more likely to create heterogeneity within the blocks than broad-scale (large range) SA. The simulations without normal error in the environmental variable (type 0) have higher power than those with normal error (types 1–3), as can be expected; this type of environmental variable was included in the study to provide a bottom line against which the effect of SA could be assessed. There is no noticeable difference in power among the three types of environmental variables that included normal error (types 1–3).

Overall mean power for a given experimental effort is shown at the bottom-right of Tables 1, C1, and C2. Power increases with the number of experimental units. This result was expected. It provides reassurance, though, about the calculation method that we used to estimate power (Fig. 3).

Lesson learned.—Short-ranged spatial autocorrelation affects the power of the ANOVA tests of significance more than large-ranged spatial autocorrelation.

DISCUSSION

The simulation results presented in this paper lead to the following recommendations:

1) Randomly positioned experimental units (i.e., completely randomized design) should only be used when the experimental area is homogeneous at broad scale. It should not be used when spatial autocorrelation, or repetitive deterministic structures such as waves, are present.

2) Blocking is an efficient way of correcting for the effect of spatial autocorrelation in the data. ANOVA that takes blocking into account should be used to analyze the results of such experiments, as suggested by Dutilleul (1993).

3) Statistic $F1$, which uses the pooled (interaction + residual) mean square in the denominator of the treatment F statistic, should be used to analyze the results of blocked ANOVA if the block × treatment interaction is not significant.

4) For constant effort, experimental designs that have a larger number of smaller-sized blocks, more widely spread across the experimental area, lead to tests of significance that have more power in the presence of spatial autocorrelation. This is because small blocks are more homogeneous than large ones.

5) In the absence of spatial autocorrelation, the six experimental designs investigated in this study lead to tests of significance with equivalent powers.

6) Fine-scale (short-ranged) spatial autocorrelation affects the power of the ANOVA tests of significance more than broad-scale (large-ranged) spatial autocorrelation.

Dutilleul (1993) wrote that a completely randomized design (our design 1) should only be used in experiments in which the field is homogeneous at broad scale. Broad-ranged spatial autocorrelation rendered the tests of significance of simple ANOVA invalid in many of our simulations, especially those involving 144 experimental units. The experiments conducted in a single completely randomized block (design 2) with $n = 81$ or 36 experimental units were also badly affected by the presence of spatial autocorrelation, which reduced the rate of type I error and the power of the tests (graphs not shown).

Except for design 1 (random positions), all our designs involved randomized complete blocks. If, for practical reasons such as field space availability, or because there are too many treatments, this cannot be done, more complex designs and analyses, called *incomplete blocks*, must be used. The results of our simulations, summarized in the *Lesson learned* paragraphs of the *Results*, can help ecologists design such experiments in the presence of spatial autocorrelation.

The value that we chose for the transfer parameter, $b = 1$ (see *Methods: Simulation setup*), is responsible for the fact that the presence and deterministic shape of the environmental variables did not play a large role in the results of the simulations. Small effects of the environmental variables may correspond to the conditions encountered in many field experiments in which researchers generally tend to minimize the differences in environmental conditions when setting up an experiment; this is one of the meanings of ''control'' in experimental design (Hurlbert 1984). We expect environmental variables with high transfer parameters to have an effect on power similar to that of spatial autocorrelation with similar range values; a linear gradient would have an effect similar to that of spatial autocorrelation with infinite range (linear variogram model), whereas four waves across the field would have an effect similar to spatial autocorrelation produced by using a spherical variogram with range 1/8 the width of the field, or 12.5 pixels in the 100×100 pixel field that we used in our simulations.

Before a field experiment, a pilot study can be extremely beneficial in helping to determine the environmental factors creating variation across the experimental area that are likely to affect the response variable. If such variables are found, they should be measured at each experimental unit and incorporated in the analysis as covariables. In the simulations reported in this paper, however, we simulated the behavior of the ecologist who does not take the environmental variables into account during analysis of experimental results.

ACKNOWLEDGMENTS

The work reported in this paper was conducted as part of the Working Group ''Integrating the Statistical Modeling of Spatial Data in Ecology'' supported by the National Center for Ecological Analysis and Synthesis (NCEAS), a Center funded by NSF (Grant #DEB-94-21535), the University of California at Santa Barbara, and the State of California. Access to the computers of the Environnement Scientifique Intégré (ESI) of Université de Montréal for our simulation work is gratefully acknowledged. Michael Hohn, West Virginia Geology and Economy Survey, and Donald E. Myers, Department of Mathematics, University of Arizona at Tucson, contributed to the discussions that led to this paper. Philip Dixon, Department of Statistics, Iowa State University, suggested the method depicted in Fig. 3 to estimate power.

LITERATURE CITED

Bartlett, M. S. 1978. Nearest neighbour models in the analysis of field experiments. Journal of the Royal Statistical Society, Series B **40**:147–174.

Bonham, C. D., and R. M. Reich. 1999. Influence of spatial autocorrelation on a fixed-effect model used to evaluate treatment of oil spills. Applied Mathematics and Computation **106**:149–162.

Casler, M. D. 1999. Spatial variation affects precision of perennial cool-season forage grass trials. Agronomy Journal **91**:75–81.

Dale, M. R. T., and M.-J. Fortin. 2002. Spatial autocorrelation and statistical tests in ecology. 2002. Écoscience **9**:162–167.

Deutsch, C. V., and A. G. Journel. 1992. GSLIB: Geostatistical software library and user's guide. Oxford University Press, New York, New York, USA.

Dutilleul, P. 1993. Spatial heterogeneity and the design of ecological field experiments. Ecology **74**:1646–1658.

Dutilleul, P. 1998. Incorporating scale in ecological experiments: study design. Pages 369–386 *in* D. L. Peterson and V. T. Parker, editors. Ecological scale: theory and applications. Columbia University Press, New York, New York, USA.

Fisher, R. A. 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture **33**:503–513.

Galton, F. 1898. Natural inheritance. Macmillan, London, UK.

Hahn, G. J. 1982. Design of experiments: an annotated bibliography. Pages 359–366 *in* S. Kotz and N. L. Johnson, editors. Encyclopedia of statistical sciences. Volume 2. Wiley, New York, New York, USA.

Hoosbeek, M. R., A. Stein, H. van Reuler, and B. H. Janssen. 1998. Interpolation of agronomic data from plot to field scale: using a clustered versus a spatially randomized block design. Geoderma **81**:265–280.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs **54**: 187–211.

Keitt, T. H., O. N. Bjornstad, P. M. Dixon, and S. Citron-Pousty. 2002. Accounting for spatial pattern when modeling organism–environment interactions. Ecography **25**: 616–625.

Legendre, P., M. R. T. Dale, M.-J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography **25**:601–615.

Sokal, R. R., N. L. Oden, B. A. Thomson, and J. Kim. 1993. Testing for regional differences in means: distinguishing inherent from spurious spatial autocorrelation by restricted randomization. Geographical Analysis **25**:199–210.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. Third edition. W. H. Freeman, New York, New York, USA.

Underwood, A. J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, UK.

van Es, H. M., and C. L. van Es. 1993. Spatial nature of randomization and its effect on the outcome of field experiments. Agronomy Journal **85**:420–428.

Ver Hoef, J. M., and N. Cressie. 2001. Spatial statistics: analysis of field experiments. Pages 289–307 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Second edition. Oxford University Press, New York, New York, USA.

APPENDIX A

Figures showing the six types of experimental designs used in this study, for $n = 144$ (Fig. A1) and $n = 36$ experimental units (Fig. A2), are available in ESA's Electronic Data Archive: *Ecological Archives* E085-108-A1.

APPENDIX B

Figures showing type I error in simulations involving $n = 144$ (Fig. B1) and $n = 36$ experimental units (Fig. B2) are available in ESA's Electronic Data Archive: *Ecological Archives* E085-108-A2.

APPENDIX C

Tables showing power estimates for $n = 81$ (Table C1) and $n = 36$ experimental units (Table C2) are available in ESA's Electronic Data Archive: *Ecological Archives* E085-108-A3.

Appendices to:

Legendre, P., M. R. T. Dale, M.-J. Fortin, P. Casgrain, and J. Gurevitch. 2004. Effects of spatial structures on the results of field experiments. *Ecology* 85: 3202–3214.

APPENDIX A

Ecological Archives E085-108-A1

FIGURES SHOWING THE SIX TYPES OF EXPERIMENTAL DESIGNS USED IN THIS STUDY, FOR $N = 144$ (FIG. A1) AND $N = 36$ EXPERIMENTAL UNITS (FIG. A2)

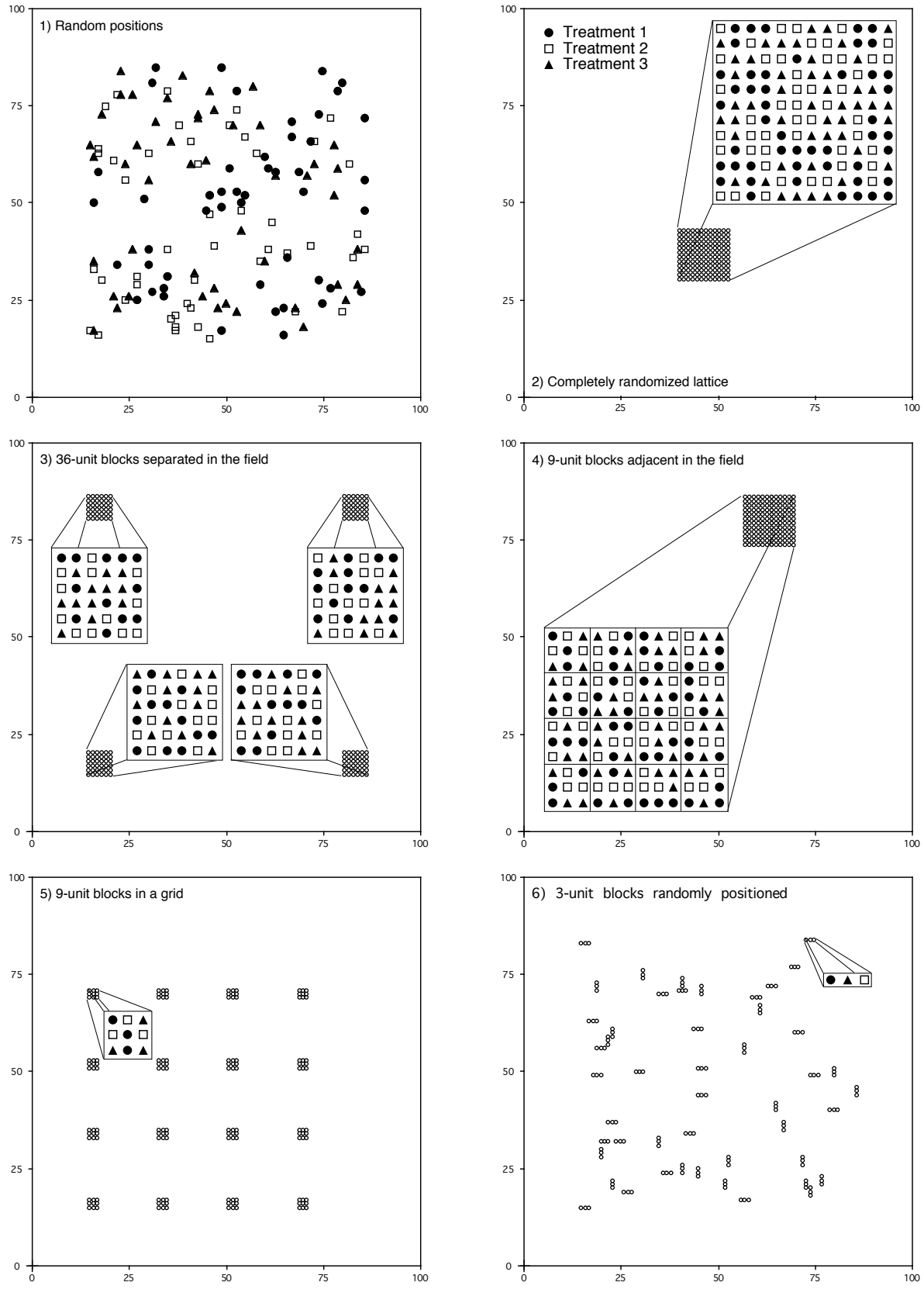

FIG. A1. The six types of experimental designs used in this study, for $n = 144$ experimental units.

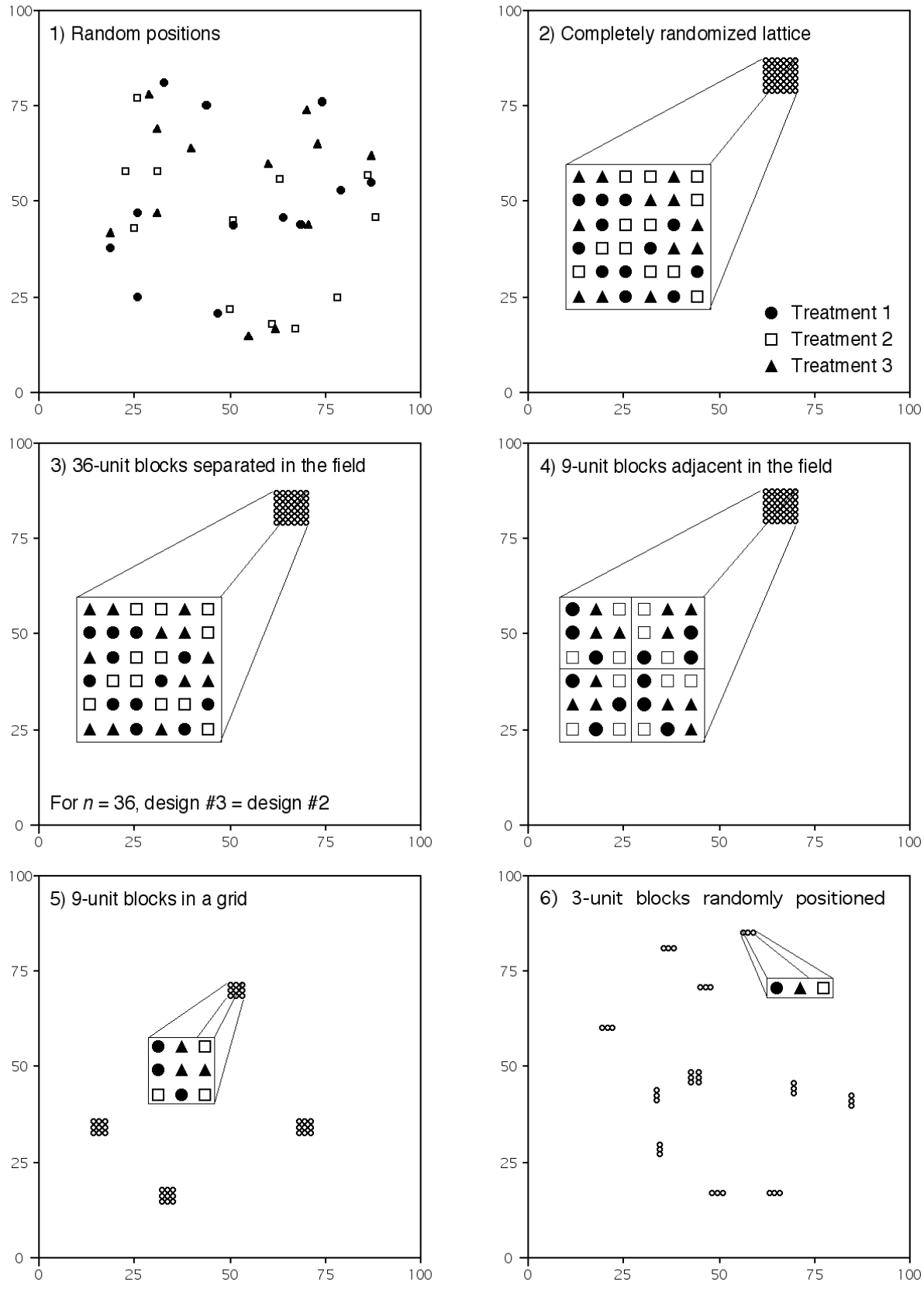

FIG. A2. The six types of experimental designs used in this study, for $n = 36$ experimental units.

APPENDIX B

Ecological Archives E085-108-A2

FIGURE SHOWING TYPE I ERROR IN SIMULATIONS
INVOLVING $N = 144$ (FIG. B1) AND $N = 36$ EXPERIMENTAL UNITS (FIG. B2)

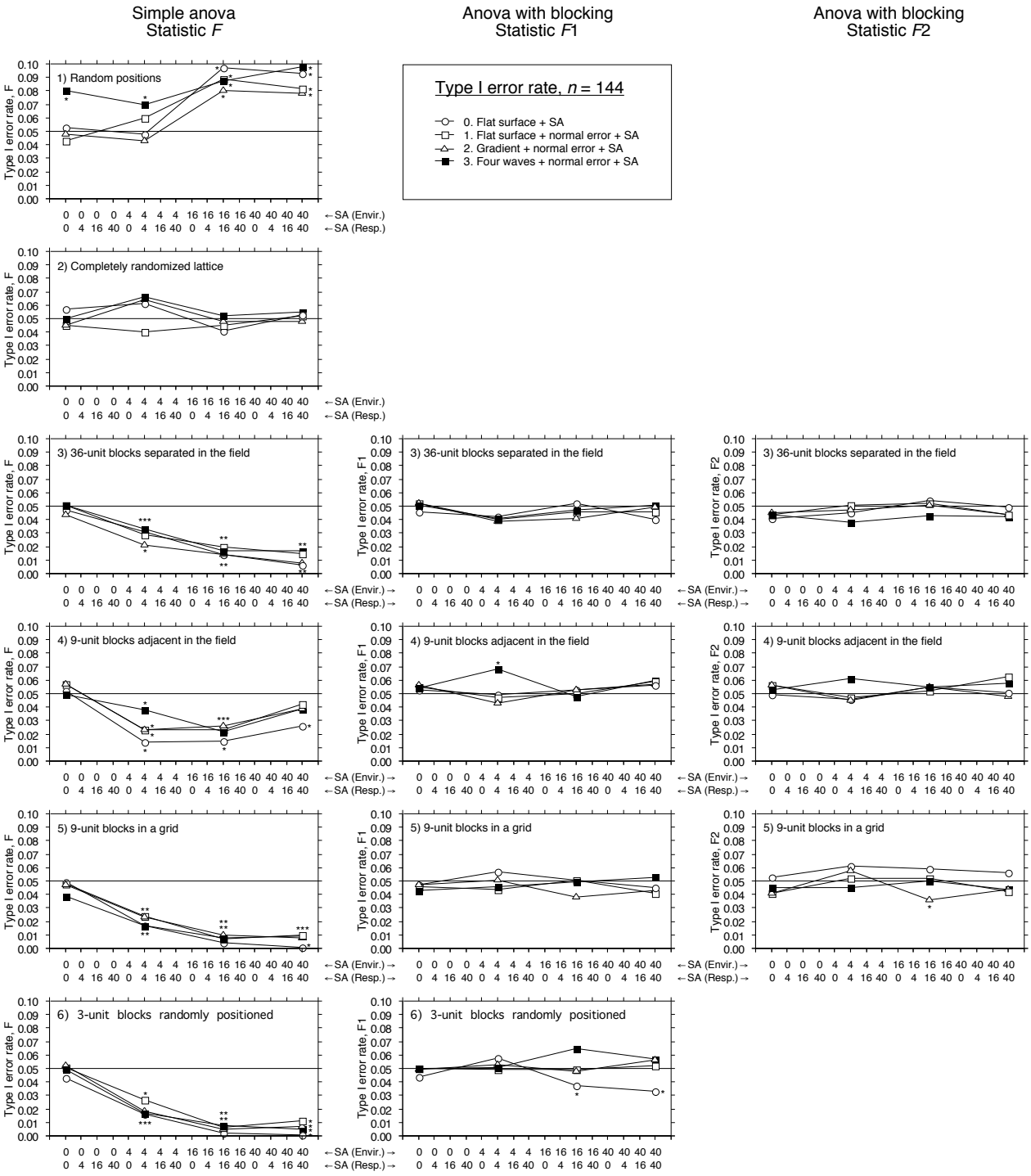

FIG. B1. Rate of type I error in simulations involving $n = 144$ experimental units after 1000 simulations. The rows of diagrams contain the results for the 6 experimental designs, numbered 1 to 6. Left-hand column: results for simple ANOVA. Middle and right-hand column: results for ANOVA taking blocking into account, statistics $F1$ and $F2$. The 4 types of environmental surfaces are represented by symbols. Abscissa: ranges of values of spatial autocorrelation (*SA*) in the environmental (Envir.) and response variables (Resp.). Asterisks: rejection rates whose confidence intervals do not include the significance level ($\alpha = 0.05$).

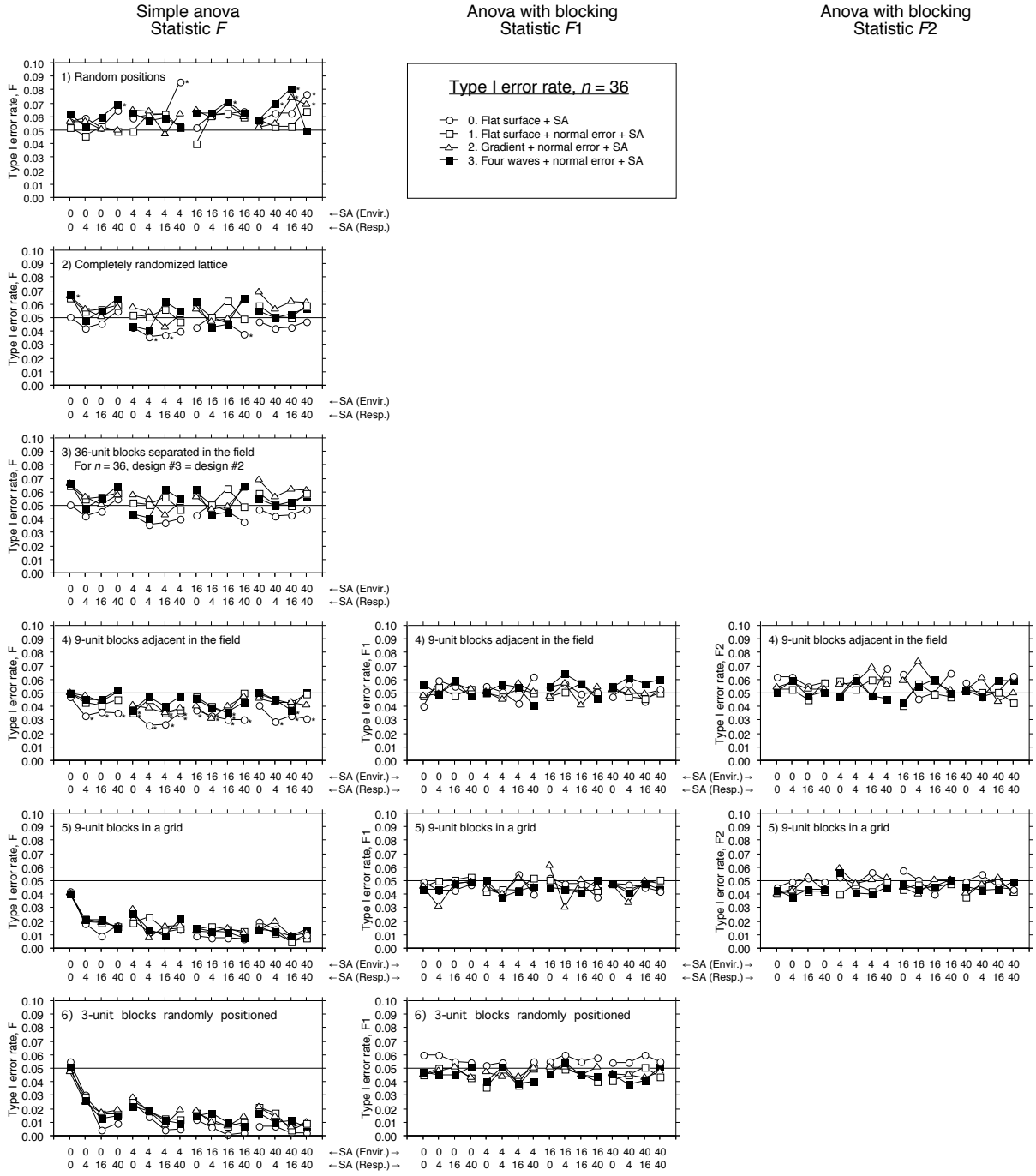

FIG. B2. Rate of type I error in simulations involving n = 36 experimental units after 1000 simulations. The rows of diagrams contain the results for the 6 experimental designs, numbered 1 to 6. Left-hand column: results for simple ANOVA. Middle and right-hand column: results for ANOVA taking blocking into account, statistics $F1$ and $F2$. The 4 types of environmental surfaces are represented by symbols. Abscissa: ranges of values of spatial autocorrelation (*SA*) in the environmental (Envir.) and response variables (Resp.). Asterisks in panels 1-4: rejection rates whose confidence intervals do not include the significance level ($\alpha = 0.05$).

APPENDIX C

Ecological Archives E085-108-A3

TABLES SHOWING POWER ESTIMATES
FOR N = 81 (TABLE C1) AND N = 36 EXPERIMENTAL UNITS (TABLE C2)

TABLE C1. Selected estimates of power for designs with 81 experimental units. Power values are the slopes of the probit-transformed rejection rates as a function of the square root of the effect sizes; larger slopes mean greater power. Mean powers are only based upon the values of *F* and *F*1.

Experimental ⇒ designs		#1 Random positions	#2 Completely randomized lattice	#3 36-U blocks separated in the field (with replic.)	#4 9-U blocks adjacent in the field (with replic.)	#5 9-U blocks arranged in a grid (with replic.)	#6 3-U blocks randomly positioned (unreplicated)	
F-statistics ⇒		*F*	*F*	*F*1 (*F*2)	*F*1 (*F*2)	*F*1 (*F*2)	*F*1	
Spatial autocorr.	Environ. variable							Mean power
0	0	4.20	4.19	4.19 (2.65)	4.39 (3.98)	4.23 (3.75)	4.10	4.217
0	1	2.71	2.80	2.86 (1.84)	2.75 (2.59)	2.78 (2.54)	2.71	2.768
0	2	2.91	2.81	2.84 (1.86)	2.75 (2.59)	2.80 (2.56)	2.71	2.803
0	3	2.85	2.68	2.51 (1.60)	2.66 (2.55)	2.73 (2.50)	2.73	2.693
4	0	2.07	2.41	2.41 (1.47)	2.80 (2.50)	2.91 (2.48)	2.52	2.520
4	1	1.71	2.05	1.79 (1.32)	2.05 (1.75)	2.16 (2.09)	1.95	1.952
4	2	1.81	2.00	1.98 (1.32)	2.06 (1.76)	2.13 (1.95)	2.12	2.017
4	3	1.80	1.83	1.76 (1.05)	1.97 (1.85)	2.14 (1.97)	2.23	1.955
16	0	2.20	3.23	3.26 (2.19)	3.69 (3.33)	3.61 (3.23)	3.47	3.243
16	1	1.82	2.47	2.41 (1.65)	2.45 (2.26)	2.60 (2.43)	2.53	2.380
16	2	1.96	2.36	2.38 (1.54)	2.59 (2.36)	2.57 (2.41)	2.54	2.400
16	3	2.10	2.27	2.20 (1.48)	2.44 (2.27)	2.54 (2.26)	2.37	2.320
40	0	2.41	3.72	3.81 (2.47)	4.07 (3.68)	3.94 (3.48)	3.79	3.623
40	1	1.73	2.56	2.61 (1.44)	2.69 (2.45)	2.73 (2.57)	2.56	2.480
40	2	1.98	2.58	2.62 (1.68)	2.69 (2.50)	2.66 (2.50)	2.61	2.523
40	3	2.02	2.42	2.18 (1.49)	2.54 (2.48)	2.64 (2.45)	2.69	2.415
Mean power		2.268	2.649	2.613	2.787	2.823	2.727	2.644

TABLE C2. Selected estimates of power for designs with 36 experimental units. For $n = 36$, columns 2 and 3 are identical. Power values are the slopes of the probit-transformed rejection rates as a function of the square root of the effect sizes; larger slopes mean greater power. Mean powers are only based upon the values of F and $F1$.

Experimental ⇒ designs		#1 Random positions	#2 Completely randomized lattice	#3 36-U blocks separated in the field (with replic.)	#4 9-U blocks adjacent in the field (with replic.)	#5 9-U blocks arranged in a grid (with replic.)	#6 3-U blocks randomly positioned (unreplicated)	
F-statistics ⇒		F	F	F	$F1$ ($F2$)	$F1$ ($F2$)	$F1$	
Spatial autocorr.	Environ. variable							Mean power
0	0	2.42	2.42	2.42	2.61 (1.66)	2.39 (1.86)	2.16	2.403
0	1	1.75	1.38	1.38	1.59 (1.19)	1.35 (1.02)	1.54	1.498
0	2	1.89	1.38	1.38	1.59 (1.19)	1.37 (1.02)	1.56	1.528
0	3	1.07	1.69	1.69	1.64 (1.35)	1.37 (1.04)	1.44	1.483
4	0	1.04	1.46	1.46	1.54 (0.99)	1.54 (1.14)	1.46	1.417
4	1	0.98	1.06	1.06	1.16 (0.70)	1.12 (0.78)	1.19	1.095
4	2	0.97	1.22	1.22	1.22 (0.81)	1.10 (0.86)	1.25	1.163
4	3	0.72	1.24	1.24	1.30 (0.97)	1.11 (0.79)	1.07	1.113
16	0	1.16	1.87	1.87	2.06 (1.55)	2.13 (1.65)	1.91	1.833
16	1	1.10	1.30	1.30	1.35 (1.11)	1.34 (0.99)	1.38	1.295
16	2	1.20	1.22	1.22	1.54 (1.02)	1.35 (0.98)	1.47	1.333
16	3	0.56	1.54	1.54	1.39 (1.02)	1.38 (0.91)	1.40	1.302
40	0	1.07	2.25	2.25	2.25 (1.60)	2.26 (1.74)	2.13	2.035
40	1	1.05	1.42	1.42	1.49 (1.28)	1.34 (0.99)	1.54	1.377
40	2	1.12	1.44	1.44	1.41 (1.13)	1.38 (1.05)	1.55	1.390
40	3	0.75	1.50	1.50	1.54 (1.19)	1.35 (1.00)	1.33	1.328
Mean power		1.178	1.524	1.524	1.605	1.493	1.524	1.475

Notes: U = experimental unit; replic. = replication. "Spatial autocorr." is the range of the spherical variogram models for generation of spatial autocorrelation in the environmental and response variables. "Environ. variable": environmental variable types, coded 0-3: see *Simulation setup* in the Methods.