Journal of Applied Ecology 2003
40, 17–31

METHODOLOGICAL INSIGHTS

Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops

JOE N. PERRY, PETER ROTHERY*, SUZANNE J. CLARK, MATT S. HEARD* and CATHY HAWES†

*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK; *NERC Centre for Ecology and Hydrology, Monks Wood, Abbots Ripton, Huntingdon, Cambridgeshire PE28 2LS, UK; and †The Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK*

Summary

1. The effects on British farmland wildlife of the management of four genetically modified herbicide-tolerant crops are currently being studied in a 5-year trial termed the Farm-Scale Evaluations (FSE), the first 4 years of which are completed. The FSE is controversial and extensive. There has been intense scrutiny of the experimental design and proposed analysis, and of the estimated statistical power to detect effects of a given magnitude, should any exist.

2. For each crop, the FSE is a form of on-farm trial with a single composite null hypothesis and a simple randomized block experimental design. This has statistical implications for the imposition of treatments by growers and the need for proper randomization. The choice of a half-field experimental unit was based on field availability, the focus on herbicide management, the need to reduce variability and efficiency gains in sampling effort. Farms and fields were selected to represent the range of variability of geography and intensiveness across Britain for each crop.

3. Results of a power analysis suggested that the planned replication of the FSE of about 60 fields per crop over 3 years would be sufficient to provide useful information, from which valid statistical inferences could be drawn. The achieved replication for spring crops in the FSE exceeded, by more than threefold, that in any of 82 comparable terrestrial manipulative ecological experiments undertaken previously.

4. Here, we exemplify a range of analyses including covariates, interactions between various factors including years and treatments, diagnostic procedures to aid selection of the most efficient statistical model, the estimation of power from coefficients of variation, a novel and apparently robust test statistic and the calculation of overall variance from within- and between-unit variability. Preliminary results indicated that a simple log-normal model appeared adequate for most analyses.

5. *Synthesis and applications.* Statistical challenges created by the scope of the FSE were resolved from a sound knowledge of good experimental design. There is an urgent need for further statistical studies to develop experimental designs or modelling approaches that allow similar studies of genetically modified (GM) crops, at reduced cost. However, this power analysis has shown that this cannot be achieved at the expense of adequate replication, essential for all risk assessment studies.

Key-words: abundance, biodiversity, biometry, GM crops, plant and invertebrate populations, statistical power analysis.

Journal of Aplied Ecology (2003) **40**, 17–31

Correspondence: Joe N. Perry, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK (fax +44 1582760981; e-mail joe.perry@bbsrc.ac.uk).

Introduction

The first genetically modified (GM) crops under consideration for commercial planting in the UK have been altered to make them less sensitive to broad-spectrum herbicides. In 1998, English Nature, the statutory body set up to promote the conservation of England's wildlife, raised concerns that the management of these genetically modified herbicide-tolerant (GMHT) crops could result in reductions of plant and invertebrate populations on which farmland birds and other farmland wildlife depend (Anonymous 1998).

There is evidence that farmland wildlife has already been affected deleteriously by the intensification of agriculture (Krebs *et al*. 1999; Robinson & Sutherland 2002). On the one hand, the introduction of GMHT crops might exacerbate this situation by allowing greater use of herbicide in farmland. This would result in fewer plants for insects to live on, and consequently fewer insect prey for farmland birds. Alternatively, the use of GMHT crops may allow more precise weed control, allowing plants to remain longer in the crop. GMHT herbicide management might thereby increase the abundance and diversity of farmland wildlife compared with herbicide use in equivalent conventional crops. To distinguish between these alternatives the Department for the Environment, Food and Rural Affairs (DEFRA) and the Scottish Executive have funded a 5-year study termed the Farm-Scale Evaluations (FSE) to provide a thorough understanding of the environmental effects of growing GMHT crops (Firbank *et al*. 1999). This is being conducted by a consortium of public sector research institutes (Firbank *et al*. 2003). It began in spring 1999 with a pilot year to develop protocols; the evaluations proper began in spring 2000. Three spring-sown crops, spring oilseed rape, fodder maize and beet (sugar and fodder), were studied in each of 3 years, 2000, 2001 and 2002. Also, one autumn-sown crop, winter oilseed rape, was sown in those years and the third year's data for this crop will be collected in 2003. The taxa studied are plants and invertebrates.

For each crop the FSE aims to test a specific null hypothesis: that there is no difference between the management of GMHT varieties and that of comparable conventional varieties, in their effect on the abundance and diversity of arable plants and invertebrates. The alternative hypothesis is that there is a difference in abundance or diversity, in either possible direction; all tests are therefore two-tailed.

Effects are likely to be indirect resulting from crop management, rather than from the direct effect of the use of GM plant breeding technology. Indeed, had herbicide resistance been introduced to the experimental crops by traditional breeding, the design of the study would have been the same. Farmers grow and manage both GM and conventional crops as closely as possible to commercial practice. The FSE is one of the most controversial ecological experiments proposed in Britain and perhaps the most extensive ever attempted; 272 fields have been sown, over four crops and 3 years. There has therefore been intense scrutiny of its design and analysis and of its estimated statistical power to detect effects of a given magnitude. This study focused on three major biometrical issues: design, analysis and power. An overview of the project is given by Firbank *et al*. (2003).

Design

THE CHOSEN DESIGN

In statistical terms the chosen design for the FSE is straightforward. It is a paired-comparison experiment in a randomized block design, with a single treatment factor at two levels. Each block is a single field; in any year, each field is sited on a different farm. The two levels of the treatment factor are GMHT crop management and conventional crop management. There are two experimental units per block, comprising two halves of the same field. The GMHT crop is allocated randomly to one half-field and the conventional crop to the other.

COMPOSITE TREATMENTS

The form of the FSE null hypothesis dictated that treatments be chosen deliberately to represent a composite of agronomic effects, not a single ecological process. Any feature of the crop itself, such as a varietal trait, and any concomitant agronomic practice linked to the crop concerned, such as recommended herbicide usage, would contribute towards the potential treatment effect being measured. Such practices, tied to the crop, had therefore to be allocated to units as part of the identical process whereby the treatments were randomized. Composite null hypotheses are often used in initial studies, to demonstrate the existence and estimate the magnitude of effects and thereby to screen out those unworthy of further interest. In such experiments, the most important property is of realism and applicability, so that the results relate unequivocally to the system that is studied. The FSE was designed as a large-scale experiment of this kind.

HALF-FIELDS VS. PAIRED-FIELDS: CHOICE OF EXPERIMENTAL UNIT

An issue for discussion before the design was finalized concerned the size and location of experimental units. Specifically, should farms act as blocks and units be whole-fields, paired within the farm to be as alike in biodiversity as possible? Alternatively, should a single field be divided into two halves, again as alike as possible, defining a unit as a half-field? The arguments in favour of the alternative approaches involved scientific, statistical and practical issues.

A strong argument for the half-field design was the potential for reduction in variability. The two halves of a field are much more likely to be similar, in previous management, soil type and surrounding habitat, than

two different fields. Residual variation is reduced by choosing blocks such that experimental units within them are matched, as far as practicable, for the measured variable (Perry 1997). Under this argument, halving fields should enhance the statistical power to detect differences between treatments, and increase the precision with which they are estimated.

However, ecological relationships measured at one spatial scale may not have the same parameters or pertain at all at other scales (Heads & Lawton 1983; Norowi *et al*. 2000). Caution is required in extrapolating the results of a study on half-fields to a larger whole-field scale. Duffield & Aebischer (1994), Perry (1997) and Kennedy *et al*. (2001) have shown how the use of relatively small plots close to one another has affected the interpretation of experiments for relatively mobile species such as carabid beetles; this argument could favour the use of paired whole-fields. Indeed, birds and small mammals were excluded from comparison within the main FSE precisely because their territories and foraging areas often extend beyond half- or whole-fields (Firbank *et al*. 2003). More generally, tritrophic interactions between the chemical ecology of plants, herbivores and their natural enemies are subtle (Vet 1999) and Schuler *et al*. (1999) highlighted many potential indirect effects of GM plants on arthropod natural enemies.

Movement of individuals between the two halves of the same field might bias the estimated difference between treatments, especially if movement was related to the effect of crop management. For example, increased mortality on one half of the field could be compensated by density-dependent immigration from the other half. An individual carabid may easily travel the order of 300 m, the breadth of a square 10-ha field, in two nights (Kennedy 1994). Duffield & Aebischer (1994) noted that the recovery from pesticide application of invertebrate populations would proceed at a slower rate when entire fields were treated, compared with within-field plots of an identical size. Despite limited replication in the largest of their plots, they suggested that small-scale within-field trials to evaluate pesticides would in many cases fail to predict accurately the impact of commercial pesticide management.

Despite these caveats, useful information may still be obtained from half-fields for highly mobile species, such as bees and butterflies, as long as direct inferences concerning abundance are not made from counts. Instead, treatment differences relate to foraging preferences towards flowering plants. These problems of interpreting data concerning bees, butterflies and, to a lesser extent, some carabids must be seen in the context of the ecology of the taxa studied, relative to the treatments imposed. Direct effects of herbicide management regimes are most likely to impinge on vegetation; effects on invertebrates will probably be indirect.

Care must be taken to avoid interference between experimental units that are close together, for example from spray drift. Here, the separation distances, of 50 m for rape and maize and 6 m for beet, between half-field units help to minimize problems. Any chosen design would have to attempt to match field-margin biodiversity between experimental units. Such margins are important habitat in arable ecosystems as reservoirs for plants and overwintering sites for insects, cover and food for birds, and may affect invertebrate distributions (Lewis 1967).

The FSE aims to compare GMHT and conventional varieties of each of the four crops grown in realistic commercial conditions, which might favour the use of whole-fields. Against this was the practical issue that in the pilot year there was a lack of candidate fields, vital to choose pairs sufficiently well-matched for previous management and cropping history; this strongly favoured the use of half-fields. Also, half-fields reduce greatly the sampling effort, as recorders travel less to collect data. Accuracy might be improved if there is less time pressure; experience during the pilot year revealed this as an important consideration at particular times of the year when sampling overlapped between taxa.

Unfortunately, very few data exist on the relative variability between whole-fields within farms and that between half-fields within whole-fields. Surveys have been used to assess the environmental effects of intensive agriculture within the UK for decades (Potts & Vickerman 1974) but designed experiments are relatively recent and lack adequate replication of realistic-sized units (Sotherton, Jepson & Pullen 1988; Aebischer 1990; Perry 1997; Moller & Raffaelli 1998; Raffaelli & Moller 2000). Lennon (1998) listed nine recent European projects on integrated pest management and noted that each suffered difficulties with inference that resulted from either inadequate replication or complications due to crop rotations. Unfortunately, the crops studied in the well-designed MAFF LINK Integrated Farming Systems Study (Ogilvy *et al*. 1995) were largely different to those of the FSE. However, some data from the Game Conservancy and Allerton Research and Educational Trusts (Boatman & Brockless 1998), from up to five winter oilseed rape fields on the demonstration farm at Loddington, Leicestershire, UK, provided information on components of variation (Perry 1989) within eight abundant suction-sampled invertebrate groups. Some fields were halved, yielding information from 1994 to 1996 on between- and within-field variation, that could be used to compare the likely efficiency of half-field and paired-field designs. The variability of paired-fields was often similar to that for half-fields, but sometimes, especially during 1995, was much greater (Fig. 1). It was not possible, due to constraints of proper randomization and insufficient replication, to use data from the FSE pilot year (1999) to inform the choice of design, although an informal inspection suggested that half-fields were inherently less variable than paired-fields.

The final choice of a half-field design was based on the availability of fields, the associated difficulty of obtaining suitably matched paired fields, the probable major effect of herbicide being on weeds rather than invertebrates, the need to reduce variability and efficiency gains in sampling effort. The choice was made

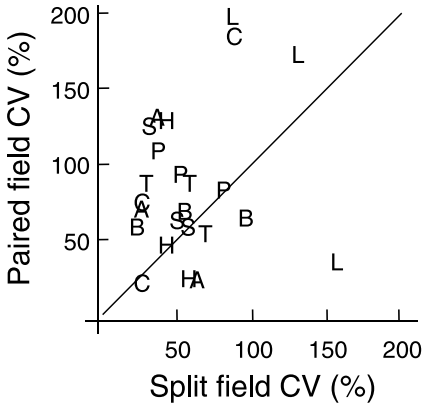

Fig. 1. Comparison of estimated coefficients of variation (CV) between half- and paired-fields from 1994 to 1996 from Loddington Farm. The data from the Allerton Project (Boatman & Brockless 1998), run by the Game Conservancy Trust for the Allerton Research and Educational Trust, were supplied by Dr Nicholas Aebischer (Game Conservancy Trust). Symbols represent annual values for the eight most abundant invertebrate groups in suction samples: Collembola (C), aphids (A), Homoptera (H), Thysanoptera (T), parasitoids (P), staphylinid larvae (S), Coleoptera adults (C) and Coleoptera larvae (L).

with the proviso that half-fields should fall within the range of field sizes used commonly for each crop, and should not compromise realistic growing conditions.

FARM AND FIELD SELECTION: REPRESENTATIVENESS AND RANGE

An important requirement of the FSE is that its results should apply to the British agricultural ecosystem and landscape as a whole. This raises the question of the representativeness of the farms included and the issue of farm and field selection. For example, it would be unsatisfactory if there were no fields within the FSE growing spring oilseed rape in Scotland, where a large acreage of the crop is grown. For the pilot study, fields came from a limited self-selected set of growers who were willing to grow GMHT crops. Within the FSE proper, the issue of representativeness was addressed by attempting to select fields that encompassed the full range of variation, in various variables, likely to be found in commercial practice. The current status within Britain for each crop was summarized with regard to its geographical distribution, usual agronomy, soil types and field sizes. This profile was then compared with more detailed information on specific candidate farms and fields, obtained from a questionnaire issued by the consortium to each grower who expressed interest in taking part (Firbank *et al*. 2003). Estimates were made of the intensiveness of the grower's inputs and the extent to which the farmers managed their land in ways that might favour biodiversity. Potential growers required early notification of whether their farm was selected, so a sequential approach was used to monitor the structure of the sample in terms of geographical spread, intensiveness and biodiversity, and to identify underrepresented strata.

The approach in the FSE was not to sample farms in proportion to their frequency of occurrence according to some factor. For example, low-intensity farms are relatively rare but they may contribute proportionately more to biodiversity than intensively managed farms (Watkinson *et al*. 2000). The consortium sought to include a disproportionately large sample of such low-intensity farms. Analyses will seek to identify a possible interaction between the treatment effect and intensity, for which there are ample degrees of freedom available.

Note that the randomization to half-fields within each field is distinct from the ability to scale-up from the experiment to some wider population, which requires that the experimental units within fields, and the fields themselves, must be representative. The larger the pool of farms, the more likely it was that a suitable set of farms could be selected. However, there is no requirement *per se* for such selection, either to ensure validity of the statistical test or for the ability to scale up.

Rather than the statistical tests of the null hypothesis, other approaches are to extrapolate the results of the FSE through explanatory, mechanistic modelling (Firbank & Forcella 2000; Watkinson *et al*. 2000) or multivariate community-based analysis; such work is not considered here.

CHOICE OF BOUNDARY TO HALVE FIELD AND TREATMENT RANDOMIZATION

Randomization of allocation of the GMHT and conventional varieties to the two halves of the field safeguarded against selection bias, for example GMHT crops being applied to the weedier half of the field. It also provided statistical validity for the test of the null hypothesis, and for the estimates of the precision of the magnitude of any differences. and it allowed differences detected to be ascribed causally as treatment effects.

The randomization protocol for the trial required a structured dialogue between the recorder from the consortium and the grower, so that the choice of boundary line to halve the field for sowing was made on scientific grounds not agronomic convenience. The optimum choice of boundary should result in two half-field units as alike as possible over the range of factors that contribute to the variability of wildlife within the field. The protocol also guarded against any preference a grower had for what side of the field should receive the GMHT treatment. Thus, treatment allocation was predetermined by project statisticians who assigned one treatment at random to the label 'A' and the other to label 'B'. This allocation was provided to the recorder (but unknown to him or her) in a sealed envelope. After the boundary line was agreed between recorder and grower, the half-field unit towards the north (for an east–west boundary line) or towards the west (for a north–south boundary line) was labelled as 'A', and the other as 'B', and drawn on a rough map. The envelope was then

opened and the treatments noted on this map. With this auditable procedure none of the recorder, statistician or grower could influence the randomization.

IMPOSITION OF CROP MANAGEMENT BY GROWERS

In some respects, the FSE experimental design has much in common with on-farm trials carried out by farmers, on their own land, in studies on third-world agriculture (Buzzard 2000). The control crop variety was selected by the farmer according to local conditions, and varied between farms. Both GMHT and conventional systems were managed by growers as closely as possible according to their current commercial practice, although within this constraint management practices were kept as similar as possible. Any pesticide seed treatment was the same on both treatments at a farm. Where non-herbicide treatments were imposed on both GMHT and the conventional varieties, they were applied at the same time unless there was good agronomic reason, for example if there were more pests on one half-field than the other. Growers took usual decisions for weed control on the conventional variety; this might or might not involve the use of consultant agronomists. However, usual practice remains difficult to define for GMHT varieties, because none has yet been grown commercially within Britain. Procedures that ensured that the treatment applied within each management regime was applicable are outlined by Firbank *et al*. (2003). Such considerations are vital to enable valid inference, and are equally important as biometrical issues of design and analysis; no treatment randomization can allow for biases arising from inappropriate management of the GMHT variety. Note that it was possible for there to be no herbicide applications to either half-field unit if, for example, there were no weeds to treat.

Some agronomic practices, such as the increased use of direct drilling or changes to normal rotations, might become associated with GMHT technology if it were commercialized. The FSE cannot, at this early stage in the use of GMHT, evaluate efficiently such events within an experimental framework of imposed treatments. However, the FSE will provide data to parameterize predictive models in which such scenarios may be studied.

VARIATES FOR ANALYSIS

Details of the range of farmland wildlife taxa studied in the FSE are given by Firbank *et al*. (2003). Both density and biomass of plants were recorded, as well as the seed bank and seed return. Invertebrates counted included carabid beetles in pitfall traps; butterflies and bees sighted along transects; other arthropods, such as Collembola and Heteroptera, in Vortis suction samples; crop pests counted on plants; and gastropods in refuge traps and verge searches. Although samples were usually identified to species or family level, analyses focused initially on totals over all species, functional groups and over large groupings such as total monocotyledons. Several samples may have been taken for each taxonomic group through the year; these may have been aggregated to give a single annual total or analysed separately. Whilst this present study is not intended to provide an exhaustive list of potential analyses, it is likely that measures of species richness and diversity will also be compared between treatments.

Statistical power

THE NEED FOR STATISTICAL POWER ANALYSIS

The choice of the number of fields in the FSE was controversial because it represented the first occasion on which GMHT crops were sown on this large scale in Europe and it was deemed inappropriate to grow a large area. Also, the cost of any publicly funded experiment must be constrained within limits and the more fields sown and sampled, the greater the cost. Against this, sufficient replication was required to detect effects. The statistical power that comes from proper control of variation and adequate replication (Perry 1986) is important in regulatory trials, which seek to study whether there are any deleterious environmental effects of new products (Anonymous & Perry 1999). A statistical power analysis, which quantifies the likely efficiency of an experiment, was essential for the FSE.

The statistical power of a significance test is the probability of rejecting the null hypothesis when some given alternative hypothesis is true. The power measures the chance of detecting an effect of a known magnitude using the specified experimental design, and varies according to the magnitude of the effect specified. It is often difficult for biologists to specify this quantitatively but without an answer to the question 'Precisely what degree of treatment effect do you consider important?' any power analysis is uninformative. Power depends also on sample size, the degree of random variation between experimental units and the chosen significance level of the test (Sokal & Rohlf 1981). Power is a continuum that varies non-linearly and gradually with sample size. There is no threshold level of replication below which an experiment is too poorly resourced to be worth conducting and above which it is satisfactory.

Power was estimated for the FSE over scenarios that encompassed a range of treatment differences, numbers of fields and degrees of random variability. For data that approximately follow a normal distribution, the power of standard tests, such as Student's *t*-test, can be calculated routinely. However, more complex calculations are required when, as here, ecological count data are collected that have an asymmetric frequency distribution and vary in relation to mean abundance. Power was estimated both for a standard, simple,

model based on a logarithmic transformation of counts, and also for an extended model developed to be more realistic for the form of ecological count data collected, with a large proportion of zero counts possible for some species.

MODEL I: THE SIMPLE, LOG-NORMAL MODEL

Suppose there were $j = 1, \ldots, n$ fields, with two experimental half-field units per field. The treatment factor, GMHT vs. conventional, had two levels, denoted $i = 1,2$. In the simple model, the observed response variate, the count c_{ij}, was transformed to $l_{ij} = \ln(c_{ij} + 1)$. Then, a standard randomized block ANOVA was done, with fields as blocks, on the transformed values, l_{ij}; the treatment effect was assessed with a t-test. This approach assumed a normal distribution for l_{ij} and therefore an approximately log-normal distribution for c_{ij}; this was termed the log-normal model (Table 1).

MODEL II: THE EXTENDED, NEGATIVE BINOMIAL MODEL

The extended model was designed to allow for many small or zero values of c_{ij} and for the observed dependence of variance, V, upon mean abundance, μ, for ecological count data, often expressed through a power-law (Taylor 1961), with parameters α and β:

$$V = \alpha\mu^{\beta} \qquad \text{eqn 1}$$

Model I above explicitly assumed that variance is homogeneous after transformation, and therefore implicitly assumed that the exponent, β, was close to 2·0. This followed from the result based on first-order Taylor series approximation (Cochran 1938), that the variance on a natural log scale is approximately equal to the variance on the untransformed scale divided by the square of the untransformed mean (i.e. $\text{var}[\ln(c_{ij})] \approx V/\mu^2$).

The systematic effects in the extended model explicitly allowed for variability in the blocking structure. The effects represented by the parameters of the extended model were: an overall mean, $\mu = e^{\gamma}$, say, a field effect F_j and a treatment effect t_i. These combined to give the expected value of the response variable, c_{ij}, on natural logarithmic scales:

$$\log_e E[c_{ij}] = \gamma + F_i + t_i = \theta_{ij}, \text{ say} \qquad \text{eqn 2}$$

Treatment and field effects were assumed multiplicative on the natural count scale.

The random component of the extended model reflected variability from unit to unit, i.e. between half-fields and within fields. A negative binomial distribution was assumed for the counts, c_{ij}, but the shape parameter of the distribution, k_{ij}, for each treatment on each field was constrained (Perry *et al*. 1998) to follow equation 1. The mean of this negative binomial distribution, for each treatment on each field, was denoted ϕ_{ij}, where $\phi_{ij} = \exp(\theta_{ij})$. For each particular submodel the value of $k_{ij}(\phi_i)$ for each count was determined from:

$$k_{ij}(\phi_{ij}) = \phi_{ij}/(\alpha\phi_{ij}^{\beta-1} - 1), \text{ for } k_{ij}(\phi_{ij}) > 0, \text{ and } k_{ij}(\phi_{ij}) = \infty, \text{ otherwise} \qquad \text{eqn 3}$$

Three submodels were studied, with a different single value of β for each submodel: $\beta = 1$, 1·5 and 2, respectively (Table 1). These values of β were chosen to incorporate relationships typical of those observed for ecological count data. When $\beta = 1$, variance was proportional to the mean. For efficient analysis, a generalized linear model (GLM) with a logarithmic link would be assumed typically, with Poisson errors and estimated scale parameter (McCullagh & Nelder 1989); this was termed the log-linear model. When $\beta = 2$, the coefficient of variation was theoretically constant, and the simple model with a logarithmic transformation provided an efficient analysis. Values close to $\beta = 1{\cdot}5$ lacked such mathematically tractable interpretation, but were typical of exponents encountered for many species in field data (Taylor, Woiwod & Perry 1978).

To generate the counts, given specified values of μ, the field effect F_i and the treatment effect t_i (see below), the value of ϕ_{ij} was found from equations 1 and 2. Then, given values of α and β for a particular submodel, the value of k_{ij} was found from equation 3. This equation

Table 1. Summary of statistical models used in the study. Models differ in their assumed variance-mean relationship, measured through the power-law function, $V \propto \mu^{\beta}$. Parametric methods use F-tests. ANOVA denotes analysis of variance. GLM denotes generalized linear model; for $\beta = 1$ with Poisson errors and log-link, for $\beta = 1.5$ with user-defined error and log-link. Entry in column detailing the non-parametric randomization test is the test-statistic used: d is mean difference in logarithmically-transformed count; r is logarithmically-transformed ratio of arithmetic means of counts; d_w is a weighted version of d. For further details see text

	Data analysis		Power estimate	
β	Parametric	Non-parametric randomization test	Parametric	Non-parametric randomization test
2	Log-normal, ANOVA	d	Log-normal	d
1·5	GLM	d_w	–	d_w
1	Log-linear, GLM	r	–	r

treated the rare case of a negative simulated value of k as indicating effectively Poisson variation, for which case the value of k was set to infinity and the Poisson distribution replaced the negative binomial for simulation. A negative binomial variate with mean μ and variance $\mu + \mu^2/k$ was simulated using the following two-step procedure (Morgan 1984). First, a random variate, say g, was sampled from a gamma distribution with mean μ and variance μ^2/k; secondly, a count from a Poisson distribution with mean g was sampled. Random gamma and Poisson variates were generated using subroutines from the NAG library (Numerical Algorithms Group 1997).

MODEL PARAMETER VALUES USED IN STATISTICAL POWER ANALYSIS

The model was used to generate sets of count data for specified combinations of parameter values and different magnitudes of the treatment effects (Table 2).

Mean abundance, μ, was studied for four values: 1, 5, 10, 50; the first three values were chosen because, for these simulations, attention was focused largely on the case of smaller counts. The field effect, F_i, modelled the effect of the variability between fields and contributed to the variability of counts, at larger than unit scales, through equation 2. Field effects were simulated as fixed effects, such that $F_i = -M + (j-1)q$, where $j = 1, \ldots, n$ and $q = 2M/(n-1)$, with $M = \log_e 10 = 2{\cdot}303$, and n specified the number of fields. For example, for $n = 20$, this resulted in the series: $F_i = -2{\cdot}303, -2{\cdot}182, \ldots, -0{\cdot}364, -0{\cdot}121, 0{\cdot}121, 0{\cdot}364, \ldots, 2{\cdot}182, 2{\cdot}303$. This scheme ensured two orders of magnitude variation in the blocking factor representing the field effect, so the expected values of the response variable for the two extreme fields varied by 100-fold. The variation in mean abundance, above, when combined with the field effect, gave expected ranges of simulated abundance of 0·10–10, 0·50–50, 1–100 and 5–500; and means of 2, 12, 23 and 115.

The degree of random variation between the experimental half-field units was varied through the parameter α. Values of α were chosen to achieve coefficients of variation (CV) of 50%, 80% and 100%. The coefficient of variation, $\sqrt{V}/\mu$, here equated to $\alpha^{1/2}\mu^{(\beta/2-1)}$. It provided a useful way of specifying baseline variability that permitted direct comparison with characteristic values for a particular taxon, perhaps calculated from previous experiments. However, the theoretical values listed above, of 50%, 80% and 100%, were different from those values actually realized by the simulations, which were subject to random variation.

Three different values of treatment effect, t_i, were specified, representing multiplicative differences of × 1·3-fold, × 1·5-fold and × 2-fold. If treatment effects were denoted as $t_1 = 0{\cdot}5 \ln(R)$ and $t_2 = -0{\cdot}5 \ln(R)$ on the natural scale, then this corresponded to a multiplicative difference of R, expressed either as a $(R-1)$% increase or a $(1 - R^{-1})$% decrease, of one treatment relative to the other. For example, values of $t_i = \pm 0{\cdot}203$ were used to represent a multiplicative difference of $R = 1{\cdot}5$ between the treatments. Then, for a mean count on the logarithmic scale of 0·0, the expected value, back-transformed onto the natural count scale under treatment 1, would be 1·225. For treatment 2 the value would be 0·816. This yielded a multiplicative difference of 1·5-fold between the two treatments; it may be viewed either as a 50% increase or as a 33·33% decrease of one relative to the other.

Table 2. Summary of runs used in statistical power simulations. In each case the variance, V, of the count is related to the mean, μ, through $V = \alpha\mu^{\beta}$. Means and coefficients of variation (CV) are expected values, which take no account of any additional variability induced by field and treatment effects

Reference number of run	α	β	μ	CV%
1	1	1·0	1	100
2	1	1·5	1	100
3	1	2·0	1	100
4	5	1·0	5	100
5	1	2·0	10	100
6	7·07	1·5	50	100
7	1·79	1·5	5	80
8	6·4	1·0	10	80
9	0·64	2·0	50	80
10	0·25	2·0	5	50
11	0·79	1·5	10	50
12	12·5	1·0	50	50

CALCULATION OF STATISTICAL POWER

The power of three statistics was computed, each based on Monte Carlo paired randomization tests (Manly 1994), and applied to each set of simulated data (Table 1). Briefly, this entailed recording the value of the 'observed' statistic computed for each generated set of $2n$ counts, and comparing this value against 199 other 'randomized' values of the statistic, recomputed after random relabelling of the treatment codes for each of the n pairs of counts. If the observed statistic exceeded the upper 5th centile of the ranked randomized values then the null hypothesis was rejected for that generated set. Two-tailed tests were performed by using the absolute value of the test-statistic. The process was then repeated for 500 sets and the power estimated as the proportion of rejections.

The three statistics studied reflected the three forms of variance–mean relationships, characterized by the exponent β. The first, d, closely related to the log-normal model, was the simple mean of the differences between the two treatments on the logarithmic scale, $d = \Sigma_i[l_{1i} - l_{2i}]/n$. This should have relatively high power when variance is proportional to the square of the mean ($\beta = 2$). The second statistic, r, closely related to the log-linear or Poisson regression model for count data (McCullagh & Nelder 1989), was the logarithm of

Table 3. Estimated type I errors (5% level) for Monte Carlo paired randomization tests using $n = 20$ pairs. Estimates based on 1000 sets of simulated data (SE = 0·7%). The three test statistics for testing treatment difference were d, r and d_w, where d denotes the mean difference in logarithmically transformed count; r denotes logarithmically transformed ratio of arithmetic mean counts; d_w is a weighted version of d. For further details see text

Reference number of run	β	μ	CV%	Type I error (%) test statistic		
				d	r	d_w
1	1·0	1	100	5·3	3·5	5·6
2	1·5	1	100	4·6	3·9	5·5
3	2·0	1	100	4·0	3·6	4·9
4	1·0	5	100	5·5	4·4	4·8
5	2·0	10	100	5·4	5·3	5·4
6	1·5	50	100	4·9	5·6	4·7
7	1·5	5	80	5·6	4·4	5·1
8	1·0	10	80	4·3	4·2	3·6
9	2·0	50	80	4·8	4·4	4·9
10	2·0	5	50	6·3	4·9	5·5
11	1·5	10	50	6·5	4·8	5·6
12	1·0	50	50	5·1	4·2	4·7
Mean				5·2	4·4	5·0

the ratio of the overall arithmetic means of the two treatments, $r = \ln[\Sigma_i c_{1j}/\Sigma_i c_{2j}]$. This should have relatively high power when variance is proportional to the mean (β = 1). The third statistic, d_w, derived by P.R., was introduced to try to accommodate the intermediate case (β = 1·5). It was a weighted version of d, with weights based on the approximate variance, assuming β = 1·5, of the difference in logarithmically transformed counts, i.e. $d_w = \Sigma_j w_j[l_{1j} - l_{2j}]/(\Sigma_j w_j)$, where $w_j = [(1 + c_{1j})^{-0·5} + (1 + c_{2j})^{-0·5}]^{-1}$.

Power was estimated for treatment differences of magnitude R = 1·3, 1·5 and 2, for 12 different combinations of the parameter values α, β and μ. It was decided to study power assuming equal sample sizes of 20 and 30 fields per year. Over a 2-year experiment these would give sample sizes of n = 40 and 60, and over 3 years n = 60 and 90. Hence, the range of sample sizes used, n = 20, 30, 40, 60 and 90, covered all combinations of 20 and 30 fields per year, for periods from 1 to 3 years. This power study simulated a total of more than 8·5 million negative binomial random variables, in a total of 90 000 sets of data. In addition, the type I error of each test was checked using 1000 sets of simulated data.

RESULTS OF STATISTICAL POWER ANALYSIS

Table 3 shows that in each case the type I error was close to its nominal value of 5%. Tables 4–6 show the estimated power for values of R = 1·3, 1·5 and 2, respectively, and, for comparison, the corresponding power of a paired t-test for the situation when l_{ij} was assumed to have a normal distribution, i.e. the simple log-normal model referred to earlier. The power of the t-test was

Table 4. Statistical power for detecting a × 1·3-fold difference for simulated count data. Details of run parameters and the three test-statistics are in Table 2. Estimates based on 500 sets of simulated data. The power for the log-normal model was based on a paired t-test. Values of power exceeding 80% shown in bold. See text for further details

Reference number of run	CV%	Test statistic	Number of pairs (n)				
			20	30	40	60	90
1	100	d	16	19	27	41	55
		r	15	22	35	48	69
		d_w	19	23	35	49	67
2	100	d	10	15	22	31	43
		r	10	12	19	27	40
		d_w	10	16	22	31	46
3	100	d	6	14	13	16	22
		r	5	10	11	15	16
		d_w	6	12	13	16	22
4	100	d	15	19	25	39	53
		r	20	25	35	53	70
		d_w	20	27	33	56	70
5	100	d	11	11	11	19	26
		r	10	11	10	16	21
		d_w	11	12	12	21	29
6	100	d	16	17	19	28	40
		r	13	19	17	30	42
		d_w	17	21	21	36	51
Log-normal	100		16	22	28	40	55
7	80	d	14	20	28	38	47
		r	14	20	26	36	44
		d_w	15	24	32	41	52
8	80	d	20	25	34	46	60
		r	30	39	53	74	**88**
		d_w	29	37	50	69	**86**
9	80	d	12	20	27	31	46
		r	9	13	17	19	28
		d_w	11	17	22	28	39
Log-normal	80		20	29	37	52	70
10	50	d	20	30	39	57	75
		r	19	25	30	43	61
		d_w	22	32	39	58	77
11	50	d	32	47	57	76	**88**
		r	31	50	59	**80**	**92**
		d_w	36	58	68	**88**	**94**
12	50	d	24	38	46	62	76
		r	60	**80**	**91**	**98**	**100**
		d_w	53	72	**87**	**97**	**100**
Log-normal	50		39	55	68	**85**	**96**

calculated using the statistical package Minitab Release 13 (Moultine & Bluman 2001). For a treatment effect of R = 1·5, n = 60 pairs and a CV = 50%, the power exceeded 90% in all but one case. When the treatment effect represented a doubling or halving of density and R = 2, for n = 60, for values of CV = 50%, 80% and 100% and for values of μ ≥ 5, the power exceeded 85% in all but one case.

As expected, the power of the r-statistic was higher than that of the d-statistic when the variance was proportional to the mean (β = 1) and vice-versa when the variance was proportional to the square of the mean (β = 2). The d_w-statistic performed best for β = 1·5, but also appeared agreeably robust, maintaining comparable power to d for β = 2 and to r for β = 1.

Table 5. Power for detecting a × 1·5-fold difference using simulated count data (see Table 4)

Reference number of run	CV%	Test statistic	Number of pairs (n)				
			20	30	40	60	90
1	100	d	32	42	53	72	**91**
		r	37	54	65	**85**	**97**
		d_w	40	53	65	**83**	**97**
2	100	d	25	31	42	55	72
		r	22	28	41	53	71
		d_w	26	34	46	59	77
3	100	d	13	22	22	34	47
		r	12	18	20	29	44
		d_w	14	23	24	35	49
4	100	d	29	46	50	75	**86**
		r	35	60	69	**88**	**97**
		d_w	37	61	70	**89**	**97**
5	100	d	16	21	27	39	58
		r	12	16	23	30	42
		d_w	16	22	28	38	55
6	100	d	21	38	43	56	77
		r	25	35	43	58	**81**
		d_w	30	42	50	66	**88**
Log-normal	100		31	45	56	75	**90**
7	80	d	28	43	51	68	**84**
		r	28	40	51	70	**86**
		d_w	33	46	59	77	**91**
8	80	d	38	53	65	77	**92**
		r	55	75	**90**	**97**	**100**
		d_w	54	74	**88**	**96**	**100**
9	80	d	25	39	46	63	**82**
		r	17	26	31	44	59
		d_w	24	35	38	59	78
Log-normal	80		41	58	71	**87**	**97**
10	50	d	44	63	77	**91**	**98**
		r	37	50	65	**80**	**93**
		d_w	47	64	77	**93**	**99**
11	50	d	57	79	**88**	**98**	**100**
		r	65	**83**	**92**	**99**	**100**
		d_w	73	**92**	**97**	**100**	**100**
12	50	d	50	71	**80**	**93**	**99**
		r	**91**	**99**	**100**	**100**	**100**
		d_w	**88**	**98**	**100**	**100**	**100**
Log-normal	50		73	**90**	**96**	**100**	**100**

Table 6. Power for detecting a × 2-fold difference using simulated count data (see Table 4)

Reference number of run	CV%	Test statistic	Number of pairs (n)				
			20	30	40	60	90
1	100	d	67	**87**	**94**	**100**	**100**
		r	75	**94**	**99**	**100**	**100**
		d_w	77	**94**	**98**	**100**	**100**
2	100	d	51	69	**82**	**96**	**100**
		r	49	66	**83**	**97**	**99**
		d_w	55	74	**86**	**98**	**100**
3	100	d	26	44	51	74	**89**
		r	21	38	46	65	**86**
		d_w	27	45	54	75	**90**
4	100	d	69	**86**	**95**	**98**	**100**
		r	77	**96**	**99**	**100**	**100**
		d_w	**82**	**97**	**99**	**100**	**100**
5	100	d	37	53	64	**85**	**95**
		r	31	42	49	71	**84**
		d_w	37	52	65	**85**	**95**
6	100	d	52	69	**84**	**95**	**100**
		r	54	73	**85**	**95**	**100**
		d_w	60	**82**	**90**	**99**	**100**
Log-normal	100		70	**88**	**95**	**99**	**100**
7	80	d	67	**83**	**92**	**99**	**100**
		r	63	**82**	**91**	**99**	**100**
		d_w	71	**90**	**96**	**99**	**100**
8	80	d	77	**93**	**97**	**100**	**100**
		r	**94**	**99**	**100**	**100**	**100**
		d_w	**92**	**99**	**100**	**100**	**100**
9	80	d	57	76	**88**	**97**	**100**
		r	40	58	69	**87**	**100**
		d_w	51	72	**85**	**97**	**100**
Log-normal	80		**84**	**96**	**99**	**100**	**100**
10	50	d	**91**	**96**	**99**	**100**	**100**
		r	**81**	**91**	**97**	**100**	**100**
		d_w	**89**	**98**	**100**	**100**	**100**
11	50	d	**97**	**100**	**100**	**100**	**100**
		r	**98**	**100**	**100**	**100**	**100**
		d_w	**100**	**100**	**100**	**100**	**100**
12	50	d	**88**	**98**	**100**	**100**	**100**
		r	**100**	**100**	**100**	**100**	**100**
		d_w	**100**	**100**	**100**	**100**	**100**
Log-normal	50		**99**	**100**	**100**	**100**	**100**

Because of its complexity, the power estimated from the negative binomial model varied more than that based on the log-normal model. Some of this variation could be accounted for by deviations of the actual realized CVs from the theoretical target values given in Tables 4–6.

To summarize this complex situation power was examined in relation to a standard statistical 'non-centrality parameter' (Pearson & Hartley 1976; section 14.5, tables 27 and 30), when a consistent pattern emerged (Fig. 2). The non-centrality parameter used, Δ, was the true difference divided by the standard error of the estimated difference, d; it therefore had much in common with the simple t-statistic, well-known in ecology. Specifically, Δ was calculated as $\log_e R(2\sigma^2/n)^{-1/2}$, where σ^2 was the variance of l_{ij}. For the negative binomial model, σ^2 was estimated from the simulated data by calculating the average residual mean square after fitting field and treatment effects in the ANOVA of l_{ij}. In addition, the solid line in Fig. 2 shows the power for the simple log-normal model, which, because it was plotted vs. Δ, was independent of the quantity R/σ and so may be utilized generally to estimate the power whenever the magnitude of effect is expressed as a multiple number of standard errors. The log-normal model therefore provided a useful baseline against which to assess the effect of assuming negative binomial counts, i.e. the effect of the discrete, variable and sometimes small counts encountered commonly in ecology. Interpretation was aided by noting that for the log-normal model the percentage coefficient of variation was equal to $100\sqrt{[\exp(\sigma^2) - 1]}$ (Hastings & Peacock 1975), so the four solid circles in Fig. 2 corresponded to CV of 50%, 80%, 100% and 156%. Clearly, most of the simulated powers fell slightly below the solid line, so the additional variability resulted, as expected, in a small

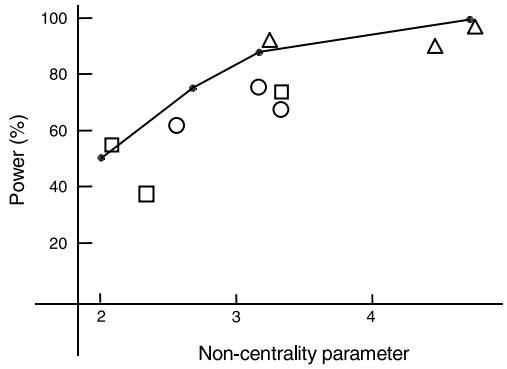

Fig. 2. Statistical power (%) for detecting an $R = \times 1{\cdot}5$-fold difference with a scheme of 20 fields over 3 years (i.e. $n = 60$ pairs), based on simulated data from a negative binomial model with theoretical, target CV of 50% (triangles), 80% (open circles) and 100% (squares), using the *d*-statistic (see Table 5), for values of $\mu \geq 5$. Solid lines show power for log-normal model; solid circles on these lines relate to CV, from left to right, of 156%, 100%, 80% and 50%. Non-centrality parameter, Δ, calculated as $\Delta = \log_e R(2\sigma^2/n)^{-1/2}$, where σ^2 is the variance of the logarithmically transformed count (see text).

reduction in power. This approach also provided a direct method of linking the theoretical power calculations to the analysis of actual data, via an estimate of σ^2 from a simple ANOVA of l_{ij}. For any desired value of R and projected value of n, we may use this future estimate of σ^2 to derive a value of Δ; an approximate prediction of power may then be made for the log-normal model, perhaps with some slight downward adjustment for the effect of extra variability. For the log-normal model the power for detecting a difference of three times the standard error was about 85%, and about 98% for a difference of four standard errors.

There are some cases, mainly for small μ and large CV, where the recommended minimum replication level of 20 fields per crop per year over 3 years ($n = 60$) had low power for $R = 1{\cdot}5$. Tables 4–6 suggested, for $\mu \geq 5$, that this reduction in power could sometimes be offset by using the *r*-statistic and the log-linear model for analysis. However, if the mean count was as small as $\mu = 1$, even in the ideal case of a completely random distribution of counts power would always be limited. For example, for the log-linear model the standard error of the estimate of log R, s, was approximated by $s \approx \{A[(\Sigma_j c_{1j})^{-1} + (\Sigma_j c_{2j})^{-1}]\}^{1/2}$, where A was an estimate of the overdispersion. Consider a scheme with $n = 60$ pairs; mean counts per half-field of $\mu_1 = 1$ and $\mu_2 = 0{\cdot}7$ for the two treatments, so $\ln(R) = 0{\cdot}36$; and with no overdispersion, so $A = 1$. Then $s = 0{\cdot}20$, and the difference of 0·36 was a mere 1·8 times the standard error. Hence it was not surprising that the power for detecting an effect would be small. This emphasizes the importance of having adopted protocols that yielded sufficiently large means per half-field unit, and of focusing on the analysis of the abundance of common individual species or groups. Rare species are important too, but their contribution will be mainly analysed through indices of diversity.

CONCLUSIONS OF STATISTICAL POWER ANALYSIS

The consortium sowed 272 fields over four crops, an average of $n = 68$ per crop. The power analysis indicated that replication of 20 fields per crop per year over 3 years ($n = 60$) should have provided adequate power (> 80%) to detect multiplicative differences of $R = 1{\cdot}5$-fold, so long as CV did not exceed 50% and mean abundance exceeded 5·0. There was no need for strictly equal replication of 20 fields per crop per year, as it is the total replication that is important. Estimates of CV from sets of data made prior to the start of the FSE were, in most cases, close to the figure of 50% used in the power analysis. Mean CV in half-fields from the Allerton project (Boatman & Brockless 1998) for different taxonomic groups were (Fig. 1): Collembola (38%), aphids (45%), Homoptera adults (46%), Thysanoptera (55%), Parasitica (59%), staphylinid larvae (47%), Coleoptera (65%) and Coleoptera larvae (128%). Frampton (1999) reported other suction sample data in winter wheat for which the total Collembola count had a CV of 51% for variation between different fields.

The weighted test-statistic d_w, may provide a promising basis for future analyses of data with a variance-mean exponent, β, between 1 and 2. The results reinforced the importance of reducing the CV between experimental units, and of the limitations in analyses of variates with small means. The need to reduce the CV supported the choice of half-field over paired-fields in the FSE design.

Analysis

THE BASIC ANALYSIS

In the FSE, the crops are considered and will be analysed, at least initially, separately. Here we illustrate a mode of analysis that follows the simple and extended models and their associated test-statistics. The rationale is to provide a range of analyses that (i) address the null hypothesis and allow estimation of treatment effects; (ii) are appropriate to the data; (iii) match the simplicity of the design; (iv) provide results that are transparent and easily understood; (v) allow for heterogeneity and other deviations from model assumptions; and (vi) have the flexibility to allow for the inclusion of covariates and multifaceted extensions to the basic analysis. Currently, a two-stage process is envisaged. The first relates to a basic analysis, which conforms to criteria (i–v) above, and will be almost identical for all variates and crops. Extensions to allow for criterion (vi) will build on this basic analysis. What follows is presented to exemplify the sort of analysis that might be performed on 3 years of data from the

FSE. The basic analysis will consist of three components, each with a parametric and non-parametric form (Table 1). In the first of the parametric analyses each count is assumed to vary proportionally to the square of its mean. This leads naturally to a logarithmic transformation for efficient analysis, and hence to the log-normal model. An analysis of variance is used to provide a test of the null hypothesis, with the fields as blocks; the magnitude of d is estimated. In the second, the variability of counts is assumed to vary proportionally to the mean. The log-linear GLM described above is used, with an analogous analysis of deviance; the magnitude of r is estimated. Both these models are used ubiquitously in the ecological literature to analyse plant and insect counts; both provide an F-statistic to test the null hypothesis. An extension of this GLM, to the case where the variability of counts is assumed to vary as a power of 1·5 of the mean, is also fitted.

The models underlying the parametric approach might lack sufficient robustness to deviations from the assumed distributions. The distributional assumptions implicit in the different variance–mean relationships adopted by the models will be tested informally, using a suite of diagnostic plots of residuals (Carroll & Ruppert 1988). In addition, formal Monte Carlo randomization tests will also be done by random relabelling of the treatment codes for each of the observed pairs of counts. Such randomization tests are useful when there are many small and/or zero counts in a data set, for which parametric tests might be too liberal. The test-statistics are, respectively, d, r and d_w (Table 1). The randomization tests use 999 random permutations within each run to estimate P-values.

SUBSIDIARY COMPARISONS AND INTERACTIONS WITH OTHER FACTORS

Here an illustration is given of an extension to the basic analysis to answer additional questions, subsidiary to the main null hypothesis but adding to the interpretative power of the FSE. One is the possible interaction between treatment and years. Another is the requirement to confirm with experimental data that sugar and fodder beet can, as claimed, be treated as effectively the same crop for the FSE; this involves testing the interaction between treatment and crop type. With so many two-factor interactions that require testing, it might be prudent to test some higher-order interactions as well, and the FSE design provides plenty of degrees of freedom. For example, consider the analysis of a variable for 60 beet fields, with, for convenience, levels equally apportioned over three factors representing intensity, years and crop type (i.e. sugar and fodder beet), with two, three and two levels, respectively. There would therefore be five fields representing each combination of levels of intensity, years and crop type, with the main treatment factor occurring at both levels on each field. In a skeleton analysis of variance (Table 7), with all possible interactions fitted up to the full four-factor interaction, all F-tests of interest are computed with 48 residual degrees of freedom. There would, of course, be sufficient flexibility to fit other covariates of interest.

Table 7. Skeleton analysis of variance for 60 beet fields over 3 years, with levels equally apportioned over three blocking factors representing intensity (I), years (Y) and crop type (sugar and fodder beet, C), with two, three and two levels, respectively. All the main effects and interactions measured by these blocking factors are estimated in the fields stratum. The main treatment factor, comparing GMHT vs. conventional, is represented by T. The main effect of T, and all two-, three- and four-factor interactions involving T are estimated in the half-field units stratum. All F-tests are based on 48 residual degrees of freedom

Source of variation	d.f.	SS	MS	F
Fields stratum				
I	1			
Y	2			
C	1			
All interactions between I, Y and C	7			
Residual	48	RSS_s	RMS_s	
Total	59			
Half-field units stratum				
Total fields (blocks, from above)	59			
Main effect of T	1			F_T
T × I interaction	1			F_{TI}
T × Y interaction	2			F_{TY}
T × C interaction	1			F_{TC}
T × I × Y interaction	2			F_{TIY}
T × I × C interaction	1			F_{TIC}
T × Y × C interaction	2			F_{TYC}
T × I × Y × C interaction	2			F_{TIYC}
Residual	48	RSS_u	RMS_u	
Total	119			

Examples of analyses using FSE data from year 2000

After the first year of the FSE, data were available to estimate CV for a range of taxa from each protocol, and to reassess the statistical power calculations. Some example analyses follow.

WITHIN- AND BETWEEN-HALF-FIELD VARIABILITY FOR WEED SEEDS AND WEED SEEDLINGS

The first data considered were weed seeds and weed seedlings on fields where no pre-emergence herbicide was used. This ensured that for all these data, at the time of sampling, each half-field had exactly the same operations; there was therefore no reason to expect any treatment effect. For the seeds, samples were taken from $n = 4$, and for the seedlings from $n = 12$ transects per half-field (Firbank *et al*. 2003). This analysis took a components of variance approach (Perry 1989) to distinguish the component governed by sampling error, arising from variation (V_t) between transects within

Table 8. Hierarchical nested analysis of variance of total weed seedling and total weed seed counts for each transect, pooled over all crops. Data pooled over all four crops sown in 2000. Counts, c, were transformed to $\log_e(c + 1)$ prior to analysis. Variance component V_t was estimated directly as the between-transect, within-half-field MS. Variance component V_h was estimated from: V_h = (between-half-field MS – within-half-field MS)/n, with n = 12 for seedlings and n = 4 for seeds. Overall variance of the total count per half-field was estimated as $V_o = V_h + V_t/n$, and approximate CV% as $100\sqrt{V_o}$

Source of variation	d.f.	SS	MS	F	(P)	Estimated variance component (% of total)
Weed seedlings						
Between-fields, within-crops	17	687·3	40·43	23·0	(< 0·001)	
Between-half-fields, within-fields	20	35·18	1·76	3·26	(< 0·001)	V_h = 0·102 (16)
Between-transects, within-half-fields	440	235·8	0·536			V_t = 0·536 (84)
Total	447	958·3				V_o = 0·147, approx. CV = 38%
Weed seeds						
Between-fields, within-crops	68	366·9	5·40	8·10	(< 0·001)	
Between-half-fields, within-fields	72	47·97	0·666	1·84	(< 0·001)	V_h = 0·075 (17)
Between-transects, within-fields	432	157·6	0·365			V_t = 0·365 (83)
Total	572	572·4				V_o = 0·166, approx. CV = 41%

half-fields, from the variation between half-fields within fields (V_h). We then studied whether within-half-field sampling intensities were sufficiently large to reduce the overall variance (V_o, where $V_o = V_h + V_t/n$) to an acceptable level, when expressed as a CV. Data were pooled over all four crops sown in 2000 (Table 8). If, say, only n = 1 transect per field been sampled, then the predicted CV would have been approximately 80% and 66% for seedlings and seeds, respectively. This justified the multiple transects used, which reduced the estimated CV to 38% and 41%, respectively.

WEED SEED ABUNDANCE FOR BEET

The next analysis was of total weed seed abundance for the 24 fields of the beet crop sown in 2000. Samples were taken after halving the field, but before sowing, so no treatment had been applied and no difference was expected between the treatments, here denoted as 1 and 2. The geometric mean abundance was 93 seeds, and the estimated CV was 52%. Such abundance and variability would be entirely satisfactory according to the power analysis. A scatterplot (Fig. 3) showed, as expected, no gross differences between the treatments. The range of values for both treatments exceeded 1·5 orders of magnitude, comparable with the two orders of magnitude assumed for the power analyses.

The estimate of R, the multiplicative factor by which one treatment was greater than the other was, for the log-normal model, ×1·33 (approximate SE = 0·199; P = 0·068; Table 9). By contrast, the estimate for the log-linear model was ×1·25 (approximate SE = 0·183; P = 0·139). The difference in these estimates of R emphasized the importance of discrimination between the two models with diagnostic residual plots. In fact, the log-linear model demonstrated a clear increase in the variability of the residuals as the fitted values increased (Fig. 4a), indicating the assumed variance–mean relationship was wrong and the need for a more skewed distribution. By contrast, the diagnostic plots implied a preference

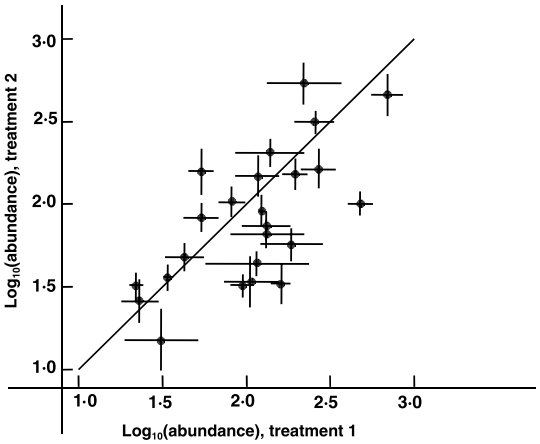

Fig. 3. Total weed seed abundance from the beet crop seed bank, for year 2000, plotted for treatment 2 vs. treatment 1, on a logarithmic (base 10) scale. Equality line shown for guidance. Horizontal and vertical lines around points show approximate standard errors for each estimated total.

Table 9. Analysis of variance and tables of means for weed seed abundance in the seed bank for the 24 beet fields in year 2000. The main treatment factor, comparing GMHT vs. conventional, is represented by the factor T. The F-test is based on 23 residual degrees of freedom. The column headed P gives the F-probability. The SED is the standard error of the difference between two means

Source of variation	d.f.	SS	MS	F	(P)
Fields	23	6·093	0·265	5·23	
T	1	0·186	0·186	3·67	(0·068)
Residual	23	1·165	0·051		
Total	47	7·443			

Tables of means (logarithmic scale, base 10)

Mean	Value		Replication	SED
Grand mean	1·972		24	
Treatments	1	2		
	2·034	1·910	12	0·0650

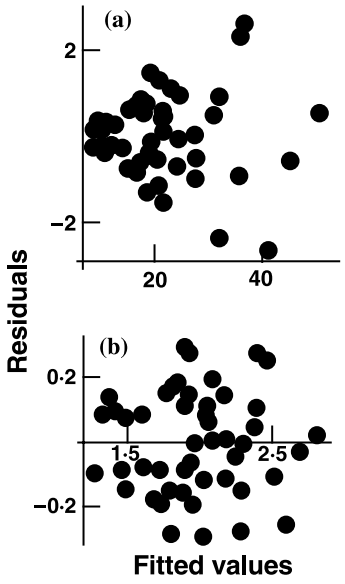

Fig. 4. (a) Standardized residuals from log-linear model analysis of total weed seed abundance from the beet crop seed bank, for year 2000, plotted against fitted values on natural scale. The lack of trend in the graph indicates there is no systematic lack of fit, but variability clearly increases with fitted values, so the value of unity for the exponent in the assumed variance–mean relationship is too small. (b) Residuals plotted against fitted values, both on a logarithmic scale, for the same data analysed using the log-normal model. There is no trend, and the logarithmic transformation has equalized the residuals, supporting the assumed value of 2 for the exponent in the variance–mean relationship.

for the log-normal model for which residuals, plotted against fitted values, appeared homogeneous (Fig. 4b). This was also indicated by the comparisons between the parametric and non-parametric analyses for the log-normal and log-linear models. The above probability of 0·068 for the log-normal model was close to the value of 0·073 estimated by the randomization test for the *d*-statistic. However, this was not the case for the log-linear model, for which the randomization probability was estimated as 0·248. This discrepancy corroborates the implication that the log-linear model should be viewed with caution for these data. Indeed, in similar analyses, the randomization tests were often more conservative than the equivalent parametric tests, particularly when there were many zero values, as for rarer species. Although the *P*-value of 0·068 was fairly close to the usual critical value of 0·05, any difference between treatments was difficult to explain on ecological grounds because the samples were taken before any treatments had been applied. The multiplicative difference of ×1·33 was probably a chance effect; data from other years should help to clarify the interpretation.

WEED SEED ABUNDANCE FOR MAIZE AND SPRING OIL SEED RAPE

Analyses of total weed seed abundance from the seed bank of the two other spring-sown crops during 2000 provided similar results. For maize, the geometric mean density over the 13 fields was 122, and the estimated CV was 42%. For spring oil seed rape there were two outlying large values that both exceeded 400, while the geometric mean density over the 14 fields was 89. The estimated CV was 37%. Again, for both maize and spring oil seed rape the diagnostic plots showed that the log-normal model appeared appropriate while the log-linear model displayed variance heterogeneity.

NUMBER OF WEED SEED SPECIES FOR BEET

All analyses considered thus far have used abundance as the response variable, but other responses are possible. The number of species, *S*, measured in the seed bank for the beet crop data above, provided an alternative quantitative comparison of biodiversity, as long as abundance was similar between treatments, as here. *S* ranged from six to 30 species over the fields, averaging 14·75 for each treatment. The estimated CV was 16%. Diagnostic plots showed both models appeared appropriate. The null hypothesis requires both abundance and diversity measures to be addressed by the FSE.

Discussion

The ambitious scope of the FSE has created numerous challenges, many of them concerned with quantitative issues. Those outlined in this study have largely been resolved from sound knowledge of good experimental design and biometrical practice in ecology (Hairston 1989; Perry 1989, 1997; McArdle 1996).

Some problems arose specifically from the large extent of the study. These included the need for database management, data verification, punching, storage, integrity and extraction. The number of protocols is large, but it is desirable to have a common approach to analysis where possible. The development of statistical models that underpin the analysis was driven by the availability of data that built up slowly over a 3-year period; this resulted in a gradual evolution of analytical methods. These have been made available through the provision of standard Genstat 5 software (Payne & members of the Genstat 5 Committee 1993), running in interactive or batch mode. The plethora of possible analyses at each stage of the project imposed a requirement for these to be audited and results stored for later comparison.

A major statistical issue is whether one of the variance–mean relationships studied will prove clearly more appropriate than others. The possible advantage of the log-normal model, evident through diagnostic plots and in its agreement between parametric and non-parametric analyses, will be examined carefully. The apparent robustness of the d_w-statistic (Rothery, Clark & Perry 2002) may result in its wide use for other ecological studies involving count data.

Some motivation for the chosen degree of replication of fields per crop might have been drawn from the literature. The planned replication for the FSE exceeds,

by more than threefold, any of the comparable 82 terrestrial manipulative ecological experiments undertaken previously, for all plot sizes, reviewed by Moller & Raffaelli (1998) and Raffaelli & Moller (2000). Those are slightly different from the FSE because they refer to what are termed 'press experiments' in animal ecology, in which animals, often predators, were added to, or removed from plots. However, they represent the best recent set of unbiased data to compare with the FSE.

However, it is the power analyses that provide the confidence that replication is neither too small to detect obvious effects that might be present, nor so great that experimental resources could easily be redirected. At any stage in the project, available data may be used to derive a current estimate of power for a particular variate by estimating σ^2 from a simple ANOVA of l_{ij}, specifying a desired value of R and a projected value of n, and thereby deriving a value of Δ. For the log-normal model, the power for detecting a difference of three (or four) times the standard error is about 85% (or 98%). All the results described in this study suggest that if data were available for about 60 fields per crop the FSE would be replicated sufficiently, and should provide useful information from which valid statistical inferences may be drawn. This may subsequently be checked by plotting the logarithm of the estimated multiplicative treatment ratio vs. the logit-transformed P-value from the randomization test. The planting of extra fields was a sensible insurance against unforeseen losses.

The fact that power is a continuous function of sample size, not a step function, does not weaken the argument for adequate replication. However, it does strengthen the argument against naive claims that an experiment would be useless if there were a marginal failure to achieve some arbitrarily chosen target level of replication. This would be the case even if there were only a single variate of interest. It is even more strongly the case when there is a very large number of variates of interest, all of which vary differently. It will always be the case that for some of these power will be large while for others it will be small. In any event, even when the magnitude of the effect required to be detected has been specified quantitatively, there remains the difficulty of interpreting what this means in terms of environmental impact, given the buffering and resilience in ecosystems. Another desirable aspect of an experiment is consistency of results, especially one such as the FSE in which there are many protocols, some of which themselves involve many taxa.

Future studies to assess the ecological effects of GM crops on a large scale may be required for other crops, in other countries, and of alternative GM traits. It is unlikely that there will often be sufficient funding for experiments as intensive as the FSE. Therefore, there is an urgent need for further statistical methodological studies to develop experimental designs or modelling approaches that allow efficient study at reduced cost. One possible initial approach, available during late 2003, might be to backcast results from the FSE to estimate how reliable a study it would have been with reduced replication.

Acknowledgements

The FSE is funded by the Department for Environment, Food and Rural Affairs and the Scottish Executive; further details are given on the website http://www.environment.defra.gov.uk/environment/fse/index.htm. We thank all of our colleagues in the consortium for supplying FSE data and for their help, encouragement and advice. We thank Professor Mick Crawley (Imperial College at Silwood Park) and Dr Nicholas Aebischer (Game Conservancy Trust) of the Scientific Steering Committee for their comments, help in formulating ideas, and for supplying data that contributed towards the initial power calculations. We thank Professor Chris Pollock for his forbearance during lengthy discussions of mind-numbing statistical questions, and for his constant encouragement to try to keep these simple and clear. Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom.

References

Aebischer, N.J. (1990) Assessing pesticide effects on non-target invertebrates using long-term monitoring and time-series modeling. *Functional Ecology*, **4**, 369–373.

Anonymous (1998) *Genetically Modified Crops Threaten Wildlife*. Press Release EN/98/20, 27 March 1998. English Nature, Peterborough, UK.

Anonymous & Perry, J.N. (1999) *Design and Analysis of Efficacy Evaluation Trials. EPPO Guideline for the Efficacy Evaluation of Plant Protection Products, PP 1/152 (2)*. Vol. 1. *Introduction, General and Miscellaneous Guidelines, New and Revised Guidelines*. EPPO/OEPP, Paris, France.

Boatman, N.D. & Brockless, M.H. (1998) The Allerton Project: farmland management for partridges (*Perdix perdix*, *Alectoris rufa*) and pheasants (*Phasianus colchicus*). *Perdix VII: International Symposium on Partridges, Quails and Pheasants* (eds M. Birkan, L.M. Smith, N.J. Aebischer, F.J. Purroy & P.A. Robertson), pp. 563–574. Gibier Faune Sauvage, Paris, France.

Buzzard, C. (2000) *Lessons Learned from on-Farm Trials: the PD/A CRSP Experience*. PD/A CRSP, Oregon State University, Corvallis, OR.

Carroll, R.J. & Ruppert, D. (1988) *Transformation and Weighting in Regression*. Chapman & Hall, New York, NY.

Cochran, W.G. (1938) Some difficulties in the statistical analysis of replicated experiments. *Empire Journal of Experimental Agriculture*, **6**, 157–175.

Duffield, S.J. & Aebischer, N.J. (1994) The effect of spatial scale of treatment with dimethoate on invertebrate population recovery in winter-wheat. *Journal of Applied Ecology*, **31**, 263–281.

Firbank, L.G. & Forcella, F. (2000) Agriculture – genetically modified crops and farmland biodiversity. *Science*, **289**, 1481–1482.

Firbank, L.G., Dewar, A.M., Hill, M.O., May, M.J., Perry, J.N., Rothery, P., Squire, G.R. & Woiwod, I.P. (1999) Farm-scale evaluation of GM crops explained. *Nature*, **399**, 727–728.

Firbank, L.G., Heard, M.S., Woiwod, I.P., Hawes, C., Haughton, A., Champion, G., Scott, R., Hill, M.O., Dewar, A., Squire, G.R., May, M., Brooks, D.R., Bohan, D., Daniels, R.E., Osborne, J.L., Roy, D., Black, H.I.J., Rothery, P. & Perry, J.N.

(2003) An introduction to the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology*, **40**, 2–16.

Frampton, G.K. (1999) Spatial variation in non-target effects of the insecticides chlorpyrifos, cypermethrin and pirimicarb on Collembola in winter wheat. *Pesticide Science*, **55**, 875–886.

Hairston, N.G.S. (1989) *Ecological Experiments: Purpose, Design and Execution*. Cambridge University Press, Cambridge, UK.

Hastings, N.A.J. & Peacock, J.B. (1975) *Statistical Distributions*. Butterworths, London, UK.

Heads, P.A. & Lawton, J.H. (1983) Studies on the natural enemy complex of the holly leaf-miner – the effects of scale on the detection of aggregative responses and the implications for biological control. *Oikos*, **40**, 267–276.

Kennedy, P.J. (1994) The distribution and movement of ground beetles in relation to set-aside arable land. *Carabid Beetles: Ecology and Evolution* (eds K. Desender, M. Dufrene, M. Loreau, M.L. Luff & J.-P. Maelfait), pp. 439–444. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Kennedy, P.J., Conrad, K.F., Perry, J.N., Powell, D., Aegerter, J., Todd, A.D., Walters, K.F.A. & Powell, W. (2001) Comparison of two field-scale approaches for the study of effects of insecticides on polyphagous predators in cereals. *Applied Soil Ecology*, **17**, 253–266.

Krebs, J.R., Wilson, J.D., Bradbury, R.B. & Siriwardena, G.M. (1999) The second silent spring? *Nature*, **400**, 611–612.

Lennon, M. (1998) *Design and analysis of multiple site, large plot field experiments*. Unpublished PhD Thesis. University of Reading, Reading, UK.

Lewis, T. (1967) The horizontal and vertical distribution of flying insects near artificial windbreaks. *Annals of Applied Biology*, **60**, 23–31.

McArdle, B.H. (1996) Levels of evidence in studies of competition, predation and disease. *New Zealand Journal of Ecology*, **20**, 7–15.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman & Hall, London, UK.

Manly, B.F.J. (1994) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. Chapman & Hall, London, UK.

Moller, H. & Raffaelli, D. (1998) Predicting risks from new organisms: the potential of community press experiments. *Statistics in Ecology and Environmental Monitoring: Risk Assessment and Decision Making in Biology* (eds D.J. Fletcher, L. Kavalieris & B.J.F. Manly), pp. 131–156. Otago University Press, Dunedin, New Zealand.

Morgan, B.J.T. (1984) *Elements of Simulation*. Chapman & Hall, London, UK.

Moultine, G. & Bluman, A.G. (2001) *Minitab Manual for Use with Elementary Statistics*. McGraw-Hill, New York, NY.

Norowi, H.M., Perry, J.N., Powell, W. & Rennolls, K. (2000) The effect of spatial scale on interactions between two weevils and their parasitoid. *Ecological Entomology*, **25**, 188–196.

Numerical Algorithms Group (1997) *The NAG Fortran Library Manual Mark 18*. Numerical Algorithms Group Ltd, Oxford, UK.

Ogilvy, S.E., Turley, D.B., Cook, S.K., Fisher, N.M., Holland, J.M., Prew, R.D. & Spink, J. (1995) LINK integrated farming systems: a considered approach to crop protection. *Integrated Crop Protection: Towards Sustainability?* (eds R.W. McKinley & D. Atkinson), pp. 331–338. BCPC, Farnham, UK.

Payne, R.W. & members of the Genstat 5 Committee (1993) *Genstat 5 Release 3 Reference Manual*. Oxford University Press, Oxford, UK.

Pearson, E.S. & Hartley, H.O. (1976) *Biometrika Tables for Statisticians*, Vol. 2. Griffin, High Wycombe, UK.

Perry, J.N. (1986) Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology*, **79**, 1149–1155.

Perry, J.N. (1989) Review: population variation in entomology: 1935–50. I. Sampling. *Entomologist*, **108**, 184–198.

Perry, J.N. (1997) Statistical aspects of field experiments. *Methods in Ecological and Agricultural Entomology* (eds D.R. Dent & M.P. Walton), pp. 171–201. CAB International, Wallingford, UK.

Perry, J.N., Parker, W.E., Alderson, L., Korie, S., Blood-Smyth, J.A., McKinlay, R. & Ellis, S.A. (1998) Simulation of counts of aphids over two hectares of Brussels sprout plants. *Computers and Electronics in Agriculture*, **21**, 33–51.

Potts, G.R. & Vickerman, G.P. (1974) Studies on the cereal ecosystem. *Advances in Ecological Research*, **8**, 107–197.

Raffaelli, D. & Moller, H. (2000) Manipulative field experiments in animal ecology: do they promise more than they can deliver? *Advances in Ecological Research*, **30**, 299–338.

Robinson, R.A. & Sutherland, W.J. (2002) Post-war changes in arable farming and biodiversity in Great Britain. *Journal of Applied Ecology*, **39**, 157–176.

Rothery, P., Clark, S.J. & Perry, J.N. (2002) Design and analysis of farm-scale evaluations of genetically modified herbicide-tolerant crops. *Proceedings of the XXith International Biometric Conference, 21–26 July 2002, Freiburg, Germany, Invited Papers* (ed. M. Schumacher), pp. 351–364. German Region of the International Biometric Society, Freiburg, Germany.

Schuler, T.H., Poppy, G.M., Kerry, B.R. & Denholm, I. (1999) Potential side effects of insect-resistant transgenic plants on arthropod natural enemies. *Trends in Biotechnology*, **17**, 210–216.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry: The Principles and Practice of Statistics in Biological Research*. W.H. Freeman, New York, NY.

Sotherton, N.W., Jepson, P.C. & Pullen, A.J. (1988) Criteria for the design, execution and analysis of terrestrial non-target invertebrate field tests. *BCPC Monograph*, **40**, 183–190.

Taylor, L.R. (1961) Aggregation, variance and mean. *Nature*, **189**, 732–735.

Taylor, L.R., Woiwod, I.P. & Perry, J.N. (1978) The density-dependence of spatial behaviour and the rarity of randomness. *Journal of Animal Ecology*, **47**, 383–406.

Vet, L.E.M. (1999) From chemical to population ecology: infochemical use in an evolutionary context. ISCE Silverstein-Simeone Lecture Award, 1 May 1998. *Journal of Chemical Ecology*, **25**, 31–49.

Watkinson, A.R., Freckleton, R.P., Robinson, R.A. & Sutherland, W.J. (2000) Predictions of biodiversity response to genetically modified herbicide-tolerant crops. *Science*, **289**, 1554–1557.

Received 5 June 2002; final copy received 26 October 2002