Long-Term Experiments with cropping systems: Case studies on data analysis

Long-term experiments with cropping systems: case studies on data analysis

ANDREA ONOFRI[1], GIOVANNA SEDDAIU[2,3], HANS-PETER PIEPHO[4]

[1]Dipartimento di Scienze Agrarie, Alimentari e Ambientali, Università degli Studi di Perugia, Italy

[2]Dipartimento di Agraria, Università di Sassari, Italy

[3]Nucleo di Ricerca sulla Desertificazione (NRD), Università di Sassari, Italy

[4]Biometrics Unit, University of Hohenheim, Stuttgart, Germany

Correspondence: Prof. Andrea Onofri, Dipartimento di Scienze Agrarie ed Ambientali, Università degli Studi di Perugia, Borgo XX Giugno 74, 06121, Perugia, Italy. Tel +39-075-5856324; fax: +39-075-5856344; e-mail: andrea.onofri@unipg.it

Abstract

Data analysis for Long-Term Experiments (LTEs) with cropping systems requires some careful thinking, especially for the most complex designs, characterised by rotations with different durations and/or a different number of test-crops per rotation cycle. This paper takes an example-based approach, built upon a number of datasets, covering the main types of LTEs, with increasing levels of complexity. A procedure is outlined to build statistical models for data analysis that is useful for all LTEs characterised by the simultaneous presence of all rotation phases in all years, together with within-year replication. This procedure is based on the assumption that correct analyses can be performed separately for each year. The use of mixed models and REML estimation is advocated for model fitting with all LTEs, due to the fact that most designs are non-orthogonal, as plots may not produce data for the test-crop under study in all years. Mixed models are also useful to account for the autocorrelation of residuals over time and hints are given for the selection of an appropriate variance-covariance structure.For all our examples, variances were not constant across years and compound symmetry correlation structures with variance heterogeneity of years proved to be the best compromise betweenparsimony and statistical accuracy. Methods are outlined to test for the need of other more complex correlation structures and examples are also given on how to test for fixed effects, model fertility trends and assess the long-term stability of cropping systems.

Key Words: rotation, repeated measures, mixed model, autocorrelation, variance heterogeneity

Highlights

- Long-Term Experiments (LTEs) with cropping systems require careful data analyses
- Five types of LTEs with increasing complexity are considered and analysed
- A correct model building procedure is outlined and the use of REML estimation is advocated
- The selection of variance-covariance structures and tests for fixed effects are discussed
- Long-term ‘fertility’ trends and stability analyses are also considered

1. Introduction

Long-Term Experiments (LTEs) represent an invaluable tool to detect possible slow changes produced by the cropping systems in the long run and reveal possible threats to the environment and to the future fertility of agricultural land. From a methodological point of view, LTEs may be regarded as a particular class of multi-environment experiments, wherein observations are repeated on the same plots for a long period of time, usually from 20 years upwards (Frye and Thomas, 1991). The experimental treatments under comparison may be highly variable (rotations, fertilisations, tillage systems ...), but in most cases they represent several cropping systems, to be investigated by repeated observations of yield or other relevant indicators of productivity, sustainability, efficiency and so on.

In general, data analysis for LTEs is not trivial, due to the peculiar characteristics of this class of experiments. In a recent review, Payne (2015) highlighted several special issues, such as: (i) non-uniform correlation structure of within-plot observations; (ii) heterogeneity of variances in different years; (iii) residual effects of agronomic practices, leading to fertility trends over time; (iv) changes in the experimental protocol, due to, e.g., the adoption of new varieties, harvesting methods or weed control methods. Unfortunately, the statistical solutions to these (and other) problems may not be unique and straightforward. Indeed, a survey of literature shows that LTEs employ a wide array of experimental designs with their own peculiarities. For example, rotation experiments may be designed so that all crops in the rotation are grown every year (Cochran, 1939). However, in order to avoid an excessive increase in the total number of plots, LTEs may also be designed without within-year replicates (Patterson, 1964), thus requiring a different approach to data analysis. Another example relates to the frequency of repeated measures: in long rotations the test crop may return on the same plot after a relatively long period of time, which may create a totally different correlation structure compared to short rotations.

We will not list all the possible peculiarities of LTEs in terms of experimental design, as this has already been accomplished in literature (see, e.g., Frye and Thomas, 1991). It should be clear that each LTE, according to its aims and experimental design, poses peculiar problems in terms of data analysis and may create difficulties in the selection of an appropriate statistical method.

In this regard, a survey of literature shows a number of early reviews, discussing ANOVA-like models for LTEs with a wide array of experimental designs (Patterson, 1964; Yates, 1954). However, these reviews use traditional methods of data analyses, while the current availability of personal computers and mixed model procedures gives scientists some new challenges and perspectives. On the other hand, more recent reviews about LTEs give examples relating to few experimental designs and do not consider the most complex situations, such as when LTEs include rotations of different lengths and/or with a different number of test crops (see for example Richter and Kroschewski, 2006; Singh and Jones, 2002). Another review about repeated measures includes only an example referring to a LTE comparing rotations of equal lengths (Piepho et al., 2004). A further recent review takes an example-based approach and presents a stepwise procedure for data analysis, though this is based on a single experiment, characterised by no within-year replication (Payne, 2015).

Considering the above background, we felt that a further example-based treatment of the statistical analysis of LTE data would be useful. Therefore, we used a number of datasets with contrasting characteristics and elaborated several case studies to give examples on how to: (i) build an appropriate model, (ii) inspect the validity of basic assumptions, (iii) test hypotheses of interest and (iv) obtain reliable measures of productivity, stability and sustainability for the cropping systems under comparison. The code to perform the analyses is made available as supplemental material, with the aim of giving practical directions when dealing with LTEs.

2. Five datasets for five types of long-term experiments

Focusing on LTEs characterised by the presence of within-year replicates, five main groups can be found in literature, with increasing levels of complexity in terms of data analysis:

1. LTEs with monocultures or perennial crops;
2. LTEs with different rotations of same length and one test crop per rotation cycle;
3. LTEs with a fixed rotation (one test crop per rotation cycle) and different treatments;
4. LTEs with more than one phase per crop and rotation cycle;
5. LTEs with several rotations of different lengths and/or different numbers of phases per crop and rotation cycle.

Relating to the terms rotation, phase, cycle, sequence, we will refer to the terminology used in Yates (1954), which is summarised in Table 1.

[Table 1 about here]

We cannot cover all these situations with datasets from real and independent experiments and, therefore, instead of using simulated data, we decided to take a 'hybrid' approach: we started from two real field experiments and formed different subsets of the data to fit the above list. In more detail, we used the datasets from two real long-term field experiments: the first was established around 1974 near Perugia (central Italy), at the Experimental Farm of the Department of Agricultural, Food and Environmental Sciences and the second was established in 1994 at the Pasquale Rosati experimental farm in Agugliano (central Italy). Both experiments are still running at present (2015) and a detailed description is given as supplemental material, while other information can be found in Perucci et al. (1997) and elsewhere in this same special issue (Bonciarelli et al., 2016).

These two real LTEs were used to extract five exemplary datasets, representing the five types of LTE listed above. It is important to point out that, for the sake of our exercise, these five datasets need not be considered in connection with the two original LTEs and may be taken to represent five 'ideal' and independent LTEs. Readers interested in full-fledged analyses of the original datasets are referred to the aforementioned references.

Table 2 summarises the contrasting characteristics of these datasets that will be described in detail afterwards. For each dataset, some records are given as supplemental materials, in order to show the exact structure of data tables.

[Table 2 about here]

2.1 LTEs with monocultures or perennial crops (Dataset 1)

This type of LTE does not involve rotations and the treatments under comparison consist of different cropping practices (e.g. fertilisations). In terms of data analysis, these LTEs are similar to experiments with perennial crops, as observations are taken all years on the same plot, for each treatment level and replicate (Eckl and Piepho, 2015; Piepho and Eckl, 2014; Smith et al., 2007).

Dataset 1: Continuous wheat cropping. Wheat is grown in continuous cropping from 1983 to 2012, with three fertilisation levels (150, 200 and 250 $kgNha^{-1}$), randomly assigned to three plots in each of three blocks. In all, there are nine plots with yearly sampling, with a total of 270 wheat yield observations in 30 years.

2.2. LTEs with different rotations of the same length and one test crop per rotation cycle (Dataset 2)

If we have, e.g., two crops in a rotation (maize and wheat), both crops will be grown in different plots in the same year and we will have two possible sequences in time (maize-wheat and wheat-maize). If we consider only one of the two crops, the main difference with respect to the previous situation is that the data obtained in two consecutive years for the same treatment and block are independent, in the sense that they are obtained in different plots. Otherwise, data obtained in a two-year interval (on different rotation cycles) on the same block are correlated, as they originate from the same plot.

*Dataset 2: Comparing wheat yield for several biennial rotations.*Wheat is grown in five types of two-year rotations, with either pea (*Pisum sativum* L.), grain sorghum (*Sorghum bicolor* (L.) Moench), sugar beet (*Beta vulgaris* L. subsp. *saccharifera*), sunflower (*Helianthus annuus* L.) and faba bean (*Vicia faba* L. subsp. *minor*). For each rotation, there are two possible sequences (wheat in odd years and wheat in even years) and the ten combinations (five rotations by two sequences) are completely randomised to ten plots per each of three blocks (Table 3). Therefore, five wheat plots out of the available ten plots are used from each block and year, for a total of 450 observations, from 1983 to 2012.

[Table 3 about here]

2.3 LTEs with a fixed rotation (one test crop per cycle) and different treatments (Dataset 3)

This type of LTE is very similar to the previous one, though we will introduce a dataset with a different experimental layout.

*Dataset 3: Comparing wheat yield in one biennial rotation with different tillage/fertilisation systems.*Durum wheat (*Triticum durum* L.) is grown in a two-year rotation with a spring crop and nine treatments, consisting of the factorial combination of three soil tillage methods (T: conventional 40 cm deep ploughing; M: scarification at 25 cm; S: sod seeding with chemical desiccation and chopping) and three N-fertilisation levels (0, 90 and 180 kg N ha^{-1}). The two possible rotation sequences (wheat-spring crop and spring crop-wheat) are arranged in two adjacent fields, which therefore host the two different crops of the rotation in the same year. Within the two fields, there are two independent randomisations, each with two blocks.Tillage levels are randomised to main-plots (1500 m^2) and N levels are randomised to sub-plots (500m^2), according to a split-plot design with two replicates.

Also for this third dataset, the response variable is wheat yield; in terms of data analysis, it is relevant to note that: (i) there are two experimental factors, laid out in a split-plot design; (ii) the two sequences are accommodated in two fields (Figure 1).

[Figure 1 about here]

2.4. LTE with a fixed rotation, different treatments and more than one phase per crop and cycle (Dataset 4)

The situation becomes more complex if we have, e.g., a three-year rotation with two wheat crops (maize-wheat-wheat). In this case there are three possible sequences and, relating to wheat, we have two distinct phases in each rotation cycle (phase difference: wheat after maize and wheat after wheat).

*Dataset 4: burial/removal of crop residues on a three-years rotation.*Wheat is grown in a three-year rotation maize-wheat-wheat, under two types of management of crop residues (burial and removal), which are randomised to main plots, while the three possible rotation sequences are randomised to subplots. This experiment has 18 plots (three sequences×two treatment levels×three blocks) and, in every year, 12 of those are cropped with wheat and six with maize.

Also in this case, the response variable is wheat yield from 1983 to 2012, i.e. twelve observations per year and 360 observations in total. Data obtained in the same plot in different years belong to two different phases (wheat after maize and wheat after wheat; Table 4).

[Table 4 about here]

2.5. LTEs with several rotations of different lengths and different number of phases per crop and rotation cycle (Dataset 5)

In some cases, it is necessary to compare several rotations with different characteristics (e.g., a different duration and/or a different number of tests crops and /or a different number of phases per crop), which may create a complex design with some degree of non-orthogonality.

*Dataset 5: comparing five maize-wheat rotations.*Wheat is grown in five maize (M) - wheat (W) rotations of different lengths, i.e., M-W, M-W-W, M-W-W-W, M-W-W-W-W, M-W-W-W-W-W. For all rotations, all phases are simultaneously present in each year, for a total of 20 plots (one for each of the possible sequences, i.e. 2 + 3 + 4 + 5 + 6 = 20) in each of three blocks. Considering wheat yield as the response variable, we find that only 15 observations are obtained in each year and block, for a total of 1350 records, from 1983 to 2012.

Experiments of this type represent a high degree of complexity. Indeed, in contrast to all other examples, after 30 years there are plots with: (i) a different number of observations for the same test crop; (ii) a different number of cycles (in some cases the last cycle is also incomplete); (iii) a different number of phases for wheat.

3. Steps to model building

Every good analysis is based on a good model that is biologically relevant and based on realistic assumptions. We will start from the general method and notation proposed by Piepho et al. (2003) that is naturally related to computer implementation by statistical packages for linear models. A similar model building procedure has been devised in Brien and Demetrio (2009), which should yieldcomparable results for the examples under investigation. With respect to these approaches, we will propose some slight changes, specific for LTEs.

In general, ANOVA-like models are based on classification variables, commonly known as factors. In order to build correct models, we need to distinguish three types of factors: (i) *treatment factors*, which are randomly allocated to randomisation units (e.g. rotations, fertilisations, management of crop residues); (ii) *block factors*, which group the experimental units according to some 'innate' (not randomly allocated) physical criterion (e.g. by position), such as the blocks, the locations, the main-plots, the sub-plots and so on. Block factors may represent the randomisation units, to which treatments are randomly allocated; (iii) *repeated factors*, which relate to time and thus cannot be randomised (e.g. years, cycles ...).

All the above factors are given in capital letters and are combined by using the following operators: (i) the 'dot' operator denotes crossed effects (e.g. $A \cdot B$ means that A and B are crossed factors); (ii) the 'nesting' operator denotes nested effects (e.g. A/B means that B is nested within A and it is equal to $A + A \cdot B$); (iii) the 'crossing' operator denotes the full factorial model for two terms ($A \times B = A + B + A \cdot B$). This brief information may be sufficient for our aims, but other operators and syntactic rules are given in the paper by Piepho et al. (2003) and in a follow-up paper about repeated measures (Piepho et al., 2004), which might be of interest to the readers.

The steps to model building may be summarised as follows:

1. Select the repeated factor.
2. Consider one fixed level of the repeated factor and build a treatment model for the randomized treatment factors.
3. Consider one fixed level of the repeated factor and build a block model for block factors.
4. Check whether randomised treatment factors might interact with block effects: if such an interaction is to be expected it should be added to the model.
5. Include the unrandomised repeated factor into the model.
6. Combine treatment model and repeated factor model, by crossing or nesting as appropriate.
7. Consider which effects in the block model reference randomisation units, i.e. those units which receive the levels of a factor or factor combination by a randomisation process. Assign to the corresponding terms a separate random effect, as explicitly recommended in Piepho et al. (2004).
8. Excluding the terms for randomisation units, nest the repeated factor in all the other terms in the block model.
9. Combine random effects for randomisation units with the repeated factor, by using the dot operator, in order to derive the correct error terms to accommodate serial correlation structures.

The key idea for the above approach is that for a properly designed experiment, valid analyses should be possible for the data at each single level of the repeated factor. Such a basic requirement should never be taken for granted, but it should be carefully checked before the beginning of model building process (see later for Dataset 3).

Dealing with rotations, steps 1 and 2 are crucial when we are only interested in one of the crops (phases) in the rotation. We will show this with an example.

3.1. Analysis of Dataset 2: choosing the repeated factor and treatment model

With Dataset 2 (LTE with wheat in five different two-year rotations; Tab. 2) we are looking only at one phase in the rotation (in this case wheat) and we should note that observations are repeated every second year on the same plot (Tab. 3), according to the sequence they belong to. In other words, observations are repeated on each rotation cycle (two years) in the same plot, while there is neither a within-cycle repetition nor a within-cycle phase difference: we have only one observation per plot per cycle. Therefore, it is natural to take the rotation cycle as the repeated factor.

As the next step, we should look at what happens in one fixed level of a two-year cycle: what did we randomize to the ten plots? It is clear that, considering only wheat, we randomised each combination of rotation and positioning in the sequence (i.e. wheat as the first crop of the sequence and wheat as the second crop of the sequence; see Table 1). Therefore, sequence and rotation should be included in the model as the treatment factors.

This approach is commonly suggested in literature (see Yates, 1954) and it is convenient, mainly because the resulting model is orthogonal and may be fitted by ordinary least squares. Indeed, for Dataset 2 (and similar experiments), there is only one observation for each block, treatment, cycle, sequence and no missing data (in our case: 3 blocks×5 rotations×2 sequences × 15 cycles = 450 observations). The phase should not enter into this model, as we are looking only at one of the two crops (only one phase).

However, the drawback is that such an approach cannot be immediately extended to the other more complex examples (e.g. rotations with different lengths and/or with a different number of test-crops). Furthermore, it partitions the year factor into three effects, i.e. 'cycles', 'sequences' and 'cycle×sequences', which might make modelling any 'fertility' trends over time less obvious. In this respect, we should note that possible differences between sequences for a given cycle (wheat as the first crop of the sequence and wheat as the second crop of the sequence, i.e. wheat in even years and wheat in odd years) do not carry any meaning that helps understand the behaviour of rotations.

Therefore, it may be more convenient to take the year as the repeated factor, considering that, for the case of Dataset 2, this effect is totally confounded with the factorial combination of 'cycle' and 'sequence' (15 cycles×2 sequences = 30 years). This gives us a good common modelling platform for all datasets, although it may be argued that the models are no longer orthogonal, as not all plots produce data in all years. It should be recognised, however, that the lack of orthogonality can easily be accommodated within mixed models. If we take the year as the repeated factor, it is more convenient to embrace the general view that in one level of year we randomise the rotations and the phases for each rotation to the plots. Therefore, in the case of Dataset 2, where we are interested only in one of the two phases, the treatment model will only contain the rotation term (see later).

The situation is totally different if we look at both the phases of the rotation (e.g. wheat and sunflower): in this case, we have a phase difference within each cycle and, considering one level of the repeated factor year, the treatment model should contain both the rotation and the phase, together with their interaction. When we introduce the year (steps 5 and 6 above), we also introduce the interactions 'year×rotation', 'year×phase' and 'year×rotation×phase', which are all meaningful when studying the behaviour of rotations. We will no longer pursue this aspect with Dataset 2, as we decided to consider only wheat as the test-crop. However, we will again consider the phase difference with Datasets 4 and 5.

On the above basis, we will now define the models for all datasets in Table 2, considering the year as the repeated factor. We will also partition the fixed model and the random model, as commonly done in several statistical packages (SAS, ASReml-R...). At first, the random model will include all random terms for randomisation units (terms at steps 7 and 9), while the fixed model will include all the other terms. Later on, we will discuss the possible extensions/changes to this basic approach.

4. Building statistical models for all datasets

4.1. Analysis of Dataset 1

If we consider Dataset 1 (LTE on continuous wheat at three N-fertilisation levels; Tab. 2), the repeated factor is the year (YEAR). In one year, the treatment factor is nitrogen fertilisation (N) and there are two block factors, i.e. the blocks (BLOCK) and the plots within each block (PLOT). Therefore, the block model is BLOCK + BLOCK · PLOT.

We now introduce the repeated factor YEAR and combine it with the treatment model, by including N × YEAR = N + YEAR + N · YEAR. The term BLOCK · PLOT references the randomisation units and receives a random effect. As the year might interact with the block, we add the term BLOCK · YEAR. We also combine the year with the random effect for plots (BLOCK · PLOT · YEAR), although this residual term is usually automatically fitted and does not need to be explicitly coded when implementing the model. The final model is (the operator ∼ means 'is modelled as'):

YIELD ∼ N + BLOCK + YEAR + N · YEAR + BLOCK · YEAR

RANDOM: BLOCK · PLOT + BLOCK · PLOT · YEAR (Model 1)

We have already mentioned that one of the problems with LTEs is that the observations taken on the same plot are not independent. Model 1 is able to account for this, owing to the presence of the two random effects BLOCK · PLOT (often called 'plot' error, with variance σ_B^2) and BLOCK · PLOT · YEAR (residual error or within-plot error, with variance σ^2). When the variance for BLOCK · PLOT is higher than 0, it is implied that observations in the same plot 'co-vary', i.e. show a positive covariance equal to σ_B^2. The variance for one observation (same plot and same year) is equal to the sum $\sigma_B^2 + \sigma^2$ and, therefore, the correlation among observations in the same plot is equal to the ratio $\rho = \sigma_B^2/(\sigma_B^2 + \sigma^2)$ (intra-class correlation). This correlation is equal for all pairs of observations in each plot, regardless of their distance in time.

This would be similar to a split-plot design (split-plot in time) with the important difference that, in this case, the sub-plot factor (year) is not randomised. The corresponding serial correlation structure is known as compound symmetry (CS). It is worth noticing that the same model may be fitted in an alternative way, i.e. by dropping the BLOCK · PLOT random effects and using the residual term BLOCK · PLOT · YEAR to accommodate the CS structure into the model. This may be done very intuitively e.g. in SAS, by using the statement:

```
repeated year/subject = block*plot type=CS;
```

This is very useful when a simple CS correlation structure is not satisfactory, because the observations in the same plot are not equally correlated (e.g. observations close in time are more correlated than those distant in time). In this case, we can implement alternative correlation structures by using the residual term BLOCK · PLOT · YEAR, e.g. by appropriately changing the specification of the 'type=' option in the above SAS statement.

4.2. Analysis of Dataset 2 (continued)

Considering Dataset 2 (LTE with wheat in five different two-year rotations; Tab. 2), we consider the YEAR as the repeated factor. In one year the treatment model is composed only by the rotation (ROT), which is randomised to plots (PLOT), within blocks. The block model for one year is BLOCK/PLOT = BLOCK + BLOCK · PLOT(the use of the nesting operator gives a more compact definition). We can combine the treatment model with the repeated factor (ROT · YEAR) and add the term BLOCK · YEAR. The final model is:

YIELD ~ ROT + BLOCK + YEAR + ROT · YEAR + BLOCK · YEAR

RANDOM: BLOCK · PLOT + BLOCK · PLOT · YEAR (Model 2)

Apart from random effects for randomisation units, this model is similar to the one used for multi-environment genotype experiments and, represents a convenient and clear platform for the analyses of LTE data. We remind the reader that the residual term (BLOCK · PLOT · YEAR) does not need to be explicitly coded when implementing the model, as it is automatically fitted.This basic model again implies the compound symmetry structure for serial correlation of plot errors;more refined modeling is possible by imposing alternative serial correlation structures.

4.3. Analysis of Dataset 3

Before proceeding to model building with Dataset 3 (LTE for wheat in one biennial rotation with different tillage/fertilisation systems; Tab. 2 and Fig. 1), we need to discuss whether valid analyses are possible at each single level of the repeated factor. Indeed, this is clearly true if we take the year as the repeated factor and consider only one of the two crops in the rotation (wheat, in this case). However, if we intended to consider both crops and compare e.g. their yields, the crop effect would be confounded with the field effect within a single year and, therefore, valid within-year analyses would not be possible. In this case, we should resort to taking the rotation cycle as the repeated factor.

Dealing only with wheat, we can therefore take the YEAR as the repeated factor and consider that, in one year, the randomised treatment factors are tillage (T) and nitrogen fertilisation (N) and the treatment model is T + N + T·N.

The block factors are the FIELDS, the BLOCKS within fields, the main plots (MAIN) within blocks and the subplots (SUB) within main plots. The block model (for one year) is FIELD/BLOCK/MAIN/SUB = FIELD + FIELD · BLOCK + FIELD · BLOCK · MAIN + FIELD · BLOCK · MAIN · SUB.

The treatment and repeated model can be combined as: (T + N + T·N) × YEAR = T + N + T · N + YEAR + T · YEAR + N · YEAR + T · N · YEAR.

At this stage, the FIELD main effect needs to be removed, as it is totally confounded with the years in this specific example. We assign a random effect to the other randomisation units, i.e. FIELD · BLOCK, FIELD · BLOCK · MAIN and FIELD · BLOCK · MAIN · SUB and combine these random terms with the repeated factor YEAR, by using the dot operator, which leads us to:

YIELD ~ T + N + T · N + YEAR + T · YEAR + N · YEAR + T · N · YEAR

RANDOM: FIELD · BLOCK + FIELD · BLOCK · MAIN + FIELD · BLOCK · MAIN · SUB + FIELD · BLOCK · YEAR + FIELD · BLOCK · MAIN · YEAR + FIELD · BLOCK · MAIN · SUB · YEAR

(Model 3)

As usual, the last term (residual) does not need to be explicitly included when implementing the model, but it can be used, together with the two previous ones (FIELD · BLOCK · YEAR + FIELD · BLOCK · MAIN · YEAR) to accommodate possible serial correlation structures into the model, by allowing year-specificity of all design effects and the residuals.

4.4. Analysis of Dataset 4

We can proceed in the same fashion as above with Dataset 4 (LTE with burial and removal of crop residues; Tab. 2), considering that the year is the repeated factor and, in one year, the treatment factors are the management of soil residues (RES, that is randomly allocated to main-plots) and the phases (P; randomly allocated to subplots); the treatment model is indeed: RES × P.

In one year, the block model is: BLOCK/MAIN/SUB = BLOCK + BLOCK · MAIN + BLOCK · MAIN · SUB. Introducing the YEAR as repeated factor, we can combine the treatment model with the repeated model as: RES + P + RES · P + YEAR + RES · YEAR + P · YEAR + RES · P · YEAR.

The terms BLOCK · MAIN and BLOCK · MAIN · SUB reference randomisation units and should receive random effects. The blocks may interact with the years (BLOCK · YEAR), while the random effects for randomisation units can be made year-specific by adding BLOCK · MAIN · YEAR and the residual term BLOCK · MAIN · SUB · YEAR. The final model is:

YIELD ~ RES + P + RES · P + YEAR + RES · YEAR + P · YEAR + RES · P · YEAR + BLOCK + BLOCK · YEAR

RANDOM: BLOCK · MAIN + BLOCK · MAIN · SUB + BLOCK · MAIN · YEAR + BLOCK · MAIN · SUB · YEAR (Model 4)

4.5. Analysis of Dataset 5

This further case study relates to Dataset 5 (comparison between five maize-wheat rotations of different lengths; Tab. 2), where the repeated factor is again the year. In one year, the treatment factors are the rotation system (ROT) and the rotation phase (P), which are randomly allocated to plots. As there is a different number of phases for each rotation, we nest the phase within the rotation, leading to the following treatment model: ROT/P = ROT + P · ROT.

In one year, the block model is BLOCK/PLOT = BLOCK + BLOCK · PLOT. The repeated factor is included and combined with the treatment model, by introducing YEAR + ROT· YEAR + ROT· P · YEAR.

The term BLOCK · PLOT references randomisation units and needs to receive a random effect, while the term YEAR · BLOCK can be added to the model, together with the residual term BLOCK · PLOT · YEAR. The final model is:

YIELD ~ ROT + P · ROT + YEAR + ROT· YEAR + ROT· P · YEAR

RANDOM: BLOCK · PLOT + BLOCK · PLOT · YEAR (Model 5)

5. Model implementation

Models 1 to 5 are fairly similar and they are very closely related to those used for multi-environment experiments. However, apart from Dataset 1, there is always a certain degree of imbalance, as plots do not produce data every year. Therefore, these models are not generally amenable to ordinary least squares estimation and traditional ANOVA and they should be implemented by using the mixed model framework, based on REstricted Maximum Likelihood (REML) or Maximum Likelihood (ML) estimation (see also Payne, 2015).

A second aspect that needs to be taken into account is that, in all cases, the above model formulations induce the aforementioned compound symmetry correlation structure ('split-plot in time'), which may not be appropriate in all circumstances. We will discuss alternative correlation structures shortly.

Last, but not least, we have already mentioned that, in the above models, only randomisation units have been given a random effect, while all the other effects have been regarded as fixed. Obviously, depending on the aims of the analyses, it might be convenient and appropriate to regard the year factoras random. In this case, all interactions containing this factor arealso random. Examples will be given later on.

Should we be happy with the above structure (random effects only for randomisation units and compound symmetry), the implementation of models 1 to 5 is straightforward in most statistical packages (R, SAS, GENSTAT . . . .). Examples of implementation in R are given as supplemental material (Code S1). Some useful SAS code can be found in Piepho (1999), while GENSTAT, R and SAS code is given in Payne (2015).

6. Model checking and extensions

Though the above models are rather easy to implement by using advanced statistical software, we should never forget that they make a number of basic assumptions that should be carefully checked before going ahead with the analyses. In particular, it is assumed that:

1. within group errors are normal and homoscedastic;
2. random effects for randomisation units are normal and independent from one another (if more than one random effect exists);
3. subjects in the same group (for example years within plots) are equally correlated, while subjects in different groups are independent.

This latter assumption (split-plot in time), in particular, seems to be rather questionable and more advanced correlation patterns have been advocated, e.g. autoregressive and unstructured (Richter and Kroschewski, 2006; Singh and Jones, 2002). In brief, the autoregressive approach assumes that observations in the same plot are not equally correlated, but their correlation depends on the time span between them. On the other hand, the unstructured approach is the least restrictive, as it does not assume any particular pattern for within-plot correlation. In principle, this latter approach is rather appealing, though it requires a large number of parameters to be estimated, which may not be supported by the data at hand and may give computational problems.

The suggested approach to the selection of variance-covariance structures is based on fitting all possible models and making an 'a posteriori' selection, based on a likelihood ratio test or, preferably, on those statistics which put a penalty on 'complexity', such as the Akaike Information Criterion (AIC: Akaike, 1974), which seeks to find the approximating model (Burnham and Anderson, 2002). Additionally, we suggest a graphical exploration of residuals; indeed, this procedure is very common with traditional linear models and gives a better 'feel' for the data.

Accommodating variance and correlation structures in a mixed model is usually done by exploiting the residual term, e.g. BLOCK · PLOT · YEAR and using the YEAR as repeated effect and the non-repeated error term BLOCK · PLOT as subject effect (Piepho et al., 2004). It is important, however, to highlight that, with split-plot designs or in presence of other randomisation structures apart from plots, it is important to allow for year-specificity of all design effects, not only the residual error term. For example, we might use the main plot random effect (e.g. BLOCK · MAIN · YEAR in Dataset 4) as above, to impose a year-specific correlation structure to main-plots.

The implementation of variance and correlation structures is straightforward with SAS. In the case of R, the implementation may be done by using the commercial package ASReml-R (Butler et al., 2009), while it becomes tricky if a freeware solution is envisaged for designs with several randomisation units. Indeed, it should be considered that the most recent function for mixed models in R, i.e. *lmer()* from the package *lme4*(Bates et al., 2015) does not yet support non-standard variance-covariance structures and it may be therefore necessary to use the function *lme()* from the package *nlme* (Pinheiro and Bates, 2000). However, this latter function is optimised for nested random effects and it may require a difficult coding for crossed random effects (see Gałecki and Burzykowski, 2013 for suggestions).

The following example is based on Dataset 1, but a similar approach can be used with every LTE.

6.1. Analysis of Dataset 1 (continued): model implementation and checking

Dataset 1 was used to parameterise Model 1 (Tab. 2). The usual diagnostic plots (i.e. a plot of residuals against fitted values and a QQ plot of residuals against a standard normal distribution) did not reveal any deviation from homoscedasticity and normality. However, a boxplot of within-group residuals plotted by year showed clear signs of heterogeneity of variances between years (Fig. 2).

[Figure 2 about here]

With reference to the possible serial correlation of within-group residuals, we decided to inspect a lag-plot of residuals, i.e. a graph showing, for each field plot, the residual observed at time i against the residual observed at time $i + 1$. This is a 'naive' approach, but it might help visualise possible problems, when a consistent linear trend is observed. With this respect, the plots in Figure 3 do not give any particular reason for concern.

[Figure 3 about here]

The code for the above graphical exploration in R is given as supplemental material (Code S2).

We concluded this preliminary analyses by introducing in the model the following correlation structures (we'll use the same notation as in Richter and Kroschewski (2006)): (i) NC: no correlation, neglecting also the random effect for randomisation units (plots); (ii) CS: compound symmetry, i.e. equal correlation for all observations in each plot; (iii) CSH: compound symmetry with heterogeneity of variances caused by years (one variance estimate per year); (iv) AR(1): autoregressive approach, where the correlation between two observations depends on the lag distance between them, according to an exponential function; (v) ARH(1): autoregressive approach, with heterogeneity of variances caused by years (one variance estimate per year); (vi) UN: totally unstructured pattern of within-plot correlations.

It is important to point out that the presence of the random term BLOCK · PLOT in combination with the random term BLOCK ·PLOT ·YEAR induces a compound symmetry structure and, in order to implement different structures, it is fundamental that the former term is dropped first, to avoid convergence / identifiability problems, especially with an unstructured model. After removing this term, we may use the residual term as explained above.

In this case, the UN approach gave convergence problems and it was clearly overly complex, with respect to the data at hand (Tab. 5). For the other approaches, however, the log-likelihood always increased when more complex variance-covariance structures were added to the model. Among these, the CSH approach appeared to give the best balance among statistical accuracy and simplicity, based on the AICstatistic(Tab. 5).

The code to reproduce the analyses in R is given as supplemental material (Code S3)

[Table 5 about here]

7. Testing for fixed effects

Testing for fixed effects has always been regarded as a fundamental part of data analyses, traditionally based on ANOVA tables and F-tests built upon expected mean squares and 'error strata'. Switching to mixed models, a different approach is required. With ML estimation, likelihood ratio tests (LRTs) would be a straightforward procedure, although the null distribution for these tests is only approximately chi-squared and such an approximation may be rather poor with small sample sizes, leading to 'anti-conservative' p-values. Better results may be achieved by using a parametric bootstrap approach that is shown later.

However, REML estimation is usually preferred with mixed models, as variance component estimates are known to have better properties (Searle et al., 1992). In this case, Wald-type F tests can be used as in the usual traditional ANOVA approach, though it is necessary to consider that, when data are unbalanced (which is almost always the case with our models), the denominator degrees of freedom are not easily obtained and some correcting method is mandatory (e.g. the Kenward-Roger approach: Kenward and Roger, 2009; Kenward and Roger, 1997).

7.1. Analysis of Dataset 2 (continued): testing for fixed effects

For this case study, we were interested in testing the significance of the 'treatment×year' interaction. We fitted Model 2 to Dataset 2 (Tab. 2), regarding all effects as fixed and considering (after preliminary analyses) a simple within-plot compound symmetry (CS) correlation structure, without heteroscedasticity of years. Working with R and the function*lmer()*in the *lme4* package, we performed a conditional F test by using the function *anova()*, which returneda F-value of 1.74, with 116 and 217 unadjusted degrees of freedom (p = 0.0002).

A Kenward-Roger approximation for degrees of freedom is more reliable and requires two additional packages in R, i.e. *lmerTest*(Kuznetsova et al., 2015) and *pbkrtest*(Halekoh and Højsgaard, 2014). In this case, we obtaineda number of denominator degrees of freedom equal to 182.56 (p = 0.0004).(See additional material Code S4).

The above solution is not general, because the *lmer()* function does not support other variance-covariance structures, apart from CS. A more general solution is to use the commercial package *asreml-R,* which works with all the main variance-covariance structures; an example is given as additional material (Code S4).

Alternatively, if a freeware solution is needed, it is necessary to consider that Wald-type F tests with Kenward-Roger adjusted degrees of freedom are not available for the function *lme()*in the *nlme* package. Therefore, the significance of the 'rotation × year' interaction should be tested by using a LRT, which consists of comparing a full model with a reduced model (dropping the 'ROT · YEAR' interaction term) in a fairly similar fashion as with the 'extra-sum-of-squares' test in traditional linear/nonlinear models.

In detail, both models (full and reduced) were fitted by full maximum likelihood (ML; REML estimation is not valid, in this particular case) and twice the log-likelihood (LOGLIK) for the full model was compared with twice the log-likelihood of the reduced model:

$$LRT = 2[LOGLIK_{full} - LOGLIK_{reduced}]$$

For large samples, the LRT statistic is assumed to have a chi-squared distribution with a number of degrees of freedom equal to the difference in the number of parameters for the two models. In this case we obtained a value of 274.99, with 116 degrees of freedom (the product of the number of years - 1 by the number of rotations - 1, i.e. 29 × 4 = 116), which corresponds to a very low p-level ($p = 6.2 \times 10^{-15}$).

This very small value leaves us quite confident about rejecting the null hypothesis. However, a parametric bootstrap approach may be used whenever one is in doubt. In this case, the procedure is as follows (Faraway, 2006):

1. simulate a new dataset under the reduced model, using the fitted parameter estimates and assuming normality for the errors and random effects;
2. fit to this dataset both the full and the reduced model;
3. compute the LRT statistic;
4. repeat steps 1 to 3 many times (e.g. 1000);
5. examine the distribution of the bootstrapped LRTs and compute the proportion of those exceeding 274.99 (empirical p-value)

An example of calculation in R is given in the supplementary material (Code S4), showing an empirical p-value of 0.000145, still significant, but much higher than the observed one.The bootstrap approach is reliable, but it may be tedious and very computer-intensive.

8. 'Fertility' trends

In contrast to other multi-environment experiments, LTEs are characterised by the same cropping practices, which are repeated on the same plots for a long period of time. This may bring to a significant 'fertility' trend over time (Berzsenyi et al., 2000) that may be reinforced by temporal changes due to the adoption of innovative cropping practices (e.g. new varieties, harvesting methods, weed control methods), as necessary for experiments lasting for a long period of time (McRae and Ryan, 1996).

Long-term 'fertility' trends are not of random nature and relate to the sustainability of each cropping system under investigation: a system with an upward trend should be more sustainable than another one with a downward trend in the long run.

The most straightforward way to model fertility trends is by fitting a simple linear regression of yield over time, as shown by Guertal et al. (1994) or Piepho et al. (2014). Nonlinear relationships might also be used, in the form of polynomials (Payne, 2015) or other intrinsically non-linear functions (Yang, 2014), whenever such a complexity is supported by the data at hand.

8.1. Analysis of Dataset 4 (continued): fitting a 'fertility' trend

For this case study we used Dataset 4 (Tab. 2), relating to the same rotation (M-W-W) with or without burial of crop residues. It might be expected that the continuous incorporation of soil residues creates a build-up of soil fertility over time.

Model 4 was used, together with a simple compound symmetry variance-covariance structure. The fixed part of the model was modified by expressing wheat yield as a linear function of a centred continuous covariate for the calendar year from the beginning of experiment (TIME_C), with treatment-dependent intercept and slope (i.e. four different average yield levels, depending on the combined levels of P and RES and two different slopes, depending on RES). All the other effects (YEAR, RES · YEAR, P · YEAR, RES · P · YEAR and all effects containing the BLOCK factor) were included as random, to represent the deviations from the regression lines. Therefore, the model was as follows:

YIELD ~ RES + P + RES · P + RES/TIME_C

RANDOM: YEAR + RES · YEAR + P · YEAR + RES · P · YEAR + BLOCK + BLOCK · YEAR + BLOCK · MAIN + BLOCK · MAIN · SUB + BLOCK · MAIN · YEAR + BLOCK · MAIN · SUB · YEAR

(Model 6)

It is useful to note that the above model partitions the year effect in two components: the fixed part relating to the fertility trend over time and the random remaining part, modelling the year-by-year variation around the trend line, as shown by Loughin et al. (2007).

The fitted fertility trends are reported in Figure 4, together with the observed yearly mean yields for the two cropping systems (with and without burial of crop residues) and phases (wheat after maize and wheat after wheat). Within managements, wheat after wheat yielded on average less than wheat after maize. The system was sustainable regardless of the burial/removal of crop residues, thanks to an upward yield trend over time, though, the slope was higher with burial (0.043; SE = 0.021) than with removal (0.031; SE = 0.021). There is, however, a remarkable random variability around these trends, as expected with this type of experiments (Laidig et al., 2014).

[Figure 4 about here]

We should point out that we use the term 'fertility' in its widest sense and the above model cannot in general separate the possible reasons for such a trend, which may relate to either possible long-term effects of agronomic practices, or technical innovation, or climate changes (Piepho et al., 2014). The R code to reproduce the analyses is given as supplemental material (Code S5)

9. Overall productivity and yield stability of cropping systems

Often, there is an interest in assessing long-term average yield levels for the cropping systems under study (broad inference space), together with their ability to minimise random fluctuations relating to the year-to-year variability in weather conditions and other biotic/abiotic factors (e.g. pests). This latter ability has been related to the concept of stability and, for LTEs, it would seem particularly appropriate to assess the 'dynamic' (Type II) stability, i.e. the ability of cropping systems to follow the mean environmental response of all cropping systems in the experiment (Piepho, 1998).

Within this frame, it would seem totally appropriate to regard the 'treatment' effect as fixed, together with the possible fertility trend across years, while the remaining part of the year effect and the year-by-treatments interactions should be regarded as random. In this case, the long duration of LTEs makes the estimation of variance components rather reliable.

Piepho (1999) has shown that an appropriate specification of the variance-covariance matrix for mixed models may be used to obtain a reliable description of stability. For example, the variance component for the random 'treatment×year' interaction might be assumed to depend on the treatment levels, which would measure their long-term 'stability variance' (Shukla, 1972). We will show this with another example.

9.1. Analysis of Dataset 3 (continued): a stability variance model

For this case study we used Dataset 3, where the factorial combination of three tillage methods and three N-fertilisation levels was evaluated on a 19-year LTE, based on a biennial rotation 'winter wheat/spring crop' (Tab. 2). Winter wheat yields were analysed to assess the long-term average of the nine cropping systems, together with their estimates of Shukla's stability variance.

Preliminary analyses showed that this dataset is characterised by a strong year-to-year heterogeneity of variances, while the existence of within-plot serial correlation is not supported by the data. Therefore, we fitted Model 3, incorporating a fertility trend (as in the previous case study) and allowed for 9 different variance components for the random 'treatment×year' interaction (where treatment is the factorial combination of tillage and N-fertilisation levels).

[Table 6 about here]

Results are reported in Table 6and clearly depict the long-term average behaviour of cropping systems, in terms of yield level, stability and sustainability. The smallest stability variance was observed for wheat grain yield under sod seeding with no N fertilization and under the other two tillage systems with 90 kg N ha^{-1}. Moreover, both the minimum tillage and sod seeding with the highest N fertilization seemed to lead to an upward fertility trend, which makes these cropping systems more sustainable than the other treatments under comparison.

The R code to reproduce the analyses is reported as supplemental material (Code S6)

10. Concluding remarks

In order to produce reliable results, the process of data analysis for LTEs with cropping systems requires some effort and careful thinking, especially for the most complex designs, characterised by rotations with different durations and a different number of test-crops per rotation cycle. Complex LTEs have been very seldom considered in literature, which motivated the example-based approach taken in this manuscript. We proposed several case-studies based on five exemplary datasets, covering the main types of LTEs, with increasing levels of complexity. Obviously, we could not cover the whole array of techniques that might be successfully used for the analysis of LTEs; for example, further research will be necessary to elucidate whether spatial models might be useful to account for possible spatial trends. For example, spatial correlation between plots errors may be combined with temporal correlation among plot errors in a single model, and there are many options for doing this.

Based on the analyses of our datasets, we would like to highlight the following issues.

1. With experiments involving rotations, where all crop phases are grown in every year and within-year replicates are available, statistical models for data analysis can be reliably built using the procedure outlined in Piepho et al. (2004), providing that correct analyses can be performed separately for each year.
2. In this paper, we propose that the year is included in the model as the repeated factor, by which all the main types of LTEs are put on an equal footing, together with multi-environment (genotype) experiments. However, particularly when we are interested only in a subset of the crop phases for each rotation, the resulting models are rarely orthogonal, as observations for the test crop of interest are not taken all years on all plots. For this reason, traditional analysis of variance is not generally appropriate, while mixed models and REML estimation should be preferred.
3. One of the main problems with LTEs is that observations are taken repeatedly on the same plot and, therefore, they are correlated. As the first step, we can account for this by including in the model random effects for randomisation units (e.g. plots). Such an approach assumes that within-plot correlation is independent on the time elapsed between two observations (the model is known as compound symmetry, also known as 'split-plot in

time'), which is rarely sufficient to appropriately describe the real pattern of variances-covariances of observations. Therefore, it is fundamental that this aspect is considered during data analysis.

4. For all our examples, we found that variances were not constant across the repeated factor, which we accounted for by allowing a different variance for each year (compound symmetry with variance heterogeneity of years). This variance-covariance structure proved to be the best compromise among simplicity and statistical accuracy, though this should not be taken as a general rule. Indeed, it is always appropriate to consider the need of other more complex correlation structures.

5. For those who are accustomed to traditional ANOVA, testing for fixed effects with mixed models and unbalanced data may require some further attention. Conditional F tests with Kenward-Roger adjusted denominator degrees of freedom may represent a good choice andseveral simulation studies have shown that they work well in practice (Richter et al., 2015; Spilke et al., 2005).If adjusted denominator degrees of freedom are not available, parametric bootstrap procedures can be used instead of simple likelihood ratio tests, as these latter tend to be anti-conservative and give unreliable p-levels.

6. Fertility trends (linear/non-linear) during LTEs may be investigated to measure the long-term sustainability of cropping systems. These trend are always of a fixed nature and may serve the purpose of partitioning the 'year' effect into its random and fixed components, as recommended by Loughin et al. (2007).

7. Several statistical methods devised for multi-environment experiments may be successfully used for the analyses of LTEs. For example, in this paper we used Shukla's stability variance model to assess the long-term stability of cropping systems.

Bibliography

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software67, 1, 1-48.

Berzsenyi, Z., Gyorffy, B., Lap, D., 2000. Effect of crop rotation and fertilisation on maize and wheat yields and yield stability in a long-term experiment. European Journal of Agronomy 13, 225–244.

Bonciarelli, U., Onofri, A., Benincasa P., Farneselli, M., Guiducci, M., Pannacci, E., Tosti, G., Tei, F., 2015. Long-term evaluation of productivity, stability and sustainability for cropping systems in Mediterranean rainfed conditions. European Journal of Agronomy, this issue.

Brien, C., Demetrio, C., 2009. Formulating mixed models for experiments, including longitudinal experiments. Journal of Agricultural, Biological and Environmental Statistics 14, 253–280.

Burnham, K.P., Anderson, D.R. 2002. Model selection and multimodel inference. A practical information-theoretic approach. Second edition. Springer, New York.

Butler, D., Cullis, B., Gilmour, A., Gogel, B., 2009. ASReml-r reference manual: mixed models for S language environments.Queensland Government, Department of Primary Industries and Fisheries, Brisbane, Australia.

Cochran, W.G., 1939. Long-term agricultural experiments. Supplement to the Journal of the Royal Statistical Society 6, 104–148.

Eckl, T., Piepho, H.-P., 2015. Analysis of series of variety trials with perennial grasses for subdivided target regions. Crop Science 55, 597–609.

Faraway, J., 2006. Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models. Chapman & Hall, Boca Raton, Florida.

Frye, W., Thomas, G., 1991. Management of long-term experiments. Agronomy Journal 83, 38–44.

Gałecki, A., Burzykowski, T., 2013. Linear mixed-effects models using R: a step-by-step approach, Springer, Berlin.

Guertal, E., WR Raun, R.W., Boman, R., 1994. Application of stability analysis for single-site, long-term experiments. Agronomy Journal 86, 1016–1019.

Halekoh, U., Højsgaard, S., 2014. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. Journal of Statistical Software, 59, 1-30.

Kenward, M.G., Roger, J.H., 1997. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood, Biometrics, 53, 983–997.

Kenward, M.G., Roger, J.H., 2009. An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood. Computational Statistics and Data Analysis, 53, 2583–2595.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2015. LmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-29. http://CRAN.R-project.org/package=lmerTest

Laidig, F., Piepho, H.-P., Drobek, T., Meyer, U., 2014. Genetic and non-genetic long-term trends of 12 different crops in German official variety performance trials and on-farm yield trends. Theoretical and Applied Genetics 127, 2599–2617.

Loughin, T.M., Roediger, M.P., Milliken, G.A., Schmidt, J.P., 2007. On the analysis of long-term experiments. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170, 29–42.

McRae, K., Ryan, D., 1996. Design and planning of long-term experiments. Canadian Journal of Plant Science 76, 595–602.

Patterson, H.D., 1964. Theory of cyclic rotation experiments. Journal of the Royal Statistical Society. Series B (Methodological) 26, 1–45.

Payne, R.W., 2015. The design and analysis of long-term rotation experiments. Agronomy Journal 107, 772–785.

Perucci, P., Bonciarelli, U., Santilocchi, R., Bianchi, A.A., 1997. Effect of rotation, nitrogen fertilization and management of crop residues on some chemical and biochemical characteristics of soil. Biology and Fertility of Soils 24, 311–316.

Piepho, H.-P., 1998. Methods for comparing the yield stability of cropping systems - a review. Journal of Agronomy and Crop Science 180, 193–213.

Piepho, H.-P., 1999. Stability analysis using the SAS system. Agronomy Journal 91, 154–160.

Piepho, H.-P., Büchse, A., Emrich, K., 2003. A hitchhiker's guide to mixed models for randomized experiments. Journal of Agronomy and Crop Science 189, 310–322.

Piepho, H.-P., Büchse, A., Richter, C., 2004. A mixed modelling approach for randomized experiments with repeated measures. Journal of Agronomy and Crop Science 190, 230–247.

Piepho, H.-P., Eckl, T., 2014. Analysis of series of variety trials with perennial grasses. Grass and Forage Science 69, 431–440.

Piepho, H.-P., Laidig, F., Drobek, T., Meyer, U., 2014. Dissecting genetic and non-genetic sources of long-term yield trend in german official variety trials. Theoretical and Applied Genetics 127, 1009–1018.

Pinheiro, J., Bates, D., 2000. Mixed-Effects models in s and S-Plus. Springer-Verlag Inc., New York.

R Core Team, 2014. R: A language and environment for statistical computing, ed. R Foundation for Statistical Computing, Vienna, Austria.

Richter, C., Kroschewski, B., 2006. Analysis of a long-term experiment with repeated-measurement models. Journal of Agronomy and Crop Science 192, 55–71.

Richter, C., Kroschewski, B., Piepho, H.-P., Spilke, J., 2015. Treatment comparisons in agricultural field trials accounting for spatial correlation. The Journal of Agricultural Science 153,1187-1207.

Searle, S.R., Casella, G., McCulloch, C.E., 2006. Variance components. John Wiley and Sons, Inc., Hoboken, New Jersey.

Shukla, G., 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29, 237–245.

Singh, M., Jones, M.J., 2002. Modeling yield sustainability for different rotations in long-term barley trials. Journal of Agricultural, Biological, and Environmental Statistics 7, 525-535.

Smith, A., Stringer, J., Wei, X., Cullis, B., 2007. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. Euphytica 157, 253–266.

Spilke, J., Piepho, H.-P., Hu, X., 2005. A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. Journal of Agricultural, Biological, and Environmental Statistics 10, 374–389.

Yang, R.-C., 2014. Analysis of linear and non-linear genotype × environment interaction. Frontiers in Genetics 5, 227.

Yates, F., 1954. The analysis of experiments containing different crop rotations. Biometrics 10, 324-346.

TABLES

Table 1. Glossary of the terms used in this paper (following Yates, 1954).

Term	Meaning
Rotation	A set of crops that are grown in successive years on the same plot (e.g. M-W-W)
Phase	Each crop component in a rotation (e.g. in the rotation M-W-W, M is phase 1, W is phase 2 and W is phase 3). In a LTE we may be interested in all phases, or (more frequently) in a specific phase (test-crop). Rotations may have the same crop in two different phases (phase difference between W1 and W2). The number of phases defines the period (duration) of each rotation
Cycle	Each single repetition of the whole rotation
Sequence	Each of the n possible arrangements for a rotation of n years, having the same crop ordering, but different initial phases (e.g M-W1-W2, W1-W2-M and W2-M-W1). Each sequences is uniquely identified by its starting phase, which needs to be randomised to each plot at the start of the experiment, according to the basic requirement that all phases of each rotation/system are grown every year. Likewise, if we look at a specific test-crop, the sequence is uniquely identified by its position in the sequence. It is also relevant to note that each sequence shows the test crop in a different position (e.g. the first sequence shows M in first position, the second shows M in third position and the third shows M in second position)

Table 2: Summary of experimental designs for the five datasets

Design	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Treatments	Three N-fertilisation levels on continuous wheat	Comparison among five two-year rotations based on wheat	Nine combination of soil tillage and N-fertilisation levels on a two-year sunflower-wheat rotation	Two types of management of crop residues on a three-year rotation maize-wheat-wheat	Five wheat (W) – Maize (M) rotations (MW – MWW – MWWW – MWWWW and MWWWWW)
Response	Wheat yield	Wheat yield	Wheat yield	Wheat yield	Wheat yield
Design	RCBD	Split-plot (main-plots: sequences; sub-plots: rotations)	Split-plot (main-plots: tillage; sub-plots: N-levels)	Split-plot (main-plots: crop residues; sub-plots: phases)	RCBD
Years	30	30	18	30	30
Blocks	3	3	2	3	3
Plots	9	30	36	18	60

N. of records per year	9	15	18	12	45
Total data	270	450	360	360	1350

Table 3. Experimental design for Dataset 2, in one block and four years (two rotation cycles). W: wheat; P: Pea; GS: grain sorghum; SB: sugar beet; SU: sunflower; FB: faba bean.

Plot No.	Rotation	Sequence	Year			
			1983	1984	1985	1986
4	3	1	W	SB	W	SB
6	1	1	W	P	W	P
12	5	1	W	FB	W	FB
13	2	1	W	GS	W	GS
14	4	1	W	SU	W	SU
20	5	2	FB	W	FB	W
21	2	2	GS	W	GS	W
22	4	2	SU	W	SU	W
27	3	2	SB	W	SB	W
29	1	2	P	W	P	W

Table 4. Experimental design for Dataset 4, in one block and three years (one rotation cycles). Three phases can be identified: W1 (wheat after maize), W2(wheat after wheat) and M (maize).

Plot No.	Sequence	Crop Residues	Year		
			1983	1984	1985
1	3	Buried	W2	M	W1
2	1	Buried	M	W1	W2
3	2	Buried	W1	W2	M
4	1	Removed	M	W1	W2
5	2	Removed	W1	W2	M
6	3	Removed	W2	M	W1

Table 5. Selection among several patterns of within-plot variance-covariance structures, following mixed model fits to Dataset 1 using REML. NC: no within-plot correlation; CS: compound symmetry; CSH: compound symmetry with heteroscedastic errors by year; AR(1): autoregressive of order 1; ARH(1): autoregressive of order 1, with heteroscedastic errors by year.

Model	df	AIC[1]	log-likelihood
NC	151	619.97	-158.98
CS	152	612.39	-154.19
AR(1)	152	619.52	-157.76
CSH	181	604.56	-121.28
ARH(1)	181	606.37	-122.18

[1]: Akaike Information Criterion (Akaike, 1974)

Table 6. Mean yield level, stability standard deviation (SD) and fertility trend for wheat, grown with nine different combinations of tillage and N-fertilisation levels (T: conventional ploughing; M: scarification; S: sod seeding), at the end of a 19-year LTE (Dataset 3).

Tillage	kg N ha^{-1}	Overall mean ($kg\ ha^{-1}$)	SEM	Stability SD ($kg\ ha^{-1}$)	Slope	SE
M	0	1488	213.4	402.0	12.0	37.95
M	90	3022	189.1	30.3	6.6	32.90
M	180	3807	214.8	414.2	42.5	38.23
S	0	1471	189.0	24.4	13.7	32.89
S	90	2924	211.9	387.4	0.3	37.63
S	180	3554	223.7	489.3	43.7	40.06
T	0	1409	218.3	445.1	-14.7	38.96
T	90	3082	193.9	169.3	-38.9	33.91
T	180	4111	278.3	851.1	16.0	51.14

FIGURES

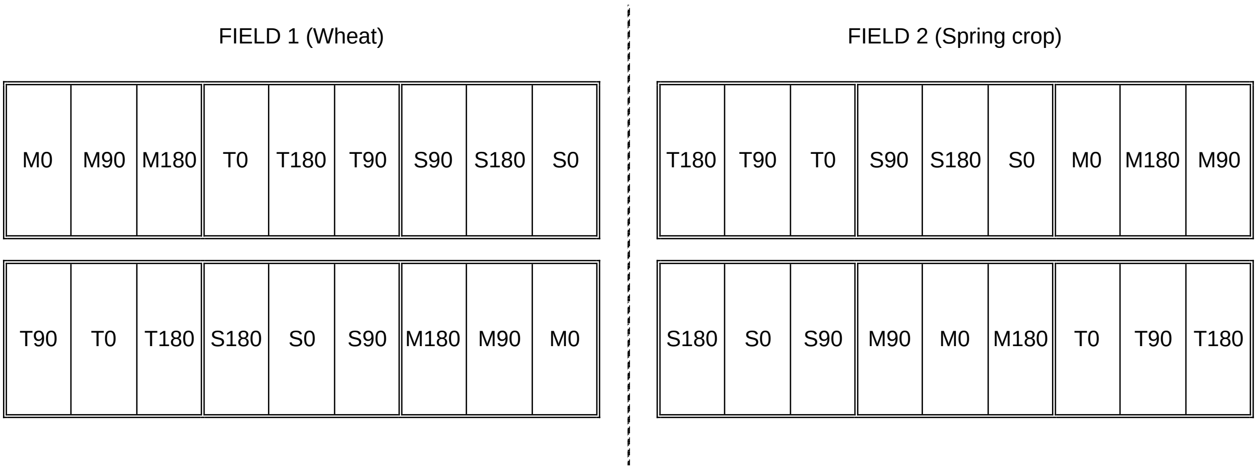

Figure 1. Experimental design for Dataset 3 in one year. The position of wheat and spring crop is exchanged in the following year (T: conventional ploughing; M: scarification; S: sod seeding; 0, 90 and 180 kg N ha^{-1})

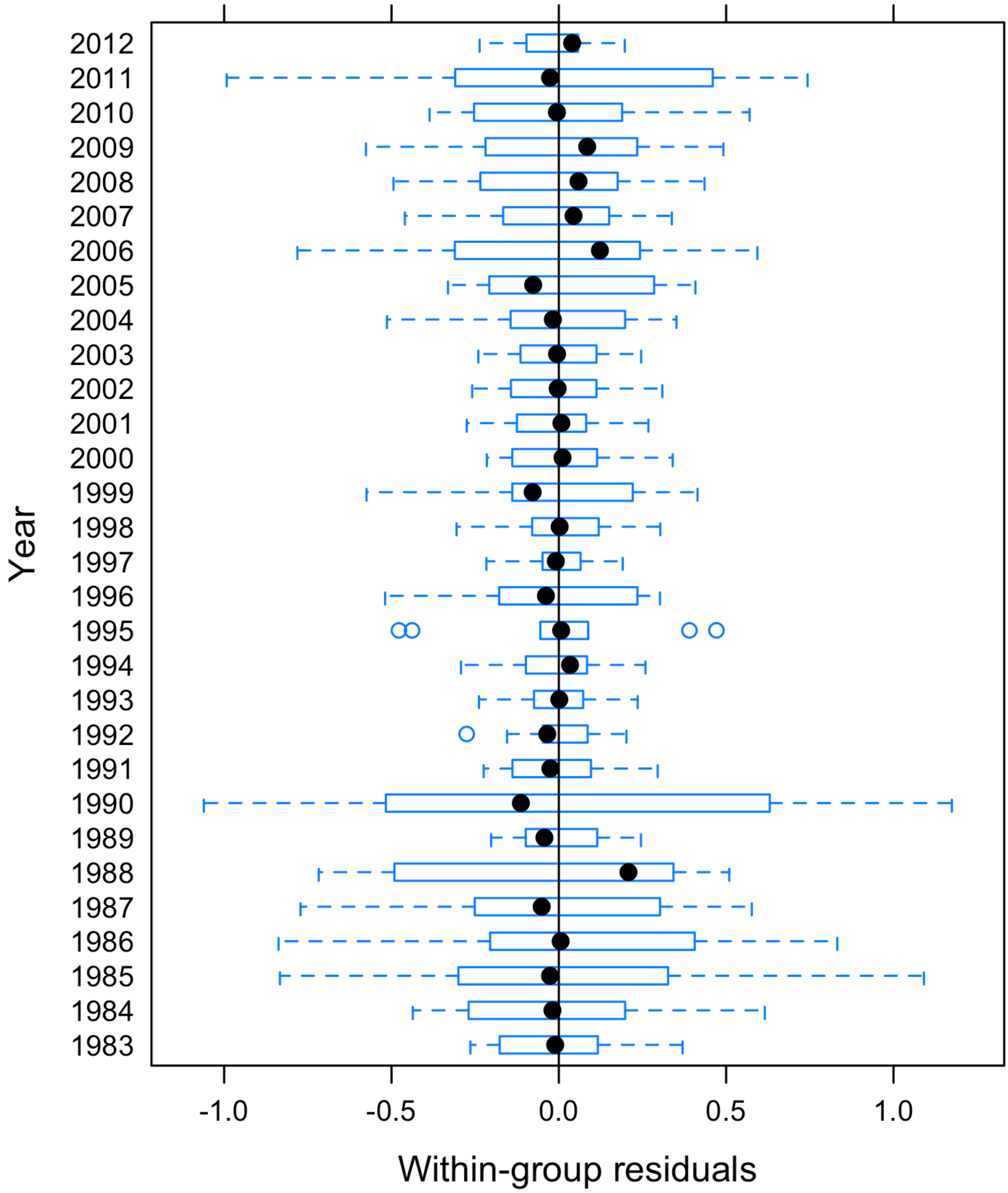

Figure 2. Boxplot of within-group residuals by year, after a mixed model fit to Dataset 1

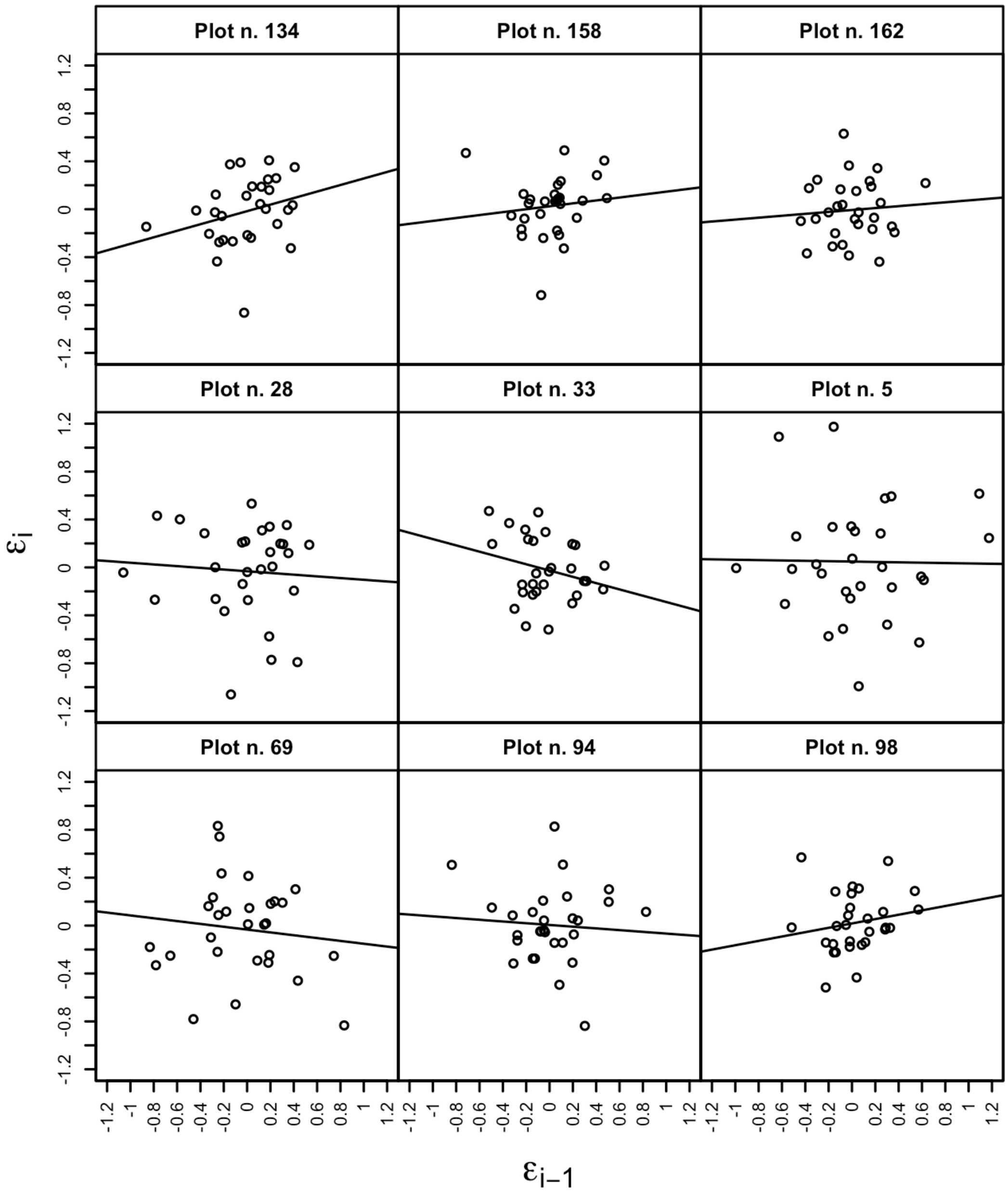

Figure 3. Lag-one residual plot, after a mixed model fit to Dataset 1.

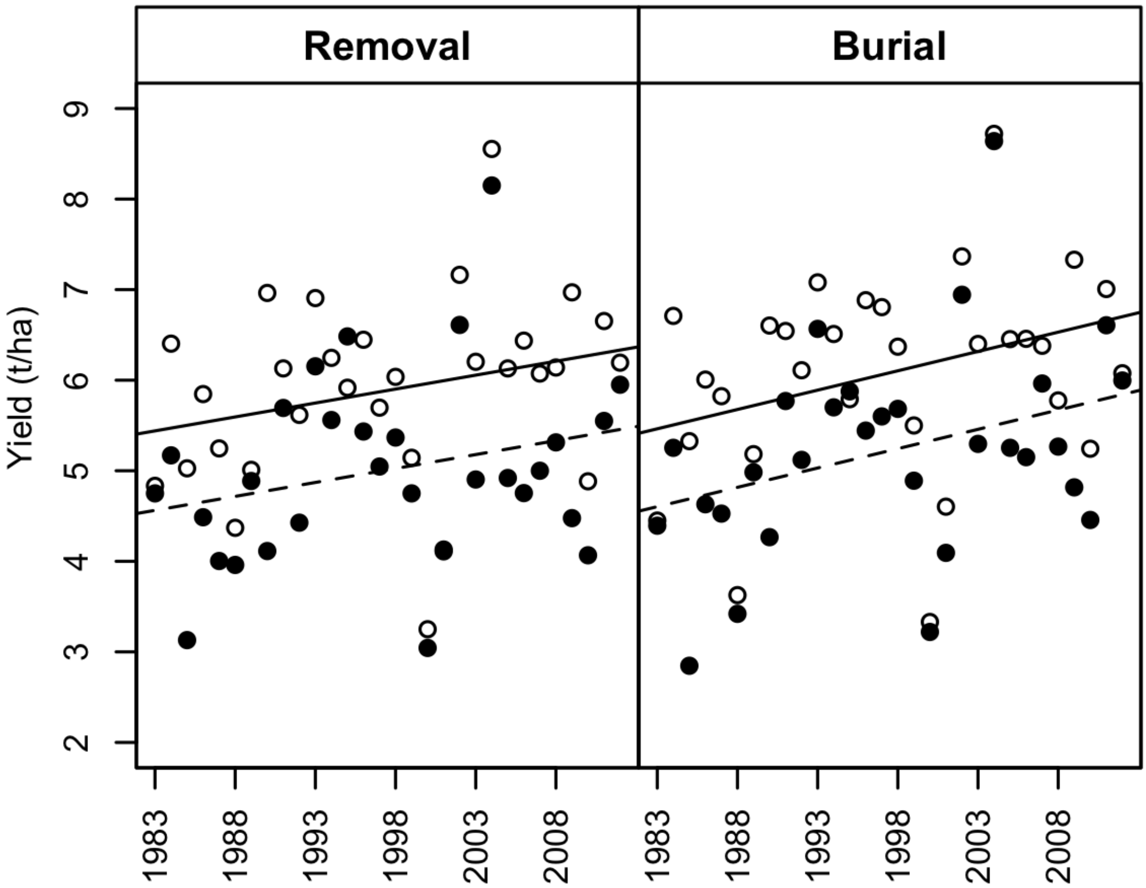

Figure 4. Fertility trends for wheat in a three-year rotation (Maize-Wheat-Wheat; Dataset 4) with either removal or burial of crop residues. For both management types, closed circles and dotted lines represent the wheat following wheat (phase 3) and open circles and solid lines represent wheat after maize (phase 2). Symbols show observed data, while lines show fitted trends.

Supplemental material

1. The two real LTEs

Two long-term field experiments are used. The first one (Exp. 1) was established around 1974 at Perugia (central Italy), at the Experimental Farm of the Department of Agricultural, Food and Environmental Sciences (Papiano, 42.9569N, 12.3767E, 196 m a.s.l.) and it is still running at present (2015). The aim is to compare different cropping systems based on winter wheat (*Triticum aestivum* L.) and resulting from the factorial combination of two different types of management of crop residues and 13 types of crop rotation and mineral N fertilisation rates. In particular, crop residues are either (i) removed, or (ii) incorporated into the soil at ploughing, while the 13 rotations/fertilisations include (i - iii) three continuous winter wheat (W) systems, differing in N fertilization rates, i.e. 150 (WN150, 'standard' rate), 200 (WN200) and 250 (WN250) kg ha^{-1}, split into 50% at tillering and 50% at shooting; (iv - viii) five types of two-year rotations, wherein wheat is rotated with either pea (*Pisum sativum* L), grain sorghum (*Sorghum bicolor* (L.) Moench), sugar beet (*Beta vulgaris* L. subsp. *saccharifera*), sunflower (*Helianthus annuus* L.) or faba bean (*Vicia faba* L. subsp. *minor*). (ix - xiii) five maize (M) - wheat (W) rotations of different lengths, i.e. MW, MWW, MWWW, MWWWW, MWWWWW. For all rotations, all phases are contemporarily present in each year, for a total of 66 plots (33 plots for each of 33 possible crop sequences by two residue management levels) in each of three blocks (198 plots in total), arranged according to a split-plot design, with crop residue management on the main plots and crop rotations/fertilisation levels on the subplots (24 m^2 each). Several measurements are taken on variable intervals, though we will only consider yearly yield measurements on each plot. Further details about this experiment can be found elsewhere (Bonciarelli et al., 2016; Perucci et al., 1997).

The second experiment (Exp. 2) was established in 1994 at the Pasquale Rosati experimental farm in Agugliano, Italy (43.32N, 13.22E, 100 m a.s.l.), located on a 10% slope. It is based on a two-year rotation with durum wheat (*Triticum durum* L.) followed either by sunflower (until 2001) or maize (from 2002 onwards). The aim was to compare several cropping systems consisting of a factorial combination of different soil tillage and N-fertilisation practices. The tillage treatments were: (T: conventional 40 cm deep ploughing; M: scarification at 25 cm; S: sod seeding with chemical desiccation and chopping). The three N fertilizer treatments were 0, 90 and 180 kg N ha^{-1}, distributed in two rates for wheat and at seeding for maize/sunflower. The crop rotation was duplicated in two adjacent fields to allow for all crops to be present each year. Specifically, the two fields receive the two sequences, so that they have the two different crops of the rotation in the same year. Within fields, there are two independent randomisations, with two blocks, tillage levels randomised to main-plots (1500 m^2) and N levels randomised to sub-plots (500 m^2), according to a split-plot design with two replicates.

2. Structure of datasets

Datasets are composed by a number of records as shown in Table 2 and columns are named according to the definition of models 1 to 5. For illustration of the general structure, the first six observations are listed here for each dataset.

Dataset 1

```
  Plot     N Year Block Yield
1   33 fn150 1983     1  4.54
2   98 fn150 1983     3  4.18
3  162 fn150 1983     2  3.70
4   33 fn150 1984     1  4.57
5   98 fn150 1984     3  5.04
6  162 fn150 1984     2  5.06
.....
.....
```

Dataset 2

```
  Block Plot Rot Year Yield
1     1    4 SBW 1983  5.10
2     3   70 SBW 1983  4.50
3     2  135 SBW 1983  4.53
4     1   27 SBW 1984  5.83
5     3   95 SBW 1984  6.26
6     2  160 SBW 1984  6.22
.....
.....
```

Dataset 3

```
  Year Field Cycle  Main Sub Block T   N Yield
1 1995     1     1 1_1_M   1     1 M   0   855
2 1995     1     1 1_1_M   2     1 M  90  2256
3 1995     1     1 1_1_M   3     1 M 180  2513
4 1995     1     1 1_1_S   4     1 S  90  2302
5 1995     1     1 1_1_S   5     1 S   0  1006
6 1995     1     1 1_1_S   6     1 S 180  3125
....
....
```

Dataset 4

```
  Block Sub Res Year P Yield
1     1  15 asp 1983 2  5.05
2     1  23 asp 1983 1  5.08
3     2  43 int 1983 2  4.00
```

```
4     2  57 int 1983 1  4.64
5     3  81 asp 1983 2  4.79
6     3  89 asp 1983 1  5.00
.....
.....
```

Dataset 5

```
  Block Plot Rot Year P Yield
1     1   16 f1m 1983 1  5.36
2     3   82 f1m 1983 1  5.20
3     2  148 f1m 1983 1  4.99
4     1    8 f1m 1984 1  6.46
5     3   75 f1m 1984 1  6.23
6     2  147 f1m 1984 1  6.35
.....
.....
```

3. R Code for all case studies

The following code has been used to perform the analysis with the R software (R Core Team, 2014) on the above mentioned datasets.

```
##########################################################################
#Code S1. Examples of implementation in R, by way of mixed models and REML
 estimation

#Load data
rm(list=ls())
dataset <- read.csv(file="Dataset2.csv", header=T)
head(dataset)

#Make all numeric variables as factors
dataset$Block <- factor(dataset$Block)
dataset$Plot <- factor(dataset$Plot)
dataset$YearF <- factor(dataset$Year)
dataset$Rot <- factor(dataset$Rot)

#Implementation with R and lmer()
#Random effect for main-plots was disregarded
library(lme4)
mod1 <- lmer(Yield ~ Block + Rot*YearF + Block:YearF +
               (1|Plot),
             data=dataset)

#Implementation with R and lme()
library(nlme)
mod2 <- lme(Yield ~ Block + Rot*YearF + Block:YearF,
            random = ~ 1|Plot,
```

```
            data=dataset)

#Implementation with R and gls()
mod3 <- gls(Yield ~ Block + Rot*YearF + Block:YearF,
            correlation=corCompSymm(form=~1|Plot),
            data=dataset)

#Implementation with ASRemL-R
library(asreml)
mod4 <- asreml(fixed = Yield ~ Block + Rot*YearF + Block:YearF,
               random = ~ Plot, data = dataset)


###################################################################
### Code S2: Checking models

#Load data
rm(list=ls())
dataset <- read.csv(file="Dataset1.csv", header=T)
head(dataset)

#Make all numeric variables as factors
dataset$Block <- factor(dataset$Block)
dataset$Plot <- factor(dataset$Plot)
dataset$YearF <- factor(dataset$Year)
dataset$N <- factor(dataset$N)

#Cumulative analyses with lme()
library(nlme)
mod <- lme(Yield ~ Block + N +
             YearF + Block:YearF + YearF:N,
           random=~1|Plot,
           data=dataset)

plot(mod) #Plot of residuals vs fitted values
qqnorm(mod, ~ resid(.)) #QQ plot of residuals

plot(mod, Plot ~ resid(.), abline=0) #Residuals by plot
#Residuals by year
plot(mod, YearF ~ resid(.), abline=0, xlab="Within-group residuals", ylab="Year")

#Other diagnostic plots
plot(mod, resid(.) ~ fitted(.)|Plot, abline=0)
plot(modMix, N ~ resid(.), abline=0)
plot(modMix, Block ~ resid(.), abline=0)

#Lag-plot 1
library(reshape)
res <- data.frame(plot=names(residuals(mod)), Year = dataset$YearF, residui=residu
als(mod))
res <- res[order(res$Year), ]
mat <- cast(Year ~ plot, data = res, value = "residui", fun=mean)
mat2 <- rbind(mat[2:30,], rep(NA,30))
par(las=3, omi=c(1,0.7,0.5,0.5), mfrow=c(3,3))
for(i in 1:9){
  par(mai=c(0,0,0.4,0))
  lab <- paste("Plot n.", dimnames(mat)[[2]][i+1], sep=" ")
  plot(0, type="n", xlim=c(-1.2,1.2), ylim=c(-1.2,1.2), axes=F, main=lab)
  points(mat[,i+1] ~ mat2[,i+1])
```

```
  abline(lm(mat[,i+1] ~ mat2[,i+1]))
  box(which = "plot")
  box(which = "figure")
  if(i>6) axis(1, at=seq(-1.2,1.2,by=0.2), labels=round(seq(-1.2,1.2,by=0.2), 2))
  if(i==1 | i==4 | i == 7) axis(2,at=seq(-1.2, 1.2, by=0.2), labels=round(seq(-1.
2,1.2,by=0.2), 2))
}
expression(varepsilon [i])
mtext(expression(epsilon [i - 1]), las=1, side=1, outer=T, at=c(0.5), padj=3.5, ce
x=1.0)
mtext(expression(epsilon [i]), outer=T, side=2, las=0, at=c(0.5), padj=-2.3, cex=
1.0)

###################################################################
### Code S3: Implementing within-plot correlation structures by using gls()
#Load data
rm(list=ls())
dataset <- read.csv(file="Dataset1.csv", header=T)
head(dataset)

#Make all numeric variables as factors
dataset$Block <- factor(dataset$Block)
dataset$Plot <- factor(dataset$Plot)
dataset$YearF <- factor(dataset$Year)
dataset$N <- factor(dataset$N)

#CS: compound symmetry
modCS <- gls(Yield ~ Block + N +
               YearF + Block:YearF + YearF:N,
             correlation=corCompSymm(form=~1|Plot),
             data=dataset)

#CSH: Compound symmetry with heteroscedastic errors by year
modCSH <- gls(Yield ~ Block + N +
                YearF + Block:YearF + YearF:N,
              correlation=corCompSymm(form=~1|Plot),
              weights=varIdent(form=~1|YearF),
              data=dataset)

#AR: autoregressive
modAR <- gls(Yield ~ Block + N +
               YearF + Block:YearF + YearF:N,
             correlation=corAR1(form=~1|Plot),
             data=dataset)

#ARH: autoregressive with heteroscedastic errors by year
modARH <- gls(Yield ~ Block + N +
                YearF + Block:YearF + YearF:N,
              correlation=corAR1(form=~1|Plot),
              weights=varIdent(form=~1|YearF),
              control=list(opt="optim", numIter=3000),
              data=dataset)

#UN: unstructured (does not converge!)
modMix4Bis <- gls(Yield ~ Block + N +
                    YearF + Block:YearF + YearF:N,
                  correlation=corSymm(form=~1|Plot),
                  weights=varIdent(form=~1|YearF),
                  control=list(opt="optim", numIter=3000),
```

```
                  data=dataset)

AIC(modCS, modCSH, modAR, modARH)

#################################################################
#Code S4. Testing for fixed effects

rm(list=ls())
dataset <- read.csv(file="Dataset2.csv", header=T)

#Making all numerical variables as factors
dataset$Block <- factor(dataset$Block)
dataset$Plot <- factor(dataset$Plot)
dataset$YearF <- factor(dataset$Year)
dataset$Rot <- factor(dataset$Rot)

#Kenward-Roger approximation (by using the lmerTest)
library(lmerTest)
library(pbkrtest)
modlF2 <- lmer(Yield ~ Block + Rot*YearF + Block:YearF +
                 (1|Plot),
               data=dataset)
anova(modlF2, ddf = "lme4")
anova(modlF2, ddf = "Kenward-Roger")


#Kenward-Roger approximation (by using the asreml-R package)
library(asreml)
modFA <- asreml(fixed = Yield ~ Block + Rot*YearF + Block:YearF,
                random = ~ Plot, data = dataset)
wald(modFA, denDF="default", ssType="conditional")



#Fitting a full and reduced mixed model by lme()
#Models are fit by maximum likelihood and a LRT test is performed
#by the function anova()
#Random effect for main-plots was disregarded
library(nlme)
modlF <- lme(Yield ~ Block + Rot*YearF + Block:YearF,
                random=~1|Plot,
              data=dataset, method="ML")
modlR <- lme(Yield ~ Block + Rot + YearF + Block:YearF,
             random=~1|Plot,
```

```
              data=dataset, method="ML")
anova(modlF, modlR)

#Parametric bootstrap (may be extremely slow!!!!!!!!!)
#This reproduces the procedure outlined in the paper, although
#the nsim argument may be used to avoid the 'for' cycle
nsim <- 100000
lrtSim <- numeric(nsim)
for (i in 1:nsim){
  y <- simulate(modlR, nsim=1, m2=modlF, method="ML")
  lrtSim <-  as.numeric(2*(y$alt$ML[,2] - y$null$ML[,2]))
  lrtSim
  print(i)
}
mean(lrtSim > 274.99)



#################################################################
#Code S5. Fertility trends

rm(list=ls())
dataset <- read.csv(file="Dataset4.csv", header=T)
head(dataset)

#Making all numerical variables as factors
dataset$Block <- factor(dataset$Block)
dataset$Main <- factor(dataset$Block:dataset$Res)
dataset$Sub <- factor(dataset$Sub)
dataset$YearF <- factor(dataset$Year)
dataset$P <- factor(dataset$P)

#Fitting a fertility trend
library(lme4)
dataset$YearC <- scale(dataset$Year, scale=F)
options(contrasts=c("contr.sum", "contr.poly"))
modLST <- lmer(Yield ~ I(Res:P) + Res/YearC +
                  (1|YearF) + (1|YearF:Res) + (1|YearF:P) + (1|YearF:Res:P) +
                 (1|YearF:Block) +
                 (1|Block/Main/Sub),
               data=dataset)
summary(modLST)


#################################################################
#Code S6. Stability variance model

rm(list=ls())
dataset <- read.csv(file="Dataset3.csv", header=T)
```

```
head(dataset)

#Making all numeric variables as factors
dataset$Field <- factor(dataset$Field)
dataset$Block <- factor(dataset$Block)
dataset$Sub <- factor(dataset$Sub)
dataset$FieldBlock <- factor(I(dataset$Block:dataset$Field))
dataset$FBM <- factor(dataset$FieldBlock:dataset$Main)
dataset$FBMS <- factor(dataset$FBM:dataset$Sub)
dataset$Till <- factor(dataset$T)
dataset$N <- factor(dataset$N)
dataset$Treat <- factor(dataset$Till:dataset$N)
dataset$YearF <- factor(dataset$Year)
dataset <- within(dataset, one1 <- one2 <- one3 <- one4 <- one5 <- 1L)
dataset$YearC <- scale(dataset$Year, scale=F)

options(contrasts=c("contr.sum", "contr.poly"))
mod <- lme(Yield ~ Treat - 1 + Treat/YearC,
           random = list(one1 = pdIdent(~Block - 1),
                         one3 = pdIdent(~FieldBlock - 1),
                         one4 = pdIdent(~FBM - 1),
                         one5 = pdIdent(~FBMS - 1),
                         YearF = pdBlocked(list(pdIdent(~1),
                                                pdDiag(~Treat - 1)))),
           weights=varIdent(form=~1|YearF),
           control=list(opt="optim", numIter=3000),
           data=dataset)

tTable <- summary(mod)$tTable
shuk <- as.numeric(VarCorr(mod)[61:69,2])
cbind(tTable[1:9,1:2], "shukla"=shuk, tTable[10:18,1:2])
```