CSIRO PUBLISHING

Crop & Pasture Science, 2015, **66**, 947–962
http://dx.doi.org/10.1071/CP14312

Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials

Joanne De Faveri[A,F], *Arūnas P. Verbyla*[B], *Wayne S. Pitchford*[C], *Shoba Venkatanagappa*[D], *and Brian R. Cullis*[E]

[A]Department of Agriculture and Fisheries, PO Box 1054, Mareeba, QLD, 4880, Australia.
[B]Data Analytics, CSIRO Digital Productivity Flagship and School of Agriculture, Food and Wine, The University of Adelaide,Atherton, QLD, 4883, Australia.
[C]School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy Campus, SA, 5371, Australia.
[D]NSW Department of Primary Industries, Tamworth, NSW, 2340, Australia. Current address: Enza Zaden Australia, Narromine, NSW, 2821, Australia.
[E]National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522, Australia.
[F]Corresponding author. Email: Joanne.DeFaveri@daf.qld.gov.au

Abstract. Variety selection in perennial pasture crops involves identifying best varieties from data collected from multiple harvest times in field trials. For accurate selection, the statistical methods for analysing such data need to account for the spatial and temporal correlation typically present. This paper provides an approach for analysing multi-harvest data from variety selection trials in which there may be a large number of harvest times. Methods are presented for modelling the variety by harvest effects while accounting for the spatial and temporal correlation between observations. These methods provide an improvement in model fit compared to separate analyses for each harvest, and provide insight into variety by harvest interactions. The approach is illustrated using two traits from a lucerne variety selection trial. The proposed method provides variety predictions allowing for the natural sources of variation and correlation in multi-harvest data.

Received 23 November 2014, accepted 28 April 2015, published online 4 September 2015

Introduction

Variety selection in perennial pasture crops is usually based on measurements taken at multiple harvest times from field trials which cover a potentially large and variable area. Critical to identifying the best varieties for selection and increasing the rate of genetic gain is the implementation of statistical methods that accurately predict the true potential of varieties (Smith and Spangenberg 2014). Statistical methods for analysing data from perennial pasture variety selection trials need to account for the spatial variation and correlation within a trial and the temporal correlation between repeated measurements. The methods also need to appropriately model the genetic effects over time.

Accounting for spatial variation in perennial pasture field trials is not a new concept. A number of papers have promoted the use of spatial analysis methods in pasture crops. For example, Casler (1999), Smith and Kearney (2002) and Smith and Casler (2004) used Nearest Neighbour methods (first introduced by Papadakis in 1937, see Bartlett (1978) for an account). In annual field crop evaluation trials, the methods of spatial analysis have been developed extensively since the early Nearest Neighbour methods, with advancements including the one dimensional models of Gleeson and Cullis (1987), where trend is modelled using time series models, and their extension to 2 dimensions by Cullis and Gleeson (1991), using a separable correlation structure. Gilmour *et al.* (1997) extended the method of Cullis and Gleeson (1991) by identifying three major components of spatial variation to be modelled, namely local and global smooth spatial trend and extraneous variation.

Gilmour *et al.* (1997) demonstrated that no one spatial model will be applicable to every trial and that models need to be formulated to incorporate the unique spatial trends and correlation that might be present at each individual trial. They presented a complete spatial modelling approach that incorporates diagnostic aids and tests for model selection. This is the approach that has been used in the analysis of cereal crop breeding trials across Australia for many years and has been shown to provide more efficient variety predictions than previous methods (see Gilmour *et al.* 1997; Cullis *et al.* 1998; Smith *et al.* 2001; Stefanova *et al.* 2009). While the application of this method in perennial pasture variety selection trials is not yet well documented, it has been implemented in other perennial crops such as sugarcane (Stringer and Cullis 2002; Smith *et al.* 2007); tea (Resende *et al.* 2006); and forestry trees (Dutkowski *et al.* 2002). The approach used in this paper will be based on these methods.

In perennial pasture variety selection trials, data is usually obtained from multiple harvests over a number of years. There is a need to account for temporal correlation in the residuals (due to the repeated measurements on each plot). This serial correlation often decreases with increasing intervals between harvests (Bjornsson 1978). Diggle (1988) presents an approach to modelling repeated measurements that accounts for variation between experimental units and serial correlation within units. In this paper this approach is extended to account for spatially referenced data.

Smith *et al.* (2007) present a method to analyse multi-harvest data in perennial crops in the case of a short sequence of measurements in sugarcane. These methods may not be entirely suitable when there are longer sequences of measurement times, as in perennial pasture field trials. In these situations, it is usually not of interest to simply obtain predictions at each harvest time, but rather to investigate the varietal response profile over time, or at specific times of interest, and to obtain an insight into variety by harvest interaction. One approach to investigating variety by harvest interaction in perennial pasture crops is to use a clustering approach as in Hayward *et al.* (1982) and Cullis *et al.* (2010). Piepho and Eckl (2014) also present an approach to analysing a series of variety trials in perennial crops which accounts for serial correlation between repeated measurements but their approach ignores any spatial correlation that may be present.

The approach used to model the genetic effects over time will depend on the aim of the experiment and the trait involved. One approach may be to model the deviations of each variety from the harvest means (Evans and Roberts 1979) rather than the actual means themselves. An alternative approach may be to model the genetic response over time. A method suitable for modelling the genetic profile over time is the random regression (or random coefficients) model, as used commonly in the animal sciences. Random regression is commonly used to model lactation curves and cattle growth data (Meyer 1998; Schaeffer 2004; Meyer and Kirkpatrick 2005) and has also been used in forestry breeding (Apiolaza *et al.* 2000).

In this paper an approach will be presented for analysing data from multi-harvest, perennial pasture field trials that accounts for both spatial and temporal variation and correlation within a trial and genetic correlation between harvests. Suitable approaches will be applied to model the genetic effects over time. The method of analysis will be applied to data from a lucerne breeding trial conducted by the Tamworth Breeding Program in the New South Wales Department of Primary Industries.

Materials and methods

Motivating data

The motivating data considered in this paper arises from a lucerne variety assessment trial conducted by the New South Wales Department of Primary Industries (NSW DPI), at Terry Hie Hie in NSW from 2003–2006. The trial was designed as a Randomized Complete Block (RCB) with 3 replicate Blocks and was laid out in a rectangular array of 180 plots consisting of 30 rows by 6 columns (with each Block consisting of 30 rows by 2 columns). The number of varieties tested in the trial was 60 (with eleven being commercial varieties). The trial was sown on 22/7/2003.

Lucerne yield

The first trait of interest was lucerne yield. Yield was measured by cutting all trial plots at a consistent defined height at each harvest time and drying the samples to obtain dry matter weights expressed as kg/ha. There were 10 harvest times. The data was transformed prior to analysis using a cube root transformation ($(y+1)^{1/3}$), to stabilize the variance and better approximate the assumed Normal distribution. The cube root transformation was chosen (over the more commonly used log transformation) after careful consideration of residual plots, in particular Normal Quantile-Quantile plots of residuals from analyses of each harvest. The cube root transformation provided a less severe transformation, that better approximated the Normal distribution, than the log transformation. The cube root transformation is often used in transforming volume data and given the lucerne yield data arose from cutting lucerne at a certain height from a plot of set length and width, it was considered sensible for this application. A plot of the transformed yield data for the 11 commercial varieties is presented in Fig. 1.

Figure 1 demonstrates that the lucerne yield response is not smooth over time even though the pattern is consistent over varieties. This is due to the nature of the trait where the yield data involves growth between cuts, with the cuts occurring at varying time spacings, and the growth being very dependent on the environment and management of the trial during these different time intervals. Knowing the actual level at each time may not be imperative but the differential impact of each variety from the overall performance of all varieties is of most interest; that is, comparative inference for varieties is required.

Lucerne persistence

The second trait recorded at Terry Hie Hie was persistence. Persistence is a critical component in perennial pasture variety

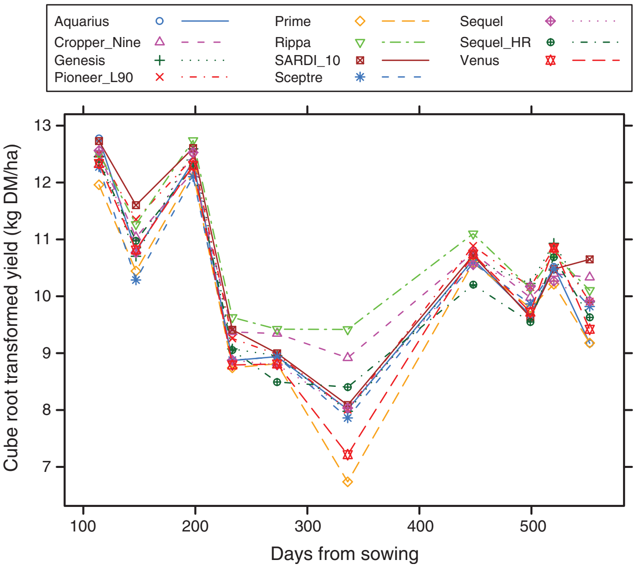

Fig. 1. Plot of mean lucerne yield (on transformed scale) for each of the 11 commercial varieties across harvests at Terry Hie Hie.

improvement as a measure of sustained productivity (Bouton 2012). The persistence of each variety was recorded as the percentage P of unit squares in a grid of 10 by 10 squares that had a lucerne plant(s) present at each of six assessment dates. The details of this method are given in Lodge and Gleeson (1984) where it is shown that this method reliably reflects changes in plant populations, for all but very high plant densities. The data was transformed prior to analysis, using a logit transformation ($\log((P+0.5)/(100-P+0.5))$). This transformation aims to serve two purposes, namely to map the percentage data from the (0,100) range to the real line, and to stabilize the variance, thereby providing a better approximation to the Normal distribution. This transformation was also chosen based on consideration of plots of residuals. A plot of the transformed persistence data for the 11 commercial varieties is presented in Fig. 2.

In the case of the lucerne persistence data, the continuous nature of the trait results in a relatively smooth profile over time (see Fig. 2). In this situation the actual level of variety response is of interest and predictions of time to a certain level of persistence are desired. Hence predictions are required for the actual variety response at times other than the harvest times. In the analysis presented in this paper the time until the persistence of each variety declines to 30% (or –0.838 on the transformed scale) is investigated.

Statistical methods

We begin by establishing some notation for the general approach presented in this paper. Consider a perennial crop variety selection trial consisting of n plots in a rectangular array of c columns by r rows ($n = cr$), in which m genotypes are grown and multiple harvests are made. Let h denote the number of harvests (or assessment dates) for the trial and let $\boldsymbol{y}$ be the $hn \times 1$ vector of data observations across all the harvests, ordered as rows within columns within harvests.

A linear mixed model for the data may be written as:

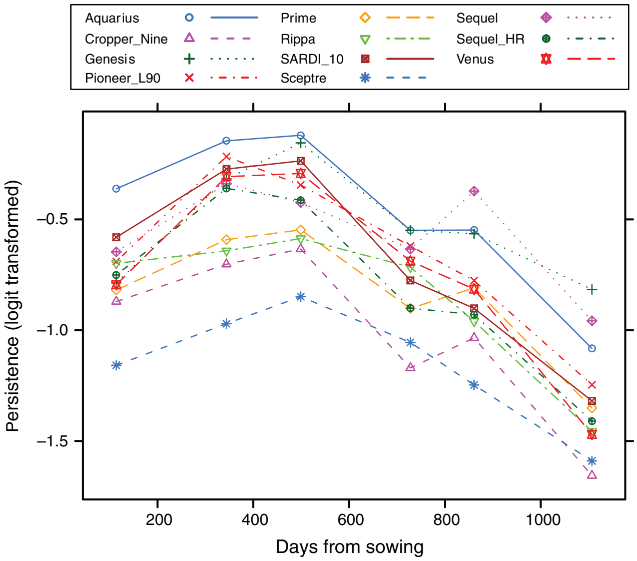

Fig. 2. Plot of mean lucerne persistence data (on transformed scale) for each of the 11 commercial varieties at Terry Hie Hie.

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{g} + \boldsymbol{Z}_o\boldsymbol{u}_o + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\tau}$ is a $(p \times 1)$ vector of p fixed effects with design matrix $\boldsymbol{X}^{(hn \times p)}$, $\boldsymbol{g}$ is the $hm \times 1$ vector of random variety (or genetic) effects for individual harvests with associated design matrix $\boldsymbol{Z}_g^{(hn \times hm)}$, $\boldsymbol{u}_o$ is a vector of other random effects with associated design matrix $\boldsymbol{Z}_o$ and $\boldsymbol{e}$ is the $hn \times 1$ vector of residuals.

The random effects from the linear mixed model (1) are assumed to follow a Normal distribution with zero mean vector and variance-covariance matrix:

$$\mathrm{var}\left(\begin{bmatrix}\boldsymbol{g}\\ \boldsymbol{u}_o\\ \boldsymbol{e}\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{G}_g & 0 & 0\\ 0 & \boldsymbol{G}_o & 0\\ 0 & 0 & \boldsymbol{R}\end{bmatrix} \tag{2}$$

Therefore, the distribution of the data $\boldsymbol{y}$ is normal with mean $\boldsymbol{X\tau}$ and variance matrix:

$$\boldsymbol{H} = \mathrm{var}(\boldsymbol{y}) = \boldsymbol{Z}_g\boldsymbol{G}_g\boldsymbol{Z}_g^T + \boldsymbol{Z}_o\boldsymbol{G}_o\boldsymbol{Z}_o^T + \boldsymbol{R}$$

This linear mixed model provides the basis for analysis.

The actual modelling approach will depend on the trait involved and the aim of the experiment.

If the response is not a smooth function over time and/or interest lies in the differences between varieties more than the actual level of a trait then it may be best to base the analysis on the deviations from the harvest means. Hence, the ideal approach to analysing the yield data is to model the variety deviations from the harvest means. In terms of the mixed model (1) this means that the vector $\boldsymbol{\tau}$ contains the main effects for harvests. If the response of the trait over time is a smooth continuous function and the actual level of the trait is of interest then the ideal approach would be to model the response over time using a smooth curve. Hence in the case of the persistence data, the ideal approach is to model the underlying overall trend over time using a smooth curve (for example using a polynomial or cubic smoothing spline) and then investigate the departures from this underlying trend for each variety. These variety departures may be modelled using linear functions or may require more complex models including splines (Verbyla *et al.* 1999).

Modelling involves a sequential process to arrive at a best model in terms of the fit to the data. The steps involve allowing for non-genetic variation, through design, management and other sources of variation such as spatial trends in the field, accounting for temporal variation and correlation that is inherent in the multiple harvests, and importantly from the breeding point of view, modelling the genetic variation through genotypes that are investigated in the trial. The following subsections detail modelling of the genetic and non-genetic variation that may be present.

Modelling non-genetic effects

In order to obtain accurate predictions of genetic effects it is essential to suitably model the non-genetic effects such as spatial variation in the field and temporal correlation between repeated measurements. Spatial variation has long been recognized as an important issue in field trials. The classic Fisherian or randomization approach (see Brien and Bailey 2006 and references therein) attempts to minimize the effect of spatial variation through careful selection of the trial sites, application of randomization of plots to varieties, and inclusion of sources of variation due to the design in the analysis. Often this randomization approach is not sufficient to account for all sources of spatial variation inherent in field trials. Variation may occur due to management practices impacting on the experiment, non-stationary spatial trend occurring across the field, and neighbouring plots being more similar than those further apart due to soil fertility or moisture levels. The spatial analysis method of Gilmour *et al.* (1997), in which spatial variation is partitioned into three components, namely local and global smooth spatial trend and extraneous variation, provides a means for accounting for these additional forms of variation. Our approach to modelling the spatial variation is to commence with the randomization model and to build on this using the method of Gilmour *et al.* (1997). As the lucerne trial was designed as a RCB, the randomization model includes a Block effect for each harvest time.

Following the approach of Gilmour *et al.* (1997) and Stefanova *et al.* (2009), diagnostics such as the sample variogram and residual plots are used to identify departures from the randomization model based on the experimental design. Identified additional sources of variation may be included in some or all of $\boldsymbol{\tau}$, $\boldsymbol{u}_o$ and $\boldsymbol{e}$ in (1), depending on their origin. For example, global and extraneous variation can be accounted for by including model-based terms (usually involving row or column co-ordinates) in $\boldsymbol{u}_o$ and $\boldsymbol{\tau}$. Local smooth spatial trend which arises because data from plots close together are more similar to those further apart, is modelled in $\boldsymbol{e}$, typically using a separable correlation structure in the residuals. For example, if a separate model for each harvest within the trial is assumed, and $\otimes$ denotes the Kronecker product, the residual variance matrix for harvest j, $\boldsymbol{R}_j$, may be written as:

$$\boldsymbol{R}_j = \sigma_j^2 \boldsymbol{\Sigma}_{cj} \otimes \boldsymbol{\Sigma}_{rj} \tag{3}$$

where $\boldsymbol{\Sigma}_{cj}$ and $\boldsymbol{\Sigma}_{rj}$ are the $c \times c$ and $r \times r$ spatial correlation matrices corresponding to the column and row dimensions respectively as in Gilmour *et al.* (1997). These two matrices are typically assumed to arise from autoregressive processes of order 1 (labelled ar1 using the notation presented in the Appendix 1), so that correlation is allowed in both column and row directions. The joint model is denoted by ar1(Column).ar1(Row).

In a multiple harvest trial it is expected that harvests will be correlated, as measurements are made on the same plots (Bjornsson 1978; Diggle 1988; Piepho and Eckl 2014). Smith *et al.* (2007) incorporate such correlation by assuming a three-way separable process for the residual variance structure, denoted by $\boldsymbol{R}$, with:

$$\boldsymbol{R} = \boldsymbol{R}_h \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r \tag{4}$$

where $\boldsymbol{R}_h$ is a $h \times h$ covariance matrix that incorporates temporal correlation (between harvests) and possibly heterogeneous variance across harvests and $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_r$ are the $c \times c$ and $r \times r$ column and row spatial correlation matrices; in this formulation these latter structures are common to all harvests within a trial.

Smith *et al.* (2007) model $\boldsymbol{R}_h$ using an unstructured matrix, the most general form of covariance matrix, requiring the estimation of $h(h+1)/2$ parameters. For a small number of harvests (as in (Smith *et al.* 2007)) this may be achievable but for more extensive sequences of repeated harvests the number of parameters to be estimated may become prohibitive. For example for yield at Terry Hie Hie this is $10 \times 11/2 = 55$ parameters. More parsimonious covariance structures such as the uniform structure (with equal variances and equal correlation between harvests) or heterogeneous covariance model (with differing variances and equal correlation) could be considered but are unlikely to be suitable in practice as correlations between the different harvest times are unlikely to be constant.

A more suitable parsimonious model can be based on the repeated measures analysis approach of Diggle (1988). The model proposed by Diggle (1988) for y_{ij} (where y_{ij} denotes the j^{th} measurement on the i^{th} unit (or plot), with t_{ij} denoting the time that measurement y_{ij} was made) is given by:

$$y_{ij} = \mu_{ij} + e_{ij}$$

with

$$e_{ij} = \zeta_i + \eta_{ij} + \xi_i(t_{ij})$$

where μ_{ij} is the mean at time t_{ij}, ζ_i is a unit (or plot) effect, η_{ij} is a measurement error and $\xi_i(t_{ij})$ is a temporally correlated process to account for the serial correlation between measurements on the same unit. The latter three effects are random. If $\boldsymbol{e}_i$ is the vector of residuals for unit i, the variance matrix of $\boldsymbol{e}_i$ is:

$$\text{var}(\boldsymbol{e}_i) = \sigma_p^2 \boldsymbol{J}_h + \sigma_m^2 \boldsymbol{I}_h + \sigma^2 \boldsymbol{R}_h^*(\phi) \tag{5}$$

where σ_p^2 is a between plot or unit variance, $\boldsymbol{J}_h$ is a $h \times h$ matrix with all elements equal to 1, σ_m^2 is the measurement error variance, $\boldsymbol{I}_h$ is the identity matrix of order h, σ^2 is the error scale parameter and $\boldsymbol{R}_h^*(\phi)$ is a smooth (typically) correlation structure over time.

Diggle (1988) assumes independent units, however in the case of multi-harvest data the units are plots in the field experiment which may be spatially correlated. Our model for multi-harvest data should include terms for plot effects, measurement error and serial dependence and allow for these effects to be spatially correlated. Thus in the linear mixed model for multi-harvest data (1) the residual term $\boldsymbol{e}$ can be partitioned into a vector of random plot effects $\boldsymbol{\zeta}$, a temporal correlation process $\boldsymbol{\xi}$ and a vector of measurement errors $\boldsymbol{\eta}$, where each of these three random effects may have their own spatial structure. Hence,

$$\boldsymbol{e} = \boldsymbol{\zeta} + \boldsymbol{\eta} + \boldsymbol{\xi} \tag{6}$$

where $(\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\xi})$ are pairwise independent, jointly normal, mean zero and have variance matrix:

$$\text{var}\left(\begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{bmatrix}\right) = \begin{bmatrix} \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{\Sigma}_c^{(p)} \otimes \boldsymbol{\Sigma}_r^{(p)} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{\Sigma}_c^{(m)} \otimes \boldsymbol{\Sigma}_r^{(m)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma^2 \boldsymbol{R}_h^*(\phi) \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r \end{bmatrix}$$

where $\boldsymbol{\Sigma}_c^{(p)}$ and $\boldsymbol{\Sigma}_r^{(p)}$ are the spatial correlation matrices (in column and row directions respectively) for the overall plot term, and $\boldsymbol{\Sigma}_c^{(m)}$ and $\boldsymbol{\Sigma}_r^{(m)}$ are similar terms for the measurement error component. We note that if the same spatial correlation structure is assumed for the three random effects ($\boldsymbol{\zeta}$, $\boldsymbol{\eta}$, $\boldsymbol{\xi}$), the variance model generalizes to:

$$\mathrm{var}(\boldsymbol{e}) = (\sigma_p^2 \boldsymbol{J}_h + \sigma_m^2 \boldsymbol{I}_h + \sigma^2 \boldsymbol{R}_h^*(\phi)) \otimes \boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c$$

which is a separable spatial extension of (5) (the model of Diggle (1988)).

Whilst this separable extension may be theoretically appealing, it may be questionable as to whether the measurement error term should be spatially correlated. For a purely spatial (single harvest time) model the measurement error is assumed to be independently and identically distributed (i.i.d) random "white noise", and similarly in the temporal case with spatially independent measurements, the measurement error is also assumed to be i.i.d. It may be more reasonable to assume the measurement error term in the spatio-temporal context is also independent and not spatially correlated. It may however, be reasonable to assume the same spatial structure for the overall plot (or unit) effect and the plot by harvest effects and hence the variance matrix can be modified to:

$$\mathrm{var}(\boldsymbol{e}) = (\sigma_p^2 \boldsymbol{J}_h + \sigma^2 \boldsymbol{R}_h^*(\phi)) \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r \quad (7)$$

Model (7) is still restrictive, and a more desirable model may be to allow $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ to have differing spatial correlation structures and to assume an independent measurement error. Thus, we may assume:

$$\begin{aligned}\mathrm{var}(\boldsymbol{e}) = {} & \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{\Sigma}_c^{(p)} \otimes \boldsymbol{\Sigma}_r^{(p)} + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r \\ & + \sigma^2 \boldsymbol{R}_h^*(\phi) \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r \end{aligned} \quad (8)$$

This model has the advantage of increased flexibility and an independent measurement error component. Other models are possible and these are discussed in De Faveri (2013).

Following Diggle (1988), $\boldsymbol{R}_h^*$ may be modelled using a decaying correlation model, which implies the correlation between harvests decreases as the time between harvests increases. For unequally spaced time points the exponential (or power model) (exp) may be appropriate, while for equally (or close to) spaced measurements the autoregressive (ar1) correlation process may be suitable. This structure may be generalized to a heterogeneous variance process (for example a heterogeneous autoregressive model (ar1h), a heterogeneous exponential model (exph), or antedependence model (ante)) to account for differing variances at each harvest. Details of models and notation used in relation to the models are given in the Appendix. Note that the serial temporal correlation models have correlations that typically decay to zero as the separation between harvests increases; however by also including an overall plot term a positive correlation is induced between plots and hence the overall temporal correlation need not decay to zero.

Modelling $\boldsymbol{R}_h^*$ (in 8) as an ar1h process, enables the spatial and temporal residual correlation structure to be modelled using a maximum of $h+7$ parameters, in comparison to the $h(h+1)/2+2$ parameters required for the separable model used in Smith *et al.* (2007). For a trial with 10 harvests this equates to a difference of 40 parameters.

The original temporal model of Diggle (1988) and the spatial model of Gilmour *et al.* (1997) were both motivated biologically. Similar motivation carries over to the spatio-temporal case. Thus the terms can be motivated as follows. The overall plot effect reflects the fact that some plots perform better than others, for example producing higher yield over all times while others have reduced performance. These overall plot effects are likely to be spatially correlated due to local spatial soil fertility or moisture trends, with plots closer together more highly correlated than those further apart. The separable spatial temporal component models the temporal serial correlation between repeated measurements (where measurements close together in time are likely to be more highly correlated than those further apart in time), the local spatial structure of the plot by harvest effects, and also the interaction between spatial and temporal components. As for the plot effects the dynamics of both spatial and temporal changes in local trends such as soil fertility and moisture are accommodated. The measurement error term represents the possibility that repeating the measurement process on a plot will not result in exactly the same value. It is assumed to be i.i.d. random white noise. Some traits may have higher measurement error than others depending on the complexity and accuracy of the measurement process. The global smooth trend and extraneous spatial variation may arise due to management practices, for example serpentine harvesting or irrigation systems aligned with rows and or columns, or global soil trends and slope effects across the trial.

Modelling genetic effects

Smith *et al.* (2007) present various models for the genetic effects in the multi-harvest situation where only a small number of harvests are involved. In particular they represent the variance matrix $\boldsymbol{G}_g$ for such effects by:

$$\boldsymbol{G}_g = \boldsymbol{G}_h \otimes \boldsymbol{I}_m \quad (9)$$

where $\boldsymbol{G}_h$ is a $h \times h$ matrix representing the harvest genetic covariance structure (variances and correlations for the harvests) with possible forms being unstructured and factor analytic models, and $\boldsymbol{I}_m$ is the assumed structure for the varieties.

In the case of perennial pasture variety selection trials the number of harvests is likely to be much greater than that in Smith *et al.* (2007). While it may be possible to estimate the $h \times h$ variance covariance matrix $\boldsymbol{G}_h$ using factor analytic models (Smith *et al.* 2001) and predict genetic effects for each harvest time, it may be more desirable for selection purposes to reduce the estimation of genetic effects to a smaller dimension. Two approaches that may reduce the dimension of the selection problem are detailed in the next section. One approach is the method of random regression. A second approach is to use factor analytic models in conjunction with clustering to group the harvests into target groups and form predictions for each group.

Random regression

A suitable model for estimating the genetic response over time is the random regression (or random coefficients) model (Laird and Ware 1982). Random regression models involve fitting regression coefficients on time (or other explanatory variables), for each variety, as random effects. This allows for variation between varieties in the shape of the response profile over time.

The covariances among random regression coefficients implicitly model the genetic covariance structure of the varieties over time. The random regression model allows the dimension of genetic effects of interest to be reduced to the smaller number of coefficients in the random regression model.

Let g_{ij} denote the random effect for variety i at harvest j (where $j = 1, \ldots, h$), and x_j represent the value for the explanatory variable x (e.g. time) at harvest j, then a polynomial random regression model of order p, over x for variety effects g_{ij} can be formulated as:

$$\begin{aligned} g_{ij} &= u_{pi0} + u_{pi1}x_j + \ldots + u_{pip}x_j^p + \epsilon_{ij} \\ &= \boldsymbol{x}_{pj}^T \boldsymbol{u}_{pi} + \epsilon_{ij} \end{aligned} \tag{10}$$

where$\boldsymbol{x}_{pj}^T = \left[1\ x_j\ x_j^2\ \ldots\ x_j^p\right]$ and $\boldsymbol{u}_{pi} = \left[u_{pi0}\ u_{pi1}\ u_{pi2}\ \ldots\ u_{pip}\right]^T$. The term $\varepsilon_{\mathrm{ij}}$ represents a residual term for genetic effects, assumed to be independent and identically distributed, with variance σ^2_{g}. This model can also be written as:

$$\boldsymbol{g}_i = \boldsymbol{X}_p \boldsymbol{u}_{pi} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{g}_i = [g_{i1}\quad g_{i2}\quad g_{i3}\quad \ldots\quad g_{ih}]^T$ and

$$\boldsymbol{X}_p = \begin{bmatrix} \boldsymbol{x}_{p1}^T \\ \vdots \\ \boldsymbol{x}_{ph}^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ \vdots & & & & \\ 1 & x_h & x_h^2 & \ldots & x_h^p \end{bmatrix}$$

If $\boldsymbol{u}_{pi} \sim N(\boldsymbol{0}, \boldsymbol{G}_p)$, where $\boldsymbol{G}_p$ is a $(p+1) \times (p+1)$ covariance matrix for the random polynomial terms, then

$$\boldsymbol{G}_h = \mathrm{var}(\boldsymbol{g}_i) = \boldsymbol{X}_p \boldsymbol{G}_p \boldsymbol{X}_p^T + \sigma_g^2 \boldsymbol{I}$$

$\boldsymbol{G}_p$ is taken as an unstructured matrix to ensure translation invariance of the model.

In the case of $p = 1$, (10) reduces to the linear random regression for modelling linear trends:

$$g_{ij} = u_{i0} + u_{i1}x_j + \epsilon_{ij}$$

where u_{i0} and u_{i1} are the random intercept and slope terms (respectively) for variety i.

By including further polynomial terms in the random regression model non-linear trends may be modelled. Alternatively it may be preferable to use natural cubic smoothing splines (Verbyla *et al.* 1999) to provide a more flexible specification.

Firstly, in some cases it is appropriate to model the mean trend over harvests using a cubic smoothing spline. This also enables the notation to be established for the variety by harvest interaction. So consider the mean effect vector $\boldsymbol{\mu}$ ($h \times 1$). Verbyla *et al.* (1999) formulate the cubic smoothing spline as a mixed model:

$$\boldsymbol{\mu} = \boldsymbol{X}_1 \boldsymbol{\tau} + \boldsymbol{Z}_s \boldsymbol{u}_s + \boldsymbol{e}_d$$

where $\boldsymbol{X}_1$ is $\boldsymbol{X}_p$ with $p = 1$; thus $\boldsymbol{X}_1\boldsymbol{\tau}$ is a fixed effects linear regression on harvest time. The random effects term $\boldsymbol{Z}_{\mathrm{s}}\boldsymbol{u}_{\mathrm{s}}$ provides the smooth nonlinear component of the cubic smoothing spline. If $d_j = x_{j+1}$-x_j, then $\boldsymbol{Z}_s = \boldsymbol{\Delta}\ (\boldsymbol{\Delta}\ \boldsymbol{\Delta}^T)^{-1}$ and $\boldsymbol{u}_s \sim N(\boldsymbol{0}, \sigma^2_s \boldsymbol{G}_s)$ where the non-zero elements of $\boldsymbol{\Delta}$ ($h \times (h$-$2)$) and $\boldsymbol{G}_s$ ((h-$2) \times (h$-$2)$) are for $j = 2, 3, \ldots, h - 1$

$$\Delta_{j-1,j} = \frac{1}{d_{j-1}}, \quad \Delta_{jj} = -\left(\frac{1}{d_{j-1}} + \frac{1}{d_j}\right), \quad \Delta_{j+1,j} = \frac{1}{d_j}$$

and

$$G_{s;j-1,j} = G_{s;j,j-1} = \frac{d_j}{6}, \quad G_{s;j-1,j-1} = \frac{d_{j-1} + d_j}{3}$$

Note that $\boldsymbol{e}_d \sim N(\boldsymbol{0}, \sigma^2_d \boldsymbol{I}_h)$ provides for the non-smooth deviations from the cubic smoothing spline. In the notation used later in the paper, the model is represented symbolically as:

$$1 + \mathtt{lin(x)} + \mathtt{spl(x)} + \mathtt{dev(x)}$$

where the first two terms are the linear regression terms, spl(x) is the curvature term in the cubic smoothing spline and dev(x) represents the non-smooth deviations.

Cubic smoothing splines can be used to model the variety by harvest effects using:

$$\boldsymbol{g}_i = \boldsymbol{X}_1 \boldsymbol{\tau}_i + \boldsymbol{Z}_s \boldsymbol{u}_{si} + \boldsymbol{e}_{di}$$

where in this model linear random coefficients are used instead of a fixed linear regression. Symbolically this may be written as:

$$\begin{aligned} &\mathtt{Variety} + \mathtt{Variety.lin(x)} + \mathtt{Variety.spl(x)} \\ &+ \mathtt{Variety.dev(x)} \end{aligned}$$

so that Variety is interacted with each term in the basic spline model.

Factor analytic models

The application of Factor Analytic (fa) models to multi-environment trials is outlined in Smith *et al.* (2001). Similar principles could apply in the case of multi-harvest data where measurements from different harvests can be regarded as separate traits and variety effects at different harvest times are assumed correlated.

An fa model (of order s) can be fitted to the variety effects at each harvest, with the genetic effects given by:

$$g_{ij} = \sum_{r=1}^{s} \lambda_{jr} f_{ir} + \delta_{ij} \tag{11}$$

where g_{ij} is the random effect for variety i, $i = 1, \ldots, m$ at harvest j, $j = 1, \ldots, h$, f_{ir} is the score for variety i in the r^{th} factor, λ_{jr} is the loading for harvest j for the r^{th} factor and δ_{ij} is a residual.

In vector notation the genetic effects are given by:

$$\boldsymbol{g} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\boldsymbol{f} + \boldsymbol{\delta}$$

where $\boldsymbol{\Lambda}$ is a $h \times s$ matrix of loadings, $[\boldsymbol{\lambda}_1\ \boldsymbol{\lambda}_2\ \ldots\ \boldsymbol{\lambda}_s]$, $\boldsymbol{f}$ is a $ms \times 1$ vector of factor scores, $[\boldsymbol{f}_1{}^T, \boldsymbol{f}_2{}^T, \ldots, \boldsymbol{f}_s{}^T]^T$, and $\boldsymbol{\delta}$ is a $mh \times 1$ vector of residuals, $\{\delta_{ij}\}$.

The joint distribution of $\boldsymbol{f}$ and $\boldsymbol{\delta}$ is assumed to be:

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{\delta} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G}_f \otimes \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi} \otimes \boldsymbol{I}_m \end{pmatrix}\right]$$

where $\boldsymbol{\Psi}$ is a diagonal matrix of specific variances, one for each harvest, that is $\boldsymbol{\Psi} = \mathrm{diag}\ (\Psi_1\ \ldots\ \Psi_h)$. The factor scores are commonly assumed to be independent and scaled to have unit variance, so that $\boldsymbol{G}_f = \boldsymbol{I}_s$. Hence the variance matrix for the variety effects at each harvest is given by:

$$\begin{aligned} \text{var}(\boldsymbol{g}) &= (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_m)\text{var}(\boldsymbol{f})(\boldsymbol{\Lambda}^T \otimes \boldsymbol{I}_m) + \text{var}(\boldsymbol{\delta}) \\ &= (\boldsymbol{\Lambda\Lambda}^T + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m \end{aligned}$$

Therefore the fa model results in $\boldsymbol{G}_h$, a $h \times h$ genetic variance matrix, of the form:

$$\boldsymbol{G}_h = \boldsymbol{\Lambda\Lambda}^T + \boldsymbol{\Psi}$$

In theory this enables predictions for each variety at each harvest ($\boldsymbol{g}$), to be calculated and these predictions can be combined to form a weighted selection index, see Smith *et al.* (2007). However, in practice when the number of harvests is large this may not be ideal and it may be desirable to form a selection index based on a smaller number of 'traits'. The approach of Cullis *et al.* (2010) may be used in which cluster analysis is performed, based on the genetic correlation matrix, to group the harvests into target groups and form variety predictions for each selection group. This method allows investigation of variety by harvest interaction.

Estimation, model selection and software

Estimation of variance parameters from the mixed model was performed using Residual Maximum Likelihood (REML) (Patterson and Thompson 1971), using the Average Information algorithm as implemented in ASReml in the R environment (Butler *et al.* 2009). The variance parameters were used to obtain Best Linear Unbiased Estimates (BLUEs) of the fixed effects and Best Linear Unbiased Predictions (BLUPs) of the random effects (Robinson 1991).

To test the significance of random effects in the linear mixed model the Residual Maximum Likelihood Ratio Test (REMLRT) can be used. The REMLRT may be used to compare the fit of two models only if they are nested and contain the same fixed effects. For two nested models, M_0 and M_1 with M_1 having p_1 variance parameters and M_0 having p_0 variance parameters, with $p_1 > p_0$, the Residual Maximum Likelihood Ratio Test Statistic (REMLRS) is calculated as $-2(l_0 - l_1)$ where l_0 is the residual log-likelihood for model M_0 and l_1 is the residual log-likelihood for model M_1.

The standard REMLRS is asymptotically distributed as a chi-squared statistic with $p_1 - p_0$ degrees of freedom. If however the test involves a null hypothesis where the parameter is on the boundary of the parameter space the REMLRT needs to be adjusted. For a test of a single variance component the theoretical asymptotic distribution of the REMLRS is a mixture of chi-squared variates where the mixing probabilities are 0.5, one with 0 degrees of freedom (a spike at 0) and the other with 1 degree of freedom. The approximate P value for the REMLRS is $0.5(1 - Pr(\chi^2 \leq d))$ where d is the observed value of the REMLRS (see Stram and Lee 1994).

To compare the goodness of fit of two models (with the same fixed effects) that may be non-nested, the Akaike Information (AIC) criterion may be used. The AIC value for a model is calculated as $-2(l - p)$, where l is the residual log-likelihood for the model and p is the number of variance parameters in the model. Models with smaller AIC values provide a better fit to the data.

Results

Lucerne yield analysis

The analysis of yield was based on the linear mixed model (1). This analysis requires appropriate variance models for the genetic ($\boldsymbol{G}_g$) and residual effects ($\boldsymbol{R}$). As it is difficult to find optimal models for both $\boldsymbol{R}$ and $\boldsymbol{G}_g$ at the same time, the sequential approach of Smith *et al.* (2007) was followed, by firstly assuming a simple genetic model in order to determine a suitable residual model. Using this residual model more complex models for the genetic effects were investigated. The initial simple genetic model has genetic variance matrix $\boldsymbol{G}_g$ given by:

$$\boldsymbol{G}_g = \text{diag}\ (\sigma_{gj}^2) \otimes \boldsymbol{I}_m \qquad (12)$$

where σ_{gj}^2 is the genetic variance for harvest j. This initial genetic model therefore allows for a different genetic variance for each harvest and assumes the genetic effects are independent between harvests. It also assumes the varieties are unrelated.

The description of the sequential model-building process has been separated into two sections below, namely modelling the non-genetic (residual) effects and modelling the genetic effects, with the residual models given in Table 1 and the genetic models given in Table 3. In all models, a fixed Harvest term is fitted in the linear mixed model (in $\boldsymbol{\tau}$), so the variety deviations from the harvest means are being modelled.

The notation for terms used in the residual and genetic models are defined in the Appendix 1.

Modelling non genetic effects

The first step in modelling the non-genetic (residual) effects was to account for any spatial variation in the data. Initially this involved investigating spatial models for each harvest separately, using Eqn 3.

The sequence of residual models fitted is presented in Table 1. The initial model fitted to each of the harvests was the RCB model (in this case $\boldsymbol{\Sigma}_{cj}$ and $\boldsymbol{\Sigma}_{rj}$ in Eqn 3 are given by the identity matrices $\boldsymbol{I}_c$ and $\boldsymbol{I}_r$ respectively) and the random effects vector $\boldsymbol{u}_o$ contains the replicate block effects for each harvest. This represents the baseline model (Y1). In this model (and Y2–Y8, see Table 1) the simple genetic model of Eqn 12 was included at the genetic level.

Model Y2 includes the local spatial correlation between plots at each harvest time, using a separable autoregressive (ar1(Column).ar1(Row)) process for each harvest. Model Y2 was a significant improvement on Y1 as can be seen formally by the Residual Maximum Likelihood Ratio Test Statistic (REMLRS) of 233.08 on 20 df ($P < 0.001$).

Using diagnostic tools such as the variogram and plots of residuals, as in Gilmour *et al.* (1997) and Stefanova *et al.* (2009), further terms were identified and included in the model to account for global and extraneous spatial variation at each harvest time. These terms included a random column effect for harvests 2, 3, 4, 5, 6, 8 (included in u_o) and a fixed effect for row 1 for harvest 10. These terms are included in models Y3-Y12 in Tables 1 and 3.

In model Y4 a common ar1(Column).ar1(Row) spatial model was included across the harvests (common spatial parameters for each harvest) and allowed for heterogeneity of variance over time. This model was not significantly different to Model Y3 (REMLRS = 17.17 on 18 d.f., $P = 0.511$). This suggests that common spatial correlation parameters across the harvests may be a reasonable assumption. Models Y1 to Y4 did not attempt to account for the temporal correlation between the multiple harvests. Subsequent models allowed for such correlation at the residual level.

To decide on an appropriate model for the residual covariance matrix, the empirical variance and covariance matrix of the residuals from Y3 was calculated. The empirical variances and correlations are presented in Table 2. This matrix suggests possible models for the residual variance structure. It is apparent that there is positive correlation between all harvests with the correlations generally decaying as the time between measurements increases, but the correlations do not decay to zero.

Based on these observations, a plausible model is to assume an overall average plot effect (Column.Row), a decaying correlation process for the plot by harvest effects and a measurement error effect as given by the extension of Diggle (1988) in Eqn 8.

To build up to this model, firstly a simpler model was fitted (Y5), using a 3 way separable process (ar1v(Harvest).ar1(Column).ar1(Row)) for the plot by harvest effects, thereby modelling the temporal correlation and the spatial correlation in each direction (row and column) each with an autoregressive process of order 1, with a common residual variance across the harvests. In this model the spatial correlation model for plots was common for all harvests. The variances for each harvest in Table 2 (ranging from 0.095 to 0.253), indicate that it may be more suitable to assume heterogeneous residual variances for each harvest, despite the transformation to cube root. Model Y6 allowed for differing variances across the harvests and was a significant improvement on model Y5 (P value <0.001).

Table 1. Summary of models (given in both statistical and ASReml notation) fitted for yield at Terry Hie Hie investigating different residual variance models. In all models the simple genetic effects model of (12) has been fitted (with 10 variance parameters). In all models a random Block effect has been included for each harvest and models Y3–Y9 include global and extraneous spatial effects as outlined in the text. Residual log-likelihoods (denoted by ℓ), number of parameters in *R* (par), number of other non-genetic variance parameters (Other par) and AIC values (given as differences from the best model) are presented for each model. Horizontal lines in the table indicate that fixed effects have been added or removed, and residual likelihood ratio tests and AIC comparisons are therefore not appropriate between models above and below the line. Model terms are detailed in the Appendix

Model	Residual variance matrix $\boldsymbol{R}$	ℓ	$\boldsymbol{R}$ par	Other par	AIC
Y1	$\mathrm{diag}(\sigma_j^2\, \boldsymbol{I}_c \otimes \boldsymbol{I}_r)$ at(Harvest).id(Col).id(Row)	443.490	10	10	998.1
Y2	$\mathrm{diag}(\sigma_j^2\, \Sigma_{cj} \otimes \Sigma_{rj})$ at(Harvest).ar1(Col).ar1(Row)	560.030	30	10	795.0
Y3	$\mathrm{diag}(\sigma_j^2\, \Sigma_{cj} \otimes \Sigma_{rj})$ at(Harvest).ar1(Col).ar1(Row)	606.156	30	16	714.8
Y4	$\mathrm{diag}(\sigma_j^2\, \Sigma_c \otimes \Sigma_r)$ diag(Harvest).ar1(Col).ar1(Row)	597.859	12	16	695.3
Y5	$\sigma^2\, \Sigma_h \otimes \Sigma_c \otimes \Sigma_r$ ar1v(Harvest).ar1(Col).ar1(Row)	801.129	4	16	272.8
Y6	$\boldsymbol{D}\Sigma_h\boldsymbol{D} \otimes \Sigma_c \otimes \Sigma_r$ ar1h(Harvest).ar1(Col).ar1(Row)	834.460	13	16	224.1
Y7	$\boldsymbol{D}\Sigma_h^{ar2}\boldsymbol{D} \otimes \boldsymbol{I}_c \otimes \Sigma_r + \sigma_p^2\boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \Sigma_r^{(p)} + \sigma_m^2\, \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ ar2h(Harvest).id(Col).ar1(Row) + id(Col).ar1v(Row) + Harvest.Col.Row	927.338	16	16	44.4
Y8	$\Sigma_h^{ante2} \otimes \Sigma_c \otimes \Sigma_r + \sigma_p^2\boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \Sigma_r^{(p)} + \sigma_m^2\boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ ante2(Harvest).ar1(Col).ar1(Row) + id(Col).ar1v(Row) + Harvest.Col.Row	965.309	32	16	0.5
Y9	$\Sigma_h^{ante2} \otimes \Sigma_c \otimes \Sigma_r + \sigma_p^2\boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \Sigma_r^{(p)}$ ante2(Harvest).ar1(Col).ar1(Row) + id(Col).ar1v(Row)	964.532	31	16	0

Table 2. Empirical/sample variances (on diagonal) and correlations (off diagonals) of residuals from model Y3

Harvest	1	2	3	4	5	6	7	8	9	10
1	0.241									
2	0.481	0.253								
3	0.475	0.597	0.155							
4	0.354	0.469	0.393	0.114						
5	0.470	0.420	0.440	0.576	0.096					
6	0.288	0.376	0.280	0.594	0.523	0.115				
7	0.438	0.359	0.405	0.501	0.550	0.409	0.130			
8	0.303	0.280	0.392	0.496	0.557	0.478	0.615	0.153		
9	0.344	0.304	0.372	0.308	0.456	0.245	0.543	0.479	0.095	
10	0.351	0.405	0.373	0.386	0.440	0.174	0.420	0.358	0.273	0.204

On closer inspection of the empirical correlations of residuals in Table 2 it appears that while the correlations are decaying they are not decaying as quickly as what would be expected from an ar1 process and it may be more reasonable to fit a correlation process of higher order (for example ar2).

The next model (Y7) includes an ar2 correlation process for the temporal correlation and allows for a spatially correlated plot effect as well as a measurement error term. In this model the spatial correlation at the overall plot level in the column direction was dropped (set to zero) in order to achieve convergence (this is not surprising as the correlation in the column direction was very small for all harvests in models Y2 and Y3.

Antedependence structure (Gabriel 1962) was also investigated for modelling the temporal covariance component. An antedependence model of order s assumes that the j^{th} observation ($j > s$), given the s proceeding observation, is independent of all other proceeding observations. The model is more flexible than the exponential or autoregressive models in that it allows the variances for each harvest time to differ and allows for different antedependence coefficients for each harvest. Initially the antedependence model of order 1 (ante) was tried. The antedependence model of order 2 (ante2) proved a better fit and model Y8 which incorporates the ante2 model was significantly better than previous models based on AIC values. In model Y8 the spatial correlation at the overall plot level in the column direction was set to zero to achieve model convergence. In the final model (Y9) the measurement error term was dropped from Y8 but did not result in a significant drop in log-likelihood ($P = 0.106$). Hence the final 'best' residual model was deemed to be model Y9 and this model was used in the following section where more complex genetic models are incorporated.

Modelling genetic effects

The genetic models fitted are presented in Table 3. The first attempt at improving the genetic model from the simple model of Eqn 12, was to fit an overall variety main effect plus variety by harvest interaction model (model Y10 in Table 3). This model allowed for heterogeneous genetic variances across the harvests and assumed a common genetic covariance between each pair of harvests. This model was not a significant improvement on the previous model.

The next genetic model to be fitted was the factor analytic model. The fitting of the factor analytic model had two purposes. Firstly, it enables the h by h genetic variance matrix to be estimated and hence variety predictions to be obtained for each of the harvest times. Secondly it may be used to identify a reduced number of traits that separate out the varieties. The number of factors required to explain a sufficient amount of the variation and the interpretation of their loadings are both of interest.

The genetic model in Y11 is a factor analytic model with a single factor (fa1). As model Y10 is nested within the fa1 model a direct comparison can be made using a REMLRT. The fa1 model provided a significant improvement in log-likelihood (REMLRT = 55.030 on 9 d.f., $P < 0.001$).

Model Y12 fitted an fa2 model, which whilst not providing a significant improvement in log-likelihood to the fa1 model, (REMLRT = 12.456 on 9 d.f., $P = 0.189$), explained a much greater percentage of variation for many of the harvests which had low percent variance accounted for (%VAF) with the fa1 model. The fa1 model also implied a very simplistic and restrictive structure in this case with the first four harvests being perfectly correlated (genetic correlation of 1) which is unlikely to be the case biologically. In the fa2 model fitted to this data the correlations are shown to be quite different to 1. The fa2 model was therefore chosen as the most suitable model as the resulting genetic correlation structure between harvests made more sense biologically and the model explained more of the total genetic variance. Beeck *et al.* (2010) and Cullis *et al.* (2014) also use the %VAF as a tool to aid in model selection for factor analytic models.

The loadings and %VAF for the two factor model is given in Table 4. The first factor shows harvest 6 (the Winter harvest) as being most important, with the highest genetic variance. This factor also shows harvest 9 being negatively correlated with the other harvests (but with low genetic variance). The second factor is more difficult to interpret but it may be interpreted as a contrast between harvests 1 and 2 and harvests 3, 5, 8 and 9, which may reflect an establishment effect.

The genetic correlations between harvests and the genetic variances for each harvest from the fa2 model are presented in Table 5. As an aid to interpreting the genetic covariance structure between harvests and investigate any variety by harvest interaction the approach of Cullis *et al.* (2010) was followed. These authors use cluster analysis and heat map representation of the genetic correlation matrix in the aim of grouping the harvests into meaningful clusters that may be used for prediction and selection.

Table 3. Summary of genetic models fitted for yield at Terry Hie Hie. Residual log-likelihoods (denoted by ℓ), number of parameters in $\boldsymbol{G}_g$ (par), AIC values (given as differences from the best model), are presented for each model

Model	Genetic model	Genetic variance matrix $\boldsymbol{G}_g$	ℓ	$\boldsymbol{G}_g$ par	AIC
Y9	diag(Harvest).Variety	$\mathrm{diag}(\sigma^2_{gj}) \otimes \boldsymbol{I}_m$	964.532	10	29.87
Y10	Variety + diag(Harvest).Variety	$(\sigma^2_g \boldsymbol{J}_h + \mathrm{diag}\ (\sigma^2_{gj})) \otimes \boldsymbol{I}_m$	964.722	11	31.49
Y11	fa(Harvest,1).Variety	$(\boldsymbol{\Lambda}_1 \boldsymbol{\Lambda}_1^T + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m$	992.237	20	–5.54
Y12	fa(Harvest,2).Variety	$(\boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_m$	998.465	29	0

All models incorporate the residual model Y9 in Table 1

$\boldsymbol{\Lambda}_1$ is a $h \times 1$ matrix of factor loadings.

$\boldsymbol{\Lambda}_2$ is a $h \times 2$ matrix of factor loadings.

$\boldsymbol{\Psi} = \mathrm{diag}\ (\psi_1, \ldots, \psi_h)$ is a diagonal matrix of specific variances.

A cluster analysis, using the agglomerative (nested) hierarchical clustering algorithm in the agnes package in R (R Development Core Team 2012), was performed using the average clustering method (Kaufman and Rousseeuw 1990). The dendrogram of the REML estimates of the dissimilarity matrix ($\boldsymbol{I}_h - \boldsymbol{C}_{mat}$), where $\boldsymbol{C}_{mat}$ is the correlation matrix obtained from the fa2 model, is presented in Fig. 3.

Cullis *et al.* (2010) suggests that clusters formed above a cutoff of approximately 0.6 may not be meaningful. Hence the dendrogram (Fig. 3) suggests possibly 2 main clusters plus harvests 7 and 9 in groups of their own (making 4 clusters). The two main clusters include one cluster of harvests with most harvests having higher genetic variance (harvests 1, 2, 4, 6, 10) and the second cluster consisting of harvests 3, 8 and 5. These conclusions on the groupings of sites are supported by the correlations in Table 5. From these correlations it can be seen that harvests 1, 2, 4, 6, 10 (group 1) are highly correlated with each other, harvests 3, 8, 5 (group 2) are highly correlated with each other and also negatively correlated with the establishment harvest 1, while there is little correlation between harvest 7 (group 3) and the other harvests and also harvest 9 (group 4) and the other harvests.

The groups obtained from the cluster analysis provide a starting point to identify potential target sets of harvests for which predictions of variety effects may be made. Cluster group 1 contains the winter harvest 6 (which has substantially higher genetic variance than the other harvests), the establishment harvest 1 (with little genetic variation) and the early harvests with genetic variance (2, 4), plus harvest 10 which also has moderate genetic variance. This would seem reasonable as winter active varieties are known to perform better in early harvests than their winter dormant counterparts. The varieties are also planted in winter so winter active varieties may establish better and perform better in their first year. The second cluster group is given by harvests 3, 8 and 5. These harvests have low genetic variance and are negatively correlated with the establishment period (harvest 1). Harvest 9 is in it's own group possibly due to the very high rainfall occurring near this harvest, with the time between harvest 8 and 9 containing the highest recorded rainfall at nearby weather stations.

Table 4. REML estimates of rotated factor loadings and percentage variance accounted for (%VAF) from the fa2 model (Y12) fitted to lucerne yield data

Harvest	Λ_1	Λ_2	Ψ	%VAF total
1	0.028	0.039	0.000	100.000
2	0.125	0.048	0.000	100.000
3	0.019	–0.040	0.000	100.000
4	0.177	–0.016	0.000	100.000
5	0.091	–0.088	0.000	98.984
6	0.545	0.006	0.025	92.271
7	0.027	–0.009	0.005	13.644
8	0.038	–0.120	0.000	100.000
9	–0.012	–0.067	0.015	23.541
10	0.137	0.037	0.005	80.461

Variety by harvest interaction is primarily seen between harvest 1 and group 2 with the harvests in group 2 being negatively correlated with harvest 1. Also harvest 9 is negatively correlated with the harvests in group 1.

Selection indices may be formed for each variety based on overall performance over the trial as well as for the cluster groups (Cullis *et al.* 2010). Figure 4 presents the predicted selection indices for each variety for each of the groupings, assuming equal weights for each harvest in each group, and also the total across all harvests. This figure can be used to see how varieties rank overall and across the groups, thereby providing insight into variety by harvest interaction. For example it can be seen that variety 05 (a new test line) performs well overall with high rankings in groups 1, 2 and 4 but is ranked lower in group 3. It can also be seen that the variety predictions for group 1 are highly correlated with the overall total predictions.

Discussions with the breeder can lead to different selection indices being formed using different weightings of the harvests within the groups.

Lucerne persistence analysis

While in some circumstances it may be sufficient and desirable, to model the differences between varieties or variety contrasts over time (as in the previous section for yield), in other instances it may be important to model the actual underlying response over time. For example prediction of actual persistence level for each variety may be required for times other than the harvest times. This prediction can not be obtained from an analysis based on deviations from the harvest means and hence an

Table 5. REML estimates of genetic variances ×100 (on diagonal) and genetic correlations (off diagonals) from the fa2 model (Y12) fitted to lucerne yield data

Harvest	1	2	3	4	5	6	7	8	9	10
1	0.229									
2	0.843	1.787								
3	–0.476	0.072	0.191							
4	0.513	0.894	0.511	3.145						
5	–0.133	0.419	0.930	0.778	1.614					
6	0.576	0.899	0.401	0.955	0.681	32.188				
7	0.116	0.286	0.253	0.361	0.330	0.336	0.577			
8	–0.589	–0.061	0.991	0.392	0.875	0.282	0.215	1.594		
9	–0.437	–0.256	0.393	–0.044	0.265	–0.090	0.023	0.427	1.998	
10	0.700	0.892	0.160	0.841	0.460	0.834	0.278	0.042	–0.188	2.490

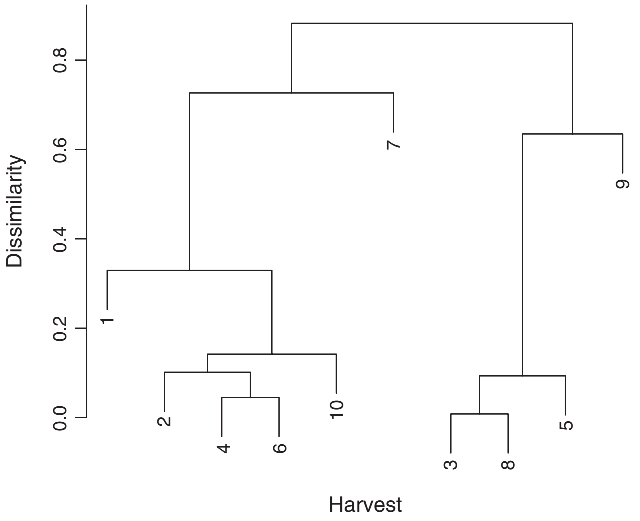

Fig. 3. Dendrogram of the dissimilarity matrix from the fa2 model (Y12) fitted to the lucerne yield data.

alternative approach must be applied. A specific example requiring the modelling of the response over time, is when the time to a certain event (for example time to persistence level dropping to 30%) is the required trait of interest. This problem will be investigated in this section.

Modelling non-genetic effects

Deciding on the most appropriate model to fit to the persistence data involved a sequential approach similar to that of the yield analysis (in the previous section). The process commenced with a simple genetic model in which harvests were treated separately in order to find suitable residual models for modelling the spatial correlation at each harvest. Then the harvests were combined and a residual model incorporating the temporal correlation between harvests was found. At this stage the genetic model was also modified to model the data across the harvests (using a cubic smoothing spline) and in subsequent models the residual model was further refined. The sequence of residual models fitted is given in Table 6 but is not described in detail. The final residual variance model was given by $\sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r^{(p)} + \boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ as in TP8. Once the final residual model was selected then more complex genetic models were investigated.

Modelling genetic effects

Table 7 presents the genetic models fitted to the persistence response over time. The first model listed is TP11. This model included an overall mean spline (1 + lin(years) + spl (years) + dev(years)) and a diagonal variance model for the variety deviations from this overall spline (diag(Harvest). Variety)).The residual variance model was given by TP8. This residual model is common for all models presented in Table 7.

In the subsequent models (TP12-14) different structures were fitted to the variety deviations from the overall underlying spline. Model TP12 includes a factor analytic structure of order 1 to the variety deviations (fa1) in a manner similar to that in the analysis of yield. In model TP13 a fa2 model was used which provided an improvement on the single factor model. Table 8 presents the estimated genetic variances for each harvest and genetic correlations between harvests based on the fa2 model. The genetic correlations were very high for successive times (ranging from 0.836 to 0.940) with the highest genetic correlation for persistence occurring between harvest times 4 and 5 (approximately two, and two and a half years after sowing). It is interesting to note that the genetic correlation between the first and final harvest times (six months and three years after sowing) was negative (–0.280) which may indicate a tendency for some varieties that performed well early in the trial do not show as high persistence later.

Subsequent models followed the approach of Verbyla *et al.* (1999), that is by modelling the overall mean profile over time using a cubic smoothing spline and then random regressions for the variety deviations from this overall mean spline. Model TP14 includes a linear random regression over time for the variety deviations. This linear random regression model correlated the intercepts and slopes for varieties. Further investigations were made to assess if there was any additional curvature from the linear random regression for each variety but for this data set there was not (see Verbyla *et al.* 1999).

Predicting time to P%

One way to investigate the persistence of varieties is to determine the time taken for each variety to decline to a fixed percentage cover. For the lucerne persistence data model TP14 was used and this required special code to be written because the model involves the cubic smoothing spline for the overall mean harvest trend. The approach uses the results from Green and Silverman (1994), in particular their equation (4.14) and an algorithm that firstly determines the interval (in terms of harvest times) where the persistence response curve achieves the set percentage, and then iterates to the estimated time. Details on the procedure and R-code are available from the authors and are presented in De Faveri (2013).

For the persistence data, and using 30% as the level of coverage, the times were estimated for the 60 varieties. The times were ranked and are presented in Fig. 5. It is interesting to note that the most persistent variety (for 30%) is an existing commercial variety, Genesis, and that there are many new varieties that also perform well. The other commercial varieties are dispersed among the new lines.

Discussion

The methods presented in this paper provide a new approach for the analysis of multi-harvest variety selection data from a single site, that accounts for both spatial variation between plots and temporal correlation between harvests, and allows the genetic effects to be modelled over time. In both data sets analysed there was substantial spatial and temporal correlation. The three components of temporal correlation of Diggle (1988) for repeated measures and extended in this paper were found across the two examples. There were large plot effects and serial correlation in both analyses and significant measurement error in the analysis of the persistence data. The models were relatively easy to fit and converged within minutes, however occasionally the measurement error

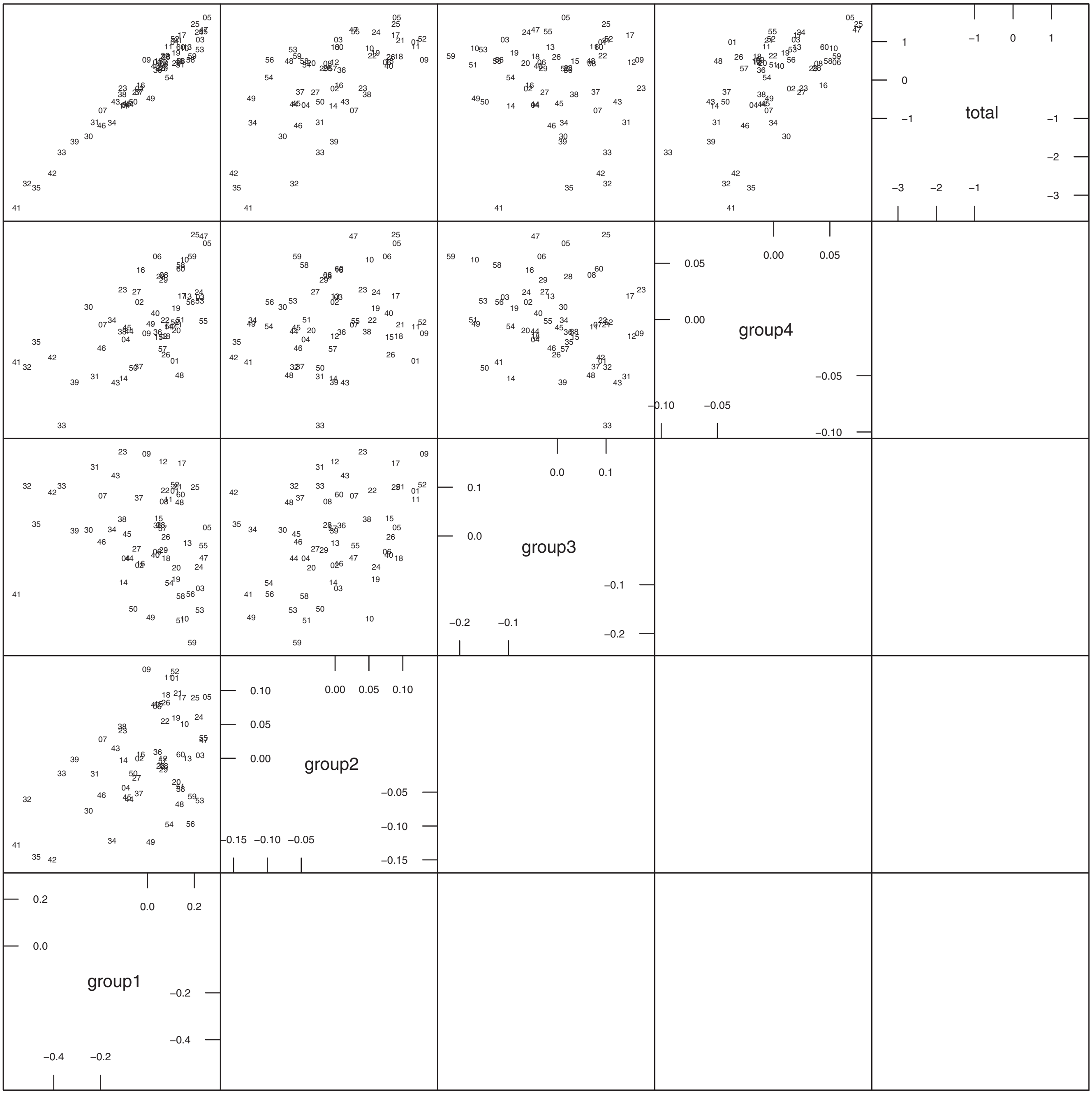

Fig. 4. Pairwise plots of the predictions for lucerne yield for each of the four groups described in the text plus the total over all harvests.

component was difficult to fit and caused convergence problems.

The combined approach of fitting the new extended spatial and temporal residual models plus the models for genetic effects over time is a new approach that builds on the models of Smith *et al.* (2007). At the residual level, it has been shown that the models presented provide an approach that is significantly better in fit than assuming independence between harvest times. However the three way separable (harvest by column by row) structure assumed for the spatio-temporal correlation may not always be appropriate (Smith *et al.* 2007). The model assumes common spatial parameters over harvest times which may not always be the case. It may be expected that the spatial variation between harvest times may differ due to seasonal changes and growth phase of the crop. Investigating alternatives for these three way separable residual models is an area of current research.

The genetic models presented relate to the repeated measures nature of the data and performance is modelled at

Table 6. Summary of models (given in both statistical and ASReml notation) fitted for persistence at Terry Hie Hie investigating different residual variance models. In all models the simple genetic effects model of (12) has been fitted (with 6 variance parameters). In all models a random Block effect has been included for each harvest and models TP3-TP10 include global and extraneous spatial effects. In models TP1-TP5 a separate mean for each Harvest has been fitted while in models TP6-TP10 the overall mean has been modelled over time using the spline model: 1 + lin(years) + spl(years) + dev(years). Residual log-likelihoods (denoted by ℓ), number of parameters in *R* (par), number of other non-genetic variance parameters (Other par) and AIC values (given as differences from the best model) are presented for each model. Groups of models above and below the horizontal line have different fixed effects and are not comparable using residual likelihood ratio tests or AIC values. Model terms are detailed in the Appendix

Model	Residual model $\boldsymbol{R}$	ℓ	$\boldsymbol{R}$ par	Other par	AIC
TP1	$\text{diag}(\sigma_j^2 \boldsymbol{I}_c \otimes \boldsymbol{I}_r)$ at(Harv).id(Col).id(Row)	673.848	6	6	1077.376
TP2	$\text{diag}(\sigma_j^2 \boldsymbol{\Sigma}_{cj} \otimes \boldsymbol{\Sigma}_{rj})$ at(Harv).ar1(Col).ar1(Row)	693.421	18	6	1062.230
TP3	$\text{diag}(\sigma_j^2 \boldsymbol{\Sigma}_{cj} \otimes \boldsymbol{\Sigma}_{rj})$ at(Harv).ar1(Col).ar1(Row)	711.817	18	12	1037.438
TP4	$\sigma^2 \boldsymbol{\Sigma}_h \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r$ ar1v(Harv).ar1(Col).ar1(Row)	1161.726	4	12	109.620
TP5	$\boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r$ ar1h(Harv).ar1(Col).ar1(Row)	1168.007	9	12	107.058
TP6	$\boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r$ ar1h(Harv).ar1(Col).ar1(Row)	1170.350	9	14	103.196
TP7	$\boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r + \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r^{(p)} + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ ar1h(Harv).ar1(Col).ar1(Row) + id(Col).ar1v(Row) + Harv.Col.Row	1222.672	12	14	1.916
TP8	$\boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r + \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r^{(p)} + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ ar1h(Harv).id(Col).ar1(Row) + id(Col).ar1v(Row) + Harv.Col.Row	1222.445	11	14	0
TP9	$\boldsymbol{D\Sigma}_h^{exp}\boldsymbol{D} \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r + \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r^{(p)} + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ exph(Harv).ar1(Col).ar1(Row) + id(Col).ar1v(Row) + Harv.Col.Row	1222.267	12	14	3.248
TP10	$\boldsymbol{\Sigma}_h^{ante} \otimes \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r + \sigma_p^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r^{(p)} + \sigma_m^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$ ante(Harv).ar1(Col).ar1(Row) + id(Col).ar1v(Row) + Harv.Col.Row	1223.968	16	14	6.564

Table 7. Summary of genetic variance models fitted to persistence at Terry Hie Hie. Residual log-likelihoods (denoted by ℓ), number of parameters in G_g (par) and AIC values (given as differences from the best model) are presented for each model.

Model	Genetic model	ℓ	$\boldsymbol{G}_g$ par	AIC
TP11	diag(Harvest).Variety	1225.536	6	19.176
TP12	fa(Harvest,1).Variety	1237.473	16	15.302
TP13	fa(Harvest,2).Variety	1241.603	21	17.042
TP14	corh(1,years).id(Variety) + idv(Harvest.Variety)	1237.124	8	0

All models include the residual variance model TP8 of Table 6 ($\boldsymbol{R} = \boldsymbol{D\Sigma}_h\boldsymbol{D} \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r + \sigma_p{}^2 \boldsymbol{J}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{\Sigma}_r{}^{(p)} + \sigma_m{}^2 \boldsymbol{I}_h \otimes \boldsymbol{I}_c \otimes \boldsymbol{I}_r$) and have the overall mean over time modelled using the spline model : 1 + lin(years) + spl(years) + dev(years).

Table 8. REML estimates of genetic variances ×100 (on diagonal) and genetic correlations between harvests (off diagonals) from the fa2 model (TP13) fitted to the lucerne persistence data

Harvest	1	2	3	4	5	6
1	0.741					
2	0.842	0.805				
3	0.741	0.864	1.065			
4	0.343	0.625	0.885	1.324		
5	0.077	0.413	0.710	0.940	0.923	
6	–0.280	0.079	0.384	0.732	0.836	0.799

specific times or as a trend over time. The genetic models aim to provide a reduced set of genetic effects to enable varietal selection to take place. In some instances the data may warrant a random regression over time approach while in others a different approach, for example predictions may be required at each harvest time or for target groups of harvests. The factor analytic models enable predictions to be made at each harvest time, or for target groups of harvests which may be identified using cluster analysis (Cullis *et al.* 2010). The random regression models for genetic effects provide intercepts and slopes for each variety giving the deviation of the variety from the overall harvest mean profile. The random regression models allow predictions to be made at any time during the trial and also predictions of the time to a certain event.

Appropriate selection indices could be developed in collaboration with breeders using the genetic parameters from these models, and hence rankings of varieties presented for selection (Kelly *et al.* 2007). Smith *et al.* (2007) present selection indices based on a weighted sum of the predictions from each harvest time with user supplied weights. This approach can be used to form selection indices for the data in this paper,

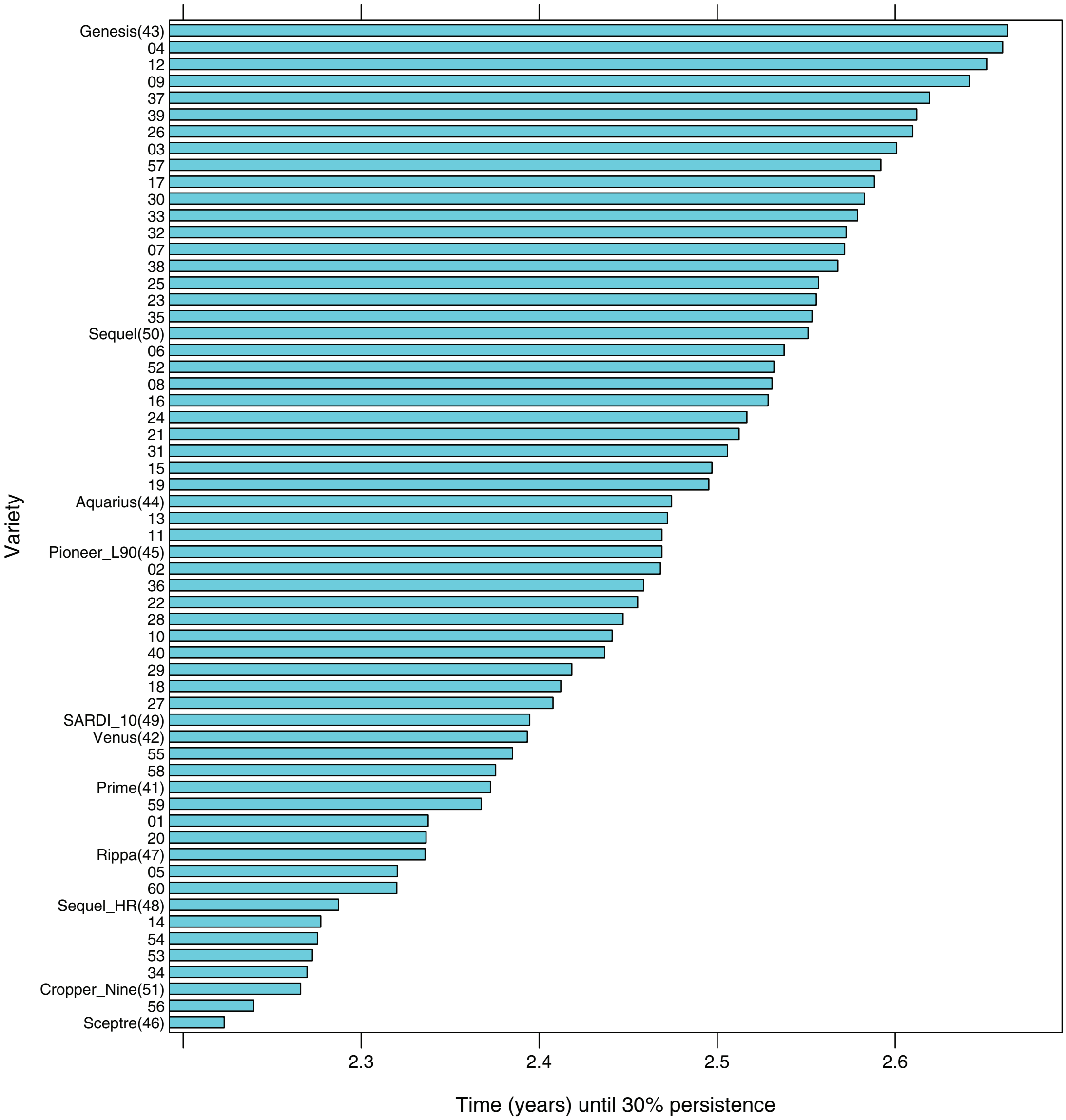

Fig. 5. Predicted time to persistence level of 30% (–0.838 on logit scale) for each variety based on Model TP14 in Table 7.

based on the predictions obtained at each harvest time, or for groups of harvests, from the factor analytic model as in Cullis *et al.* (2010).

Together the residual and genetic models proposed in this paper have the potential to have a significant impact on improving the accuracy of variety selections made in perennial crops.

Acknowledgements

The first author gratefully acknowledges a postgraduate research scholarship contributed by the New South Wales Department of Primary Industries and Fisheries and supplementary scholarships by the Division of Mathematical and Information Sciences, CSIRO, and the School of Agriculture, Food and Wine, The University of Adelaide. The authors acknowledge the financial support of Grains Research and Development Corporation and New South Wales Department of Primary Industries for the lucerne breeding program.

References

Apiolaza L, Gilmour A, Garrick D (2000) Variance modelling of longitudinal height data from a *Pinus radiata* progeny test. *Canadian Journal of Forest Research* **30**, 645–654.

Bartlett MS (1978) Nearest neighbour models in the analysis of field experiments (with discussion). *Journal of the Royal Statistical Society, Series B* **40**, 147–174.

Beeck CP, Cowling WA, Smith AB, Cullis BR (2010) Analysis of yield and oil from a series of canola breeding trials, part 1: Fitting factor analytic models with pedigree information. *Genome* **53**, 992–1001.

Bjornsson H (1978) Analysis of a series of long-term grassland experiments with autocorrelated errors. *Biometrics* **34**, 645–651.

Bouton J (2012) Breeding lucerne for persistence. *Crop & Pasture Science* **63**, 95–106.

Brien CJ, Bailey RA (2006) Multiple randomizations (with discussion). *Journal of the Royal Statistical Society, Series B* **68**, 571–609.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R, Reference Manual. Release 3. Queensland Department of Primary Industries and Fisheries, Toowoomba.

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

Cullis B, Gogel B, Verbyla A, Thompson R (1998) Spatial analysis of multi-environment early generation trials. *Biometrics* **54**, 1–18.

Cullis BR, Gleeson AC (1991) Spatial analysis of field experiments – an extension to two dimensions. *Biometrics* **47**, 1449–1460.

Cullis BR, Jefferson P, Thompson R, Smith AB (2014) Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. *Theoretical and Applied Genetics* **127**, 2193–2210.

Cullis BR, Smith AB, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part 11: Exploring variety by environment interaction using factor analysis. *Genome* **53**, 1002–1016.

De Faveri J (2013) Spatial and temporal modelling of perennial crop variety selection trials. PhD Thesis, The University of Adelaide, SA, Australia. http://digital.library.adelaide.edu.au/dspace/handle/2440/83114.

Diggle P (1988) An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–971.

Dutkowski G, Costa e Silva J, Gilmour A, Lopez G (2002) Spatial analysis methods for forest genetic trials. *Canadian Journal of Forest Research* **32**, 2201–2214.

Evans J, Roberts E (1979) Analysis of sequential observations with applications to experiments on grazing animals and perennial plants. *Biometrics* **35**, 687–693.

Gabriel KR (1962) Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics* **33**, 201–212.

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 269–273.

Gleeson AC, Cullis BR (1987) Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics* **43**, 277–288.

Green PJ, Silverman BW (1994) 'Nonparametric regression and generalised linear models.' (Chapman and Hall: London)

Hayward M, Delacey I, Tyler B, Drake D (1982) The application of pattern analysis for the recognition of adaptation in a collection of *Lolium multiflorum* populations. *Euphytica* **31**, 383–396.

Kaufman L, Rousseeuw PJ (1990) 'Finding groups in data: An introduction to cluster analysis.' (Wiley: New York).

Kelly A, Smith AB, Eccleston J, Cullis BR (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Science* **47**, 1063–1070.

Laird N, Ware J (1982) Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lodge G, Gleeson A (1984) A comparison of methods of estimating lucerne population for monitoring persistence. *Australian Journal of Experimental Agriculture and Animal Husbandry* **24**, 174–177.

Meyer K (1998) Estimating covariance functions for longitudinal data using a random regression model. *Genetics Selection Evolution* **30**, 221–240.

Meyer K, Kirkpatrick M (2005) Up hill, down dale: quantitative genetics of curvacoeus traits. *Philosophical Transactions of The Royal Society B* **360**, 1443–1455.

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 100–109.

Piepho HP, Eckl T (2014) Analysis of series of variety trials with perennial crops. *Grass and Forage Science* **69**, 431–440.

R Development Core Team (2012) 'R: A language and environment for statistical computing.' (R Foundation for Statistical Computing: Vienna, Austria)

Resende MDV, Thompson R, Welham S (2006) Multivariate spatial statistical analysis of longitudinal data in perennial crops. *Brazilian Journal of Mathematics and Statistics* **24**, 147–169.

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–51.

Schaeffer L (2004) Application of random regression models in animal breeding. *Livestock Production Science* **86**, 35–45.

Smith AB, Cullis BR, Thompson R (2001) Analysing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**, 1138–1147.

Smith AB, Stringer JK, Wei X, Cullis BR (2007) Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. *Euphytica* **157**, 253–266.

Smith K, Casler M (2004) Spatial analysis of forage grass trials across locations, years, and harvests. *Crop Science* **44**, 56–62.

Smith K, Kearney G (2002) Improving the power of pasture cultivar trials to discriminate cultivars on the basis of differences in herbage yield. *Australian Journal of Agricultural Research* **53**, 191–199.

Smith KF, Spangenberg G (2014) Forage breeding for changing environments and production systems: an overview. *Crop & Pasture Science* **65**, i–ii.

Stefanova KT, Smith AB, Cullis BR (2009) Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological, and Environmental Statistics* **14**, 392–410.

Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stringer JK, Cullis BR (2002) Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. *Australian Journal of Agricultural Research* **53**, 911–918.

Verbyla AP, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–312.

Appendix 1. Model functions and their associated correlation/covariance matrices that are used in analysis of the lucerne trial. $\boldsymbol{I_h}$ is the $\boldsymbol{h \times h}$ identity matrix and $\boldsymbol{J_h}$ is the $\boldsymbol{h \times h}$ matrix of 'ones'

Name	Model Function	Correlation or variance	Example	correlation or covariance matrix
identity	id	correlation	id(Harvest)	$\boldsymbol{I}_h$
identity variance	idv	variance	idv(Harvest)	$\sigma^2{}_h\boldsymbol{I}_h$[1]
diagonal	diag	variance	diag(Harvest)	diag (σ^2_{hj})[1,4]
heterogeneous correlation	corh	variance	corh(Harvest)	diag (σ_{hj}) $\{\boldsymbol{I}_h+\rho_t(\boldsymbol{J}_h-\boldsymbol{I}_h)\}$ diag (σ_{hj})[2,3]
unstructured	us	variance	us(Harvest)	$\boldsymbol{G}_h$[5]
factor analytic, order k	fa	variance	fa(Harvest,k)	$\Lambda_h\,\Lambda_h{}^T+\boldsymbol{\Psi}_h$[6]
autoregressive, order 1	ar1	correlation	ar1(Row)	$\boldsymbol{\Sigma}_r^{\mathrm{ar1}}=\boldsymbol{I}_r+\sum_{j=1}^{r\text{-}1}\rho^j{}_r\,\boldsymbol{F}_j$[7]
autoregressive variance, order 1	ar1v	variance	ar1v(Row)	$\sigma_r^2\,\boldsymbol{\Sigma}_r^{\mathrm{ar1}}$[1,3,7]
autoregressive heterogeneous variance, order 1	ar1h	variance	ar1h(Row)	$\boldsymbol{D}\boldsymbol{\Sigma}_r^{\mathrm{ar1}}\boldsymbol{D}$[1,3,7,8]
exponential heterogeneous variance	exph	variance	exph(HarvestTime)	$\boldsymbol{D}\boldsymbol{\Sigma}_h^{\mathrm{exp}}\boldsymbol{D}$[8] where $\boldsymbol{\Sigma}_h^{\mathrm{exp}}{}_{\mathrm{ij}}=\phi^{\lvert t_i\text{-}t_j\rvert}$
antedependence, order s	ante	variance	ante(HarvestTime,s)	$\boldsymbol{\Sigma}_h^{\mathrm{ante}}$ where $\boldsymbol{\Sigma}_h^{\mathrm{ante}\text{-}1}=\boldsymbol{U}^T\boldsymbol{D}^*\boldsymbol{U}$[9]
at	at	compound	at(Trait)	separate structure for each level of Trait

[1] $\sigma^2{}_h$, $\sigma^2{}_{hj}$ and $\sigma^2{}_r$ are variances.

[2] σ_{hj} are standard deviations.

[3] ρ_t, ρ_r and ϕ are correlations.

[4] diag () is a diagonal matrix with elements specified.

[5] $\boldsymbol{G}_t$ is a fully parameterized covariance matrix of order *t*.

[6] $\boldsymbol{\Lambda}_t$ is a matrix of factor loadings, $\boldsymbol{\Psi}_t$ is a diagonal matrix of order *t*.

[7] $\boldsymbol{F}_j$ is a matrix which has 'ones' on the *j*th sub and super-diagonals and zeros elsewhere.

[8] $\boldsymbol{D}$ is a diagonal matrix of standard deviations.

[9] $\boldsymbol{U}$ is a lower triangular matrix with 'ones' on the diagonal and $-u_{ij}$ on $(i\text{-}j)^{\mathrm{th}}$ sub-diagonal (where $(i\text{-}j\leq s)$ and zeroes elsewhere and where $Y_j=\sum_{k=1}^{s_j}u_{j(j-k)}Y_{j-k}+e_j$, with $s_j=min(j-1,\ s)$. D* is a diagonal matrix.