A Bayesian Approach to Regional and Local-Area Prediction From Crop Variety Trials

Chris M. THEOBALD, Mike TALBOT, and Fabian NABUGOOMU

The inclusion of covariates in models for analyzing variety × environmental data sets allows the estimation of variety yields for specific locations within a region as well as for the region as a whole. Here we explore a Bayesian approach to the estimation of such effects and to the choice of variety using a possibly incomplete variety × location × year data set that includes location × year covariates. This approach allows expert knowledge of the crop and uncertainty about local circumstances to be incorporated in the analysis. It is implemented using Markov chain Monte Carlo simulation. An example is used to illustrate the approach and investigate its robustness.

Key Words: Bayesian inference; Decision theory; Local-area estimation; Markov chain Monte Carlo; Mixed-effects model; Residual maximum likelihood; Variety by environment data.

1. INTRODUCTION

New crop varieties are evaluated by testing them alongside established varieties in trials distributed throughout the region in which they are to be eventually grown by farmers. Typically, the results from such trials are combined to produce measures of average variety performance. These averages are the basis on which varieties are recommended for growing in the whole region. However, the regions can be heterogeneous, for example with differing levels of soil fertility or varying in earliness of season, and varieties will often respond differentially to such covariates. As a consequence, reliance on over-trials averages may not be most efficient when selecting varieties in all parts of the region. The following question then arises: If farmers can identify, with some level of uncertainty, the position of

Chris M. Theobald is Lecturer in the Department of Mathematics and Statistics, University of Edinburgh, The King's Buildings, Edinburgh, Scotland EH9 3JZ, U.K. (E-mail: cmt@maths.ed.ac.uk). Mike Talbot was Head of Technology at Biomathematics and Statistics Scotland, The King's Buildings, Edinburgh, Scotland EH9 3JZ, U.K. (E-mail: mike@bioss.ac.uk). Fabian Nabugoomu is Senior Lecturer in the Department of Mathematics, Makerere University, P.O. Box 7062 Kampala, Uganda (E-mail: fnabugoomu@yahoo.com).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 3, Pages 403–419
DOI: 10.1198/108571102230

their farms on the scales on which heterogeneity is measured, can we then use trials data to provide a local-area predictor of likely variety performance in their conditions that is better than the regional performance average?

The use of environmental covariates to analyze variety × environment data sets in order to estimate relative variety performance in target locations is considered by several workers, for example, Freeman and Perkins (1971), Hardwick and Wood (1972), Wood (1976), van Eeuwijk, Denis, and Kang (1996), and Piepho, Denis, and van Eeuwijk (1998). Their approaches largely concentrate on dealing with complete variety × environment data sets with only a single error component, though Piepho et al. (1998) include random effects for locations and years and also allow several quantitative and qualititative covariates.

This article proposes a Bayesian approach to the more typical situation of a variety × location × year data set that is incomplete and includes one or more location × year covariates. This approach incorporates expert knowledge of the crop and takes account of possible uncertainty in the values of the covariates at the location where the crop is to be grown.

Section 2 introduces data from a series of maize variety trials that include covariate information. Section 3 considers models for variety yields in multilocation trials that allow for such information. Section 4 describes our Bayesian approach to these problems. Section 5 illustrates its application to the maize data and examines its robustness. Section 6 considers some of the issues involved in applying Bayesian methods to such data.

2. AN EXAMPLE

Table 1 shows the silage dry-matter yields of 10 varieties of maize (*Zea mays*) grown in the maritime provinces of eastern Canada in the years 1990–1994. Trials were carried out in seven districts, but different fields were used in successive years. For simplicity of exposition, we treat the locations as nested within years rather than use the full variety × year × district classification. For each trial, there is also available a covariate, corn heat units (CHU). This represents accumulated average daily temperatures recorded during the growing season and is expected to influence the performance of varieties (Kiniry, Ritchie, Musser, Flint, and Iwig 1983); recorded CHU values range from 2.31 to 2.77 thousands of °C. The yield data are incomplete since some varieties were not sown in all years and trials were not conducted in 6 of the 35 possible combinations of year and location.

An important characteristic of CHU is that its value is not known when the choice of variety is to be made for a particular location. In contrast, covariates such as soil type or altitude may be accurately known at this time. We regard the CHU values in Table 1 as being measured without error, but inferences about future yields and variety choices have to be based on predictions of CHU; these inferences should reflect in a realistic way the uncertainty in the prediction of CHU.

Table 1. Maize Variety Yields (t/ha) and Corn Heat Units ('000 °C) for Each Trial

		Variety										
Year	*District*	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*	*V7*	*V8*	*V9*	*V10*	*CHU*
1990	1	6.27	6.62	6.07	7.20	6.65	6.71	6.54	8.38	—	—	2.57
	2	5.57	6.95	5.42	7.31	6.82	5.86	6.04	6.37	—	—	2.53
	3	8.45	8.19	6.84	8.19	7.18	8.36	8.85	9.80	—	—	2.72
	4	7.35	6.43	6.59	7.16	6.24	6.43	7.44	7.24	—	—	2.72
	7	6.50	7.09	5.99	7.35	6.72	7.15	6.85	6.88	—	—	2.48
1991	1	6.71	6.73	6.04	7.11	—	6.58	6.73	7.33	6.80	7.62	2.44
	2	5.59	5.01	5.21	5.74	—	5.18	5.42	4.66	5.25	6.00	2.55
	3	8.36	7.62	7.43	7.47	—	8.31	6.88	9.35	7.67	7.43	2.75
	4	7.25	6.70	7.46	7.53	—	6.69	6.90	7.40	7.72	7.22	2.75
	5	8.09	8.45	6.73	8.92	—	7.22	7.54	7.82	7.47	7.29	2.61
	6	6.66	6.97	6.43	7.07	—	7.04	6.64	6.13	5.96	7.28	2.50
	7	5.68	5.53	5.08	6.32	—	5.73	5.21	3.42	5.89	5.48	2.39
1992	1	5.09	6.05	5.24	5.37	5.70	4.86	6.50	5.94	5.79	6.37	2.37
	2	5.93	6.05	5.37	5.89	5.75	5.40	6.24	4.98	5.61	6.58	2.45
	3	7.63	8.07	5.96	6.71	7.94	7.04	8.63	7.89	6.25	7.66	2.64
	4	8.02	7.99	6.84	7.89	7.79	7.37	8.60	9.10	7.81	8.70	2.64
	5	6.69	6.99	6.04	6.62	6.80	5.73	6.79	6.22	6.62	6.78	2.51
	6	7.14	7.53	6.15	6.90	6.86	7.27	7.46	6.60	6.83	6.91	2.39
1993	3	6.82	6.75	4.83	6.47	6.11	6.27	6.26	—	6.08	5.24	2.63
	4	5.69	5.47	5.74	5.62	5.04	5.48	5.16	—	5.96	6.60	2.63
	5	5.18	5.55	4.27	5.74	5.41	4.41	4.53	—	4.65	5.60	2.46
	6	4.88	4.72	3.70	4.40	4.26	4.05	4.15	—	4.17	3.63	2.45
	7	6.20	6.09	5.79	6.62	5.86	6.00	6.32	—	5.78	6.21	2.31
1994	1	6.67	8.28	5.11	7.47	7.61	7.05	—	—	5.81	8.75	2.65
	2	3.35	3.64	3.14	4.82	4.12	3.54	—	—	4.19	5.12	2.51
	3	7.48	7.05	5.50	7.48	7.35	7.63	—	—	5.87	6.46	2.77
	5	6.41	6.90	6.44	7.78	6.28	6.51	—	—	5.66	7.62	2.64
	6	6.76	6.46	5.25	6.93	5.50	6.56	—	—	5.29	7.61	2.58
	7	6.47	6.70	5.57	7.70	6.80	6.13	—	—	6.86	6.75	2.49

3. MODELS FOR VARIETY YIELDS

3.1 REGIONAL MODEL

Let us first consider the problem of modeling variety yields from a single year's trials within a region. Analysis of the individual trials provides a table of means, classified by varieties and locations, that may be incomplete. If y_{ij} denotes the mean yield for the ith of v varieties at the jth of l locations, then an additive model for the data is

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = \alpha_i + \theta_j + \varepsilon_{ij} \qquad (i = 1, \ldots, v;\ j = 1, \ldots, l), \tag{3.1}$$

where the systematic part μ_{ij} of y_{ij} is the sum of effects α_i and θ_j for variety i and location j, while the error terms ε_{ij} are assumed to be independent with a normal distribution $\mathrm{N}(0, \sigma^2)$. Variety effects are usually treated as fixed in a frequentist analysis on the grounds that the varieties are individuals with distinct identities rather than a sample representative of the species. In contrast, the location effects $\theta_1, \theta_2, \ldots, \theta_l$ may be assumed to form a random

sample from $N(0, \sigma_L^2)$, independent of the ε_{ij}. Thus, we consider a model with variance components σ_L^2 and σ^2.

To generalize model (3.1) to a series of trials carried out over several years, we might replace the term `locations' by `environments,' the latter describing a combination of years and locations. However, it is more efficient and informative to take full account of the structure of the data. So, if y_{ijk} denotes the mean yield for variety i at location j in the kth of m years and if locations are treated as nested within years, then (3.1) can be extended to the model

$$\begin{aligned} y_{ijk} &= \mu_{ijk} + \varepsilon_{ijk} \\ &= \alpha_i + \gamma_k + \delta_{j(k)} + \eta_{ik} + \varepsilon_{ijk} \qquad (i = 1, \ldots, v;\ j = 1, \ldots, l;\ k = 1, \ldots, m), \end{aligned} \tag{3.2}$$

where γ_k denotes the effect of year k, $\delta_{j(k)}$ the effect of location j within year k, η_{ik} the interaction of variety i with year k, and ε_{ijk} an error term. The γ_k, $\delta_{j(k)}$, and η_{ik} are assumed to be independent of each other and of the ε_{ijk}, with the distributions $N(0, \sigma_Y^2)$, $N(0, \sigma_{LY}^2)$, and $N(0, \sigma_{VY}^2)$ (see Patterson 1997).

If the same locations are used for several years, we have a crossed classification with effects for years and locations and for variety $\times$ year, variety $\times$ location, and location $\times$ year interactions, all taken as random. Often, in practice, different locations are chosen for some trials each year while for others the same locations are used repeatedly. To complicate matters further, trials conducted at the same nominal location may not be in the same fields. In what follows, we treat locations as nested within years.

Variety trial data are usually incomplete or unbalanced for several reasons:

- the numbers of trials vary from year to year,
- new varieties are introduced each year while others are dropped,
- varieties may be sown only at a subset of locations.

Thus, it is essential that estimation procedures cope with unbalanced data sets.

3.2 LOCAL-AREA MODEL INCORPORATING COVARIATE INFORMATION

For many species, there are substantial variety $\times$ environment interactions (see Talbot 1984). If environmental covariates (such as soil fertility or temperature) are available for trial locations and for the target location where a variety is to be grown, then such information can be included in a local-area model in order to improve decisions on variety choice.

Let $\mathbf{x}_{jk}$ denote a p-vector of covariate values recorded at location j in year k. Model (3.2) can then be modified to include dependence on these as

$$y_{ijk} = \mu_{ijk} + \varepsilon_{ijk} = \alpha_i + \beta_i^{\mathrm{T}}(\mathbf{x}_{jk} - \mathbf{x}) + \gamma_k + \delta_{j(k)} + \eta_{ik} + \varepsilon_{ijk}, \tag{3.3}$$

where β_i is the p-vector of regression coefficients associated with variety i and $\mathbf{x}$ is the mean covariate vector. Separate coefficient vectors allow for differences between varieties in sensitivity to changes in environmental conditions.

4. BAYESIAN MODELING OF VARIETY YIELDS

4.1 HIERARCHICAL MODELS

In a Bayesian formulation, expert opinion on the likely values of model parameters is expressed by treating them as random variables and specifying a prior probability distribution for them. A frequent objection to such a formulation is that a prior distribution cannot always be adequately specified. However, in the case of crop variety performance testing, there is usually considerable experience of the magnitudes of mean yields and of variance components. Because our intention is to use expert knowledge, improper prior distributions (supposedly representing ignorance about parameter values) are not considered here.

The normal distribution assumptions, already stated for the random effects $\theta_j, \gamma_k, \delta_{j(k)}$, and η_{ik} in models (3.1), (3.2), and (3.3), form part of the specification of corresponding Bayesian models. These models are hierarchical in that the distributions of some of the unknown parameters are defined in terms of other parameters; for example, the θ_j in (3.1) are assumed independent with common distribution $\mathrm{N}(0, \sigma_L^2)$ depending on the unknown σ_L.

In contrast with the conventional frequentist analysis, we take the variety effects α_i to form a random sample from a normal distribution $\mathrm{N}(\mu_V, \sigma_V^2)$, independent of the errors ε_{ij} and the location effects θ_j. For model (3.3), the vectors of regression coefficients β_i are also assumed normal and independent of the α_i. Although these parameters are necessarily treated as random variables in a Bayesian formulation, there are other prior assumptions that could be made about them. Different formulations might be used if, for example,

- subsets of the varieties were known to be genetically similar,
- more precise information were available on control varieties than on others,
- some varieties were known to be more resistant to disease than others.

To complete the specification of the prior distribution, we take μ_V to be from a distribution $\mathrm{N}(m_V, r_V^2)$ and σ^{-2}, σ_V^{-2}, and σ_L^{-2} to be mutually independent with scaled χ^2 distributions. Past experience of variety trial data may be used to specify the quantities m_V and r_V by considering percentiles of the distribution of μ_V; the prior distributions for the three (inverse) variances can be specified by giving a prior estimate of each variance and corresponding degrees of freedom. For example, an estimate s^2 of σ^2 with d d.f. corresponds to $ds^2\sigma^{-2}$ having the distribution $\chi^2(d)$, so the prior expectation of σ^{-2} is s^{-2} and d measures the precision of this estimate. Equivalently, if $\mathrm{Ga}(a, b)$ denotes a gamma distribution with probability density function $b^a z^{a-1} e^{-bz}/\Gamma(a)$ $(z > 0)$, then σ^{-2} has the distribution $\mathrm{Ga}(d/2, ds^2/2)$. Similarly, estimates s_V^2, s_L^2, and degrees of freedom d_V and d_L are required to specify the prior distributions of σ_V^{-2} and σ_L^{-2}. The quantities required to define the joint prior distribution for model (3.1) are thus m_V, r_V^2, s^2, s_V^2, s_L^2, d, d_V, and d_L; the convention of using Roman symbols for this purpose is followed for other models.

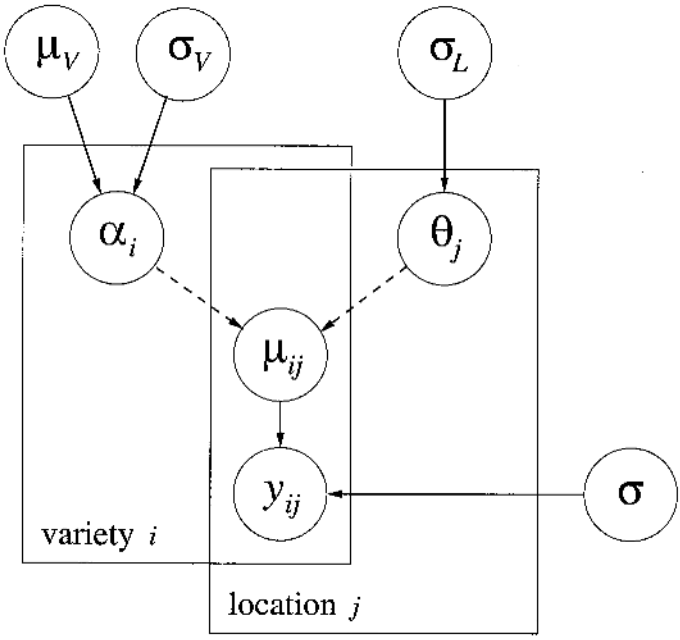

Figure 1. Directed Acyclic Graph for Model (3.1).

It is useful for both interpretation and computation to represent each Bayesian model using a directed acyclic graph: Figure 1 shows such a graph for model (3.1). The nodes of the graph (represented by circles) correspond to the unknown parameters in the model and the yield data (along with any covariates). The inclusion of nodes within a rectangle indicates that they are repeated over the levels of the factor named at the foot of the rectangle, such as the variety and location classifications. The arrows between nodes (each leading from a parent node to a child) indicate the relationships between them: solid and dashed arrows show stochastic and deterministic dependence, respectively. Thus, for example, the distribution of the α_i is defined in terms of μ_V and σ_V and the value of μ_{ij} is determined from α_i and θ_j. The graph is acyclic because there is no way to return to any node by following the directions of the arrows.

4.2 IMPLEMENTATION

The use of Markov chain Monte Carlo (MCMC) methods for investigating posterior and predictive distributions is too widespread to require more than a brief introduction here; Gilks, Richardson, and Spiegelhalter (1996) and Brooks (1998) provide accessible accounts of MCMC methods. To implement the calculations required for prediction of variety performance, we use Gibbs sampling via the BUGS package (Spiegelhalter, Thomas, Best, and Gilks 1996, 1997), which is freely available from http://www.mrc-bsu.cam.ac.

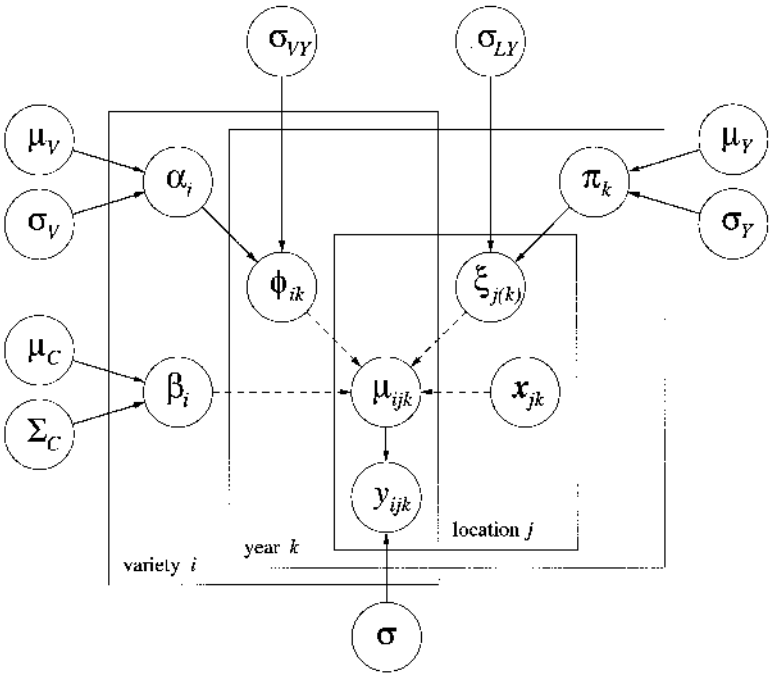

Figure 2. Directed Acyclic Graph for Model (4.1).

uk/bugs. The user of BUGS defines the model to be fitted by specifying the conditional distribution for each stochastic node in a directed acyclic graph and the formula for each deterministic one.

For complex models, there may be several choices of parameterization available with possibly different convergence properties. Model (3.3) expresses the systematic part μ_{ijk} of y_{ijk} as a sum of terms for variety, regression, year, location (within year), and variety $\times$ year interaction. An alternative parameterization is

$$\mu_{ijk} = \beta_i^{\mathrm{T}} (\mathbf{x}_{jk} - \mathbf{x}) + \xi_{j(k)} + \phi_{ik}. \tag{4.1}$$

In using (4.1), it may appear that we are fitting interactions without the corresponding main effects, but the effects for varieties and years are included in the hierarchical definition of the parameters and their joint prior distribution; this definition is as follows (the notation $\sim$ is to be interpreted as `is distributed as' or `are distributed independently as,' according to context):

$$\left.\begin{array}{lll} \beta_i \sim \mathrm{N}_p(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C) & \boldsymbol{\mu}_C \sim \mathrm{N}_p(\mathbf{m}_C, \mathbf{R}_C) & \boldsymbol{\Sigma}_C^{-1} \sim W_p(d_C, d_C^{-1}\mathbf{S}_C^{-1}) \\ \xi_{j(k)} \sim \mathrm{N}(\pi_k, \sigma_{LY}^2) & \pi_k \sim \mathrm{N}(\mu_Y, \sigma_Y^2) & \mu_Y \sim \mathrm{N}(0, r_Y^2) \\ \phi_{ik} \sim \mathrm{N}(\alpha_i, \sigma_{VY}^2) & \alpha_i \sim \mathrm{N}(\mu_V, \sigma_V^2) & \mu_V \sim \mathrm{N}(m_V, r_V^2) \\ ds^2\sigma^{-2} \sim \chi^2(d) & d_V s_V^2 \sigma_V^{-2} \sim \chi^2(d_V) & d_Y s_Y^2 \sigma_Y^{-2} \sim \chi^2(d_Y) \\ d_{VY} s_{VY}^2 \sigma_{VY}^{-2} \sim \chi^2(d_{VY}) & d_{LY} s_{LY}^2 \sigma_{LY}^{-2} \sim \chi^2(d_{LY}) & \end{array}\right\}. \tag{4.2}$$

Figure 2 shows the corresponding directed acyclic graph; the nesting of locations within

years is shown by the inclusion of the rectangle for locations inside that for years.

In model (4.1), the definitions of the $\xi_{j(k)}$ relative to the π_k and of the ϕ_{ik} relative to the α_i are forms of hierarchical centering, shown by Gelfand, Sahu, and Carlin (1995) to improve the convergence of MCMC; note that the year effects π_k differ from the γ_k by being centered on μ_Y rather than on zero. With a single covariate, $\mathbf{\Sigma}_C$ reduces to σ_C^2 and $d_C s_C^2 \sigma_C^{-2}$ has the prior distribution $\chi^2(d_C)$.

4.3 PREDICTION FOR A TARGET LOCATION IN A FUTURE YEAR

To compare potential crop yields at a target location in another year, we consider the posterior predictive distribution of the yields, i.e., their joint distribution given the data after integrating out any unknown parameters. To infer this predictive distribution (or any of its marginal distributions), we add nodes for these yields and for any necessary ancestors that are not already in the directed acyclic graph. Using a subscript asterisk to denote variables specific to the target location, the potential yields are denoted by $y_{1*}, \ldots, y_{v*}$, and the additional ancestors under model (4.1) are $\mathbf{x}_*$, π_*, ξ_*, the ϕ_{i*}, and the μ_{i*}. The distribution of the covariate vector $\mathbf{x}_*$ must be specified if some or all of its elements are not accurately known, for example if they are weather measurements for the coming season. The variables π_*, ξ_*, ϕ_{i*}, and y_{i*} are given distributions $\mathrm{N}(\mu_Y, \sigma_Y^2)$, $\mathrm{N}(\pi_*, \sigma_{LY}^2)$, $\mathrm{N}(\alpha_i, \sigma_{VY}^2)$, and $\mathrm{N}(\mu_{i*}, \sigma^2)$, respectively, and μ_{i*} has parents $\mathbf{x}_*$, ξ_*, β_i, and ϕ_{i*}.

The posterior predictive distributions of the y_{i*} under model (4.1) may be estimated by including the additional nodes in the MCMC analysis. If a succession of target locations is to be considered (e.g., as part of an advisory service for farmers) then a simpler alternative to repeating the analysis for each one would be to retain the values of μ_Y, σ_Y, σ_{LY}, σ_{VY}, σ, and the α_i and β_i from a suitably large MCMC sample and to simulate values of the μ_{i*} and hence the y_{i*} for the given distribution of $\mathbf{x}_*$. Thus, if superscript (r) denotes the rth iteration in this sample and $\mathbf{x}_*$ is assumed to have distribution $\mathrm{N}(\mathbf{m}_\mathbf{x}, \mathbf{S}_\mathbf{x})$ at the new location, then for any r, the simulated $\mu_{i*}^{(r)}$ are jointly normally distributed with expectations $\alpha_i^{(r)} + \beta_i^{(r)\mathrm{T}}(\mathbf{m}_\mathbf{x} - \mathbf{x}) + \mu_Y^{(r)}$, variances $\beta_i^{(r)\mathrm{T}} \mathbf{S}_\mathbf{x} \beta_i^{(r)} + \sigma_Y^{2(r)} + \sigma_{LY}^{2(r)} + \sigma_{VY}^{2(r)}$ $(i = 1, \ldots, v)$, and covariances $\beta_h^{(r)\mathrm{T}} \mathbf{S}_\mathbf{x} \beta_i^{(r)} + \sigma_Y^{2(r)} + \sigma_{LY}^{2(r)}$ $(h \neq i)$.

Assessments of likely variety performance at the new location can be made by comparing the posterior distributions of the μ_{i*} or of the y_{i*}, particularly their posterior expectations; we denote these by $\mathrm{E}(\mu_{i*}|D)$, using D to represent the trial data. These expectations may be approximated by the means over iterations of the $\mu_{i*}^{(r)}$. Differences between the posterior expectations might be assessed by using credible intervals or by comparing the posterior expected differences with their posterior standard errors.

Although the posterior standard error of any variety difference can be found, the number of these differences, $v(v-1)$, increases rapidly with the number of varieties. It may be convenient to calculate the average posterior variance by defining the mean μ_* of the μ_{i*} and using the identity

$$\frac{1}{v(v-1)} \sum_{h \neq i} \mathrm{var}(\mu_{h*} - \mu_{i*} \mid D) \equiv \frac{2}{v-1} \left\{ \sum_i \mathrm{var}(\mu_{i*} \mid D) - v\, \mathrm{var}(\mu_* \mid D) \right\}. \quad (4.3)$$

The average posterior variance of the differences between the yields y_{i*} themselves (rather than between the μ_{i*}) is given by adding $2\mathrm{E}(\sigma^2 \mid D)$ to (4.3).

We can avoid carrying out any additional simulations for a new location if we are content to have only the first and second posterior moments of the μ_{i*} and the y_{i*}. These moments may be calculated from those of α_i, β_i, μ_Y, σ^2_{VY}, and σ^2 by exploiting the linearity of model (4.1). If the new vector of covariates $\mathbf{x}_*$ has expectation $\mathbf{m_x}$ and variance matrix $\mathbf{S_x}$ (and is uncorrelated with the unknown parameters), then the posterior expectation of μ_{i*} (and of y_{i*}) may be calculated as

$$\mathrm{E}(\mu_{i*} \mid D) = \mathrm{E}(\alpha_i \mid D) + \mathrm{E}(\beta_i \mid D)^{\mathrm{T}}(\mathbf{m_x} - \mathbf{x}) + \mathrm{E}(\mu_Y \mid D). \tag{4.4}$$

Formulas can also be derived for the posterior variances and covariances of the μ_{i*} at a new location and for the average posterior variance of the variety differences. Introducing the means α of the α_i and β of the β_i, the right-hand side of (4.3) is shown in the Appendix to equal

$$\begin{aligned}\frac{2}{v-1}\Bigg(& \sum_i \mathrm{var}(\alpha_i \mid D) - v\,\mathrm{var}(\alpha \mid D) \\ & + 2\,(\mathbf{m_x} - \mathbf{x})^{\mathrm{T}} \left\{ \sum_i \mathrm{cov}(\beta_i, \alpha_i \mid D) - v\,\mathrm{cov}(\beta, \alpha \mid D) \right\} \\ & + \mathrm{tr}\left[\left\{ (\mathbf{m_x} - \mathbf{x})\,(\mathbf{m_x} - \mathbf{x})^{\mathrm{T}} + \mathbf{S_x} \right\} \left\{ \sum_i \mathrm{var}(\beta_i \mid D) - v\,\mathrm{var}(\beta \mid D) \right\} \right] \\ & + \mathrm{tr}\left[\mathbf{S_x} \left\{ \sum_i \mathrm{E}(\beta_i \mid D)\,\mathrm{E}(\beta_i \mid D)^{\mathrm{T}} - v\,\mathrm{E}(\beta \mid D)\,\mathrm{E}(\beta \mid D)^{\mathrm{T}} \right\} \right] \Bigg) \\ & + 2\mathrm{E}(\sigma^2_{VY} \mid D). \end{aligned} \tag{4.5}$$

A reexpression of the regional model (3.2) analogous to (4.1) is

$$\mu_{ijk} = \xi_{j(k)} + \phi_{ik}. \tag{4.6}$$

Omitting terms relating to β_i and β from (4.4) and (4.5), the posterior expected yields and the average posterior variance of their differences are given for model (4.6) by

$$\mathrm{E}(\mu_{i*} \mid D) = \mathrm{E}(\alpha_i \mid D) + \mathrm{E}(\mu_Y \mid D) \tag{4.7}$$

and

$$\frac{2}{v-1}\left\{ \sum_i \mathrm{var}(\alpha_i \mid D) - v\,\mathrm{var}(\alpha \mid D) \right\} + 2\mathrm{E}(\sigma^2_{VY} \mid D). \tag{4.8}$$

5. BAYESIAN ANALYSIS OF THE MAIZE DATA

We now apply the methods of Section 4 to the maize variety data introduced in Section 2. The results shown for the regional and local-area models are based on 1-in-10 samples

Table 2. Parameter Values Defining the Prior Distributions for Regional Model (4.6)

		Variance				Degrees of freedom		
Parameter	*Mean*	*Expert*	*Diffuse*	*V diffuse*	*Estimate*	*Expert*	*Diffuse*	*V diffuse*
μ_V	6.0	0.05	0.05	0.25	—	—	—	—
μ_Y	0.0	0.25	0.25	1.25	—	—	—	—
σ^2_V	—	—	—	—	0.25	50	10	0.2
σ^2_Y	—	—	—	—	0.25	10	2	0.2
σ^2_{VY}	—	—	—	—	0.05	50	10	0.2
σ^2_{LY}	—	—	—	—	1.00	50	10	0.2
σ^2	—	—	—	—	0.20	50	10	0.2

from 100,000 Gibbs iterates following a burn-in of 10,000 iterates. Files of BUGS code and the corresponding data for the two models are available from the first author.

Joint prior distributions of the parameters in the regional model (4.6) are defined by lines 2 to 5 of (4.2) and the values given in Table 2. The means for the expectation parameters, the estimates of variance components, and the variances and degrees of freedom under the heading `Expert' are based on experience with the crop in general and long-term estimates of components of genotype × environment variation, in particular. The few degrees of freedom for years reflects the relatively short-term nature of our experience. The degrees of freedom under the heading `Diffuse' are all reduced by a factor of five, allowing one aspect of the robustness of the analysis to be investigated. Much less informative priors are also considered (with parameters defined under the headings `V diffuse'); for these, the prior variances of expectation parameters are increased by a factor of five and the degrees of freedom for the variance components all equal 0.2. Anticipating poorer mixing of the Markov chain with this prior, we use 1-in-100 samples from 1,000,000 iterates.

Table 3 gives the expected posterior predictive yields for a future year (and the deviations from their mean) under the regional model. The ordering of the varieties is by the expected yields under the three prior distributions. The second and third rows show the expected yields under the more diffuse prior distributions. The changes in the prior cause most of the expected posterior predictive yields to be pulled further toward the prior expected value of 6.0, but the deviations of the expected yields from their means—and hence comparisons between varieties—are scarcely affected.

Table 3 also shows residual maximum likelihood (REML) estimates of the variety effects (and their deviations) when these effects are taken to be fixed. The REML estimates give the same ranking of the varieties, but they are more widely distributed.

Table 4 gives the prior distributions used for the parameters of the local-area model defined in (4.1) and (4.2). Because of our lack of experience of the local-area analysis, we use the same estimates as in Table 2 for the prior distributions of σ^2_V, σ^2_Y, σ^2_{LY}, σ^2_{VY}, and σ^2, but fewer degrees of freedom are specified for the latter two variance components. As with the regional model, the effects of using more diffuse prior distributions are investigated by dividing all the degrees of freedom for the variance components by five and also by reducing

Table 3. Expected Posterior Predictive Yields and REML Estimates of Variety Effects (With Deviations From Their Means) for the Regional Model

	Variety									
	V4	*V10*	*V8*	*V2*	*V7*	*V1*	*V5*	*V6*	*V9*	*V3*
					Means					
Posterior expectations										
Expert prior	6.68	6.67	6.53	6.52	6.44	6.40	6.28	6.20	6.11	5.67
Diffuse variance prior	6.67	6.65	6.52	6.51	6.43	6.39	6.27	6.19	6.11	5.68
Very diffuse prior	6.72	6.71	6.57	6.57	6.50	6.46	6.36	6.28	6.20	5.81
REML estimates	6.81	6.80	6.66	6.63	6.54	6.50	6.37	6.28	6.18	5.71
					Deviations from means					
Posterior expectations										
Expert prior	0.33	0.32	0.18	0.17	0.09	0.05	−0.07	−0.15	−0.24	−0.68
Diffuse variance prior	0.33	0.31	0.18	0.17	0.09	0.05	−0.07	−0.15	−0.24	−0.66
Very diffuse prior	0.30	0.29	0.16	0.15	0.08	0.04	−0.06	−0.14	−0.22	−0.61
REML estimates	0.36	0.35	0.21	0.18	0.09	0.05	−0.08	−0.17	−0.27	−0.74

all of them to 0.2 and increasing the prior variances of expectation parameters by a factor of five.

Recall from Section 2 that inferences about future yields at a target location must be based on values of CHU for that location and growing season that are only predictions. We therefore specify the expected value, m_x, and standard deviation, s_x, of the predicted CHU at that site.

Figure 3 shows the expected posterior predictive yields for the 10 varieties as functions of the expected CHU, emphasizing variety V8. This variety (which was grown only in the first 3 years) shows a markedly higher increase in yield with CHU than the others; it appears to have a considerable yield advantage when sown in areas with high CHU. The vertical lines at the bottom of the plot are meant to facilitate comparisons between varieties; their heights are equal to the average posterior standard deviations of differences in variety yields and of differences in expected variety yields (as given by expression (4.5)) assuming that the prediction of CHU has standard deviation 0.075.

Table 4. Parameter Values Defining the Prior Distribution for Local-Area Model (4.1)

		Variance				*Degrees of freedom*		
Parameter	*Mean*	*Expert*	*Diffuse*	*V diffuse*	*Estimate*	*Expert*	*Diffuse*	*V diffuse*
μ_C	5.0	0.25	0.25	1.25	—	—	—	—
μ_V	6.0	0.05	0.05	0.25	—	—	—	—
μ_Y	0.0	0.25	0.25	1.25	—	—	—	—
σ^2_C	—	—	—	—	4.00	20	4	0.2
σ^2_V	—	—	—	—	0.25	50	10	0.2
σ^2_Y	—	—	—	—	0.25	10	2	0.2
σ^2_{VY}	—	—	—	—	0.05	25	5	0.2
σ^2_{LY}	—	—	—	—	1.00	50	10	0.2
σ^2	—	—	—	—	0.20	25	5	0.2

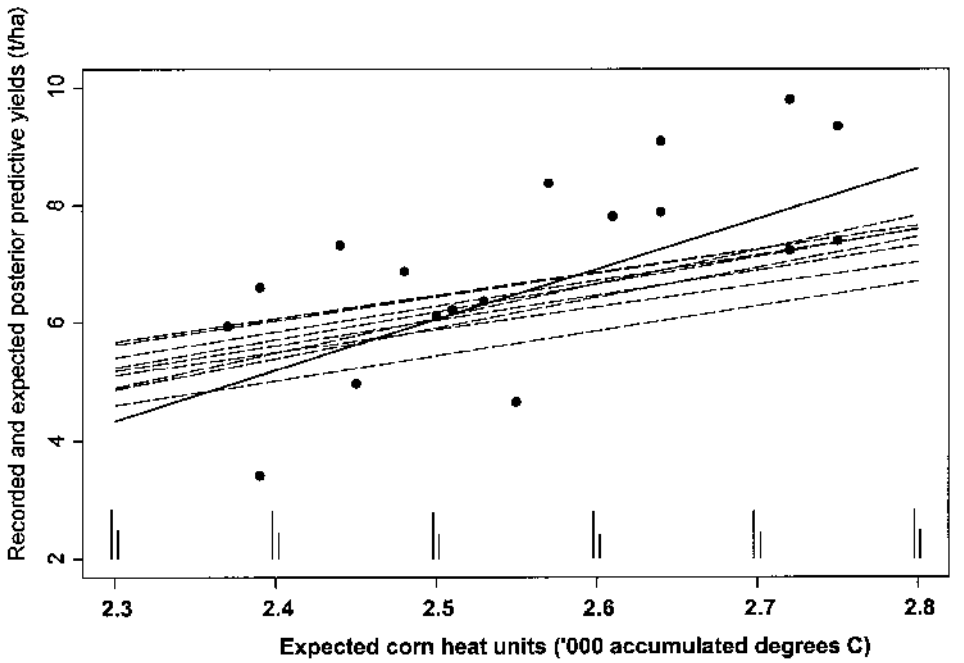

Figure 3. Dependence of Expected Posterior Predictive Yields (Based on Expert's Prior) From Local-Area Model (4.1) on Expected CHU for Canadian Maize Variety Trials. Average posterior standard deviations of variety differences and expected variety differences are indicated at the foot of the plot assuming standard deviation 0.075 for the prediction of CHU. Predictions for variety V8 are shown with a filled line, and the data points are shown only for this variety.

Table 5 compares posterior expectations and REML estimates for a future year under the local-area model at low and high values of expected CHU; the varieties are again ordered by their expected posterior predictive yields under the regional model. A further aspect of the robustness of the Bayesian method is investigated here, i.e., the model is extended to allow the slopes β_i and intercepts α_i to be correlated by changing the expectation of β_i from μ_C to $\mu_C + \gamma(\alpha_i - \mu_V)$ and giving the regression coefficient γ prior distribution $\mathrm{N}(0, 1)$, which is fairly diffuse in this context, while the remaining parameters follow the `Expert' prior distribution.

As is already apparent from Figure 3, the rankings of the varieties depend strongly on expected CHU. The five methods shown in Table 5 give identical rankings when CHU equals 2.4, and the posterior expectations give the same rankings for CHU equal to 2.7; the REML estimates give slightly different rankings for the higher CHU value and are generally more variable and slightly higher.

Table 6 shows how the assumed variation in CHU contributes to the average posterior standard deviations of differences (SEDs) between variety yields and between variety effects when the expert prior is assumed. These SEDs are also robust to the changes in the prior distribution considered above: using the diffuse prior gives values (not shown) within 1% of those for the expert prior, and the very diffuse prior gives values within 5%.

Table 5. Expected Posterior Predictive Yields and REML Estimates of Variety Effects for the Local-Area Model for Two Expected Values of CHU

Expected CHU		Variety									
		V4	*V10*	*V8*	*V2*	*V7*	*V1*	*V5*	*V6*	*V9*	*V3*
2.4	Posterior expectations										
	Expert prior	6.07	6.03	5.20	5.84	5.71	5.49	5.61	5.39	5.50	5.02
	Diffuse variance prior	6.06	6.02	5.23	5.84	5.71	5.50	5.61	5.40	5.50	5.03
	Very diffuse prior	6.10	6.07	5.37	5.91	5.79	5.60	5.70	5.50	5.60	5.17
	Correl slope, intercept	6.06	6.02	5.20	5.83	5.70	5.48	5.60	5.39	5.49	5.01
	REML estimates	6.27	6.23	5.23	6.02	5.88	5.62	5.78	5.52	5.65	5.13
2.7	Posterior expectations										
	Expert prior	7.26	7.25	7.77	7.15	7.13	7.24	6.89	6.94	6.65	6.28
	Diffuse variance prior	7.25	7.25	7.74	7.15	7.12	7.24	6.90	6.95	6.66	6.30
	Very diffuse prior	7.30	7.29	7.71	7.20	7.17	7.27	6.98	7.01	6.76	6.41
	Correl slope, intercept	7.25	7.24	7.77	7.14	7.11	7.24	6.89	6.93	6.64	6.27
	REML estimates	7.30	7.29	7.98	7.19	7.17	7.31	6.90	6.98	6.63	6.25

6. DISCUSSION

Yates and Cochran (1938) show that models for variety × environment data need to allow for differences between varieties in sensitivity to environmental changes. They generalize the additive model (3.1) to a multiplicative one of the form

$$y_{ij} = \alpha_i + \lambda_i \theta_j + \varepsilon_{ij} \qquad (i = 1, \ldots, v;\ j = 1, \ldots, l), \tag{6.1}$$

where the location effects θ_j are multiplied by parameters λ_i reflecting the sensitivity of variety i to differences between environments. Methods for fitting model (6.1) to incomplete tables assuming fixed and random location effects are described by Digby (1979) and Nabugoomu, Kempton, and Talbot (1999), respectively. Using one or more covariates to represent environmental differences rather than estimating the effects θ_j from trial data has the advantage that it allows the estimation of variety yields for new locations at which the same covariates can be measured or predicted, even with some uncertainty.

Bayesian and REML approaches to this local-area estimation problem in general provide similar point predictions unless the prior information is very influential. While REML

Table 6. Dependence of Average Posterior SEDs for Variety Yields and for Variety Effects on Expectation and SD of CHU

Expected CHU	Differences	Average posterior SED: Standard deviation of CHU				
		0	*0.025*	*0.050*	*0.075*	*0.100*
2.4	Variety yields	0.800	0.802	0.808	0.818	0.832
	Variety effects	0.414	0.418	0.430	0.448	0.473
2.7	Variety yields	0.799	0.801	0.807	0.817	0.831
	Variety effects	0.413	0.417	0.428	0.447	0.472

is widely used for analyzing variety trials data (Patterson 1997), some reasons for preferring a Bayesian approach here are as follows:

(i) The agronomist may have substantial information about the likely values of average yields and of variance components. A Bayesian formulation allows such knowledge to be incorporated into the assessment of variety performance in a systematic way by using a proper prior distribution.

(ii) The Bayesian modeling strategy formally incorporates information about local conditions into the modeling process along with the effects of sampling variability. As a consequence, the estimates of the precision of variety differences are more realistic measures of the uncertainty underlying farmers' decisions.

(iii) By using genuine prior information and expectations over the posterior distribution of the parameters rather than point estimation, we can alleviate problems associated with estimating complex mixed models, such as zero estimates of variance components.

Obvious disadvantages of the Bayesian methodology are the effort required to specify the prior distribution and the fact that simulation is required to calculate posterior distributions and summaries. However, we are implementing MCMC methodology in a decision-support system for use by farmers. In these circumstances, it is feasible to interactively elicit from users information about local conditions and the uncertainties involved, while the general experience of crop experts is more permanently embedded in the system.

Although the BUGS software is fairly straightforward to use (once the prior distribution has been chosen) and includes tests of convergence, no MCMC method is guaranteed to converge. Our experience has been that obvious failures to converge have been due to misspecification of the model being fitted.

Under the Bayesian approach, the choice of the variety to be sown at a target location can be treated as a formal decision problem. The utility for choosing each of the varieties is then specified, and the optimum variety is the one whose posterior expected utility is largest. If we are concerned only with yield, an obvious choice for the utility of growing variety i would be a multiple of the expected future yield μ_{i*}. This leads to choosing the variety with the highest expected posterior predictive yield. In practice, the choice of a variety is made after weighing the importance of several characteristics in addition to yield. In these circumstances, a Bayesian-derived utility index may have much to offer.

Although we have illustrated the use of the local-area model (3.3) with a quantitative covariate, this model allows qualitative variables to be considered as well as, or instead of, quantitative ones. For example, if the soil type were known for each location, we might include additive effects for the types and possibly allow the regression coefficients for any quantitative covariates to depend on the type also. Knowledge of the effect of soil type on yield should be sufficient for us to specify different prior means for the additive effects rather than assuming them to be drawn at random from a normal distribution.

Potentially useful generalizations of our Bayesian approach would include

(a) location effects that are spatially correlated, allowing, to some extent, for the

effects of spatial variation in soil type and weather variables not included in the model;

(b) a utility function that incorporates the cost of measuring any subset of a number of possible covariates, allowing an optimum set to be chosen.

The model of Piepho et al. (1998) allows several covariates; they suggest selecting covariates in order to minimize an estimate of the mean square error of prediction of cultivar differences at a new location, averaged over all pairs of cultivars. The dependence of the resulting set of selected covariates on the covariate values at the new location means that the prediction is a discontinuous function of the covariates, which might be seen as a disadvantage of the method.

APPENDIX: AVERAGE POSTERIOR VARIANCE OF VARIETY DIFFERENCES

To derive a formula for the average posterior variance of predicted variety differences, we define $\mathbf{z} = \mathbf{x}_* - \mathbf{x}$, $\mathbf{m_z} = \mathbf{m_x} - \mathbf{x}$, $d_{\phi i} = \phi_{i*} - \alpha_i$, and

$$\omega_i = \alpha_i + d_{\phi i} + \beta_i^{\mathrm{T}}\mathbf{z} \qquad (i = 1, \ldots, v) \tag{A.1}$$

so that $\mathbf{z}$ has expectation $\mathbf{m_z}$ and variance matrix $\mathbf{S_x}$ independent of α_i, $d_{\phi i}$, and β_i, $d_{\phi i} \sim \mathrm{N}(0, \sigma^2_{VY})$ independent given the α_i, and

$$\mu_{i*} - \mu_{h*} = \omega_i - \omega_h.$$

The average posterior variance of the differences $\omega_i - \omega_h$ $(i \neq h)$ is given by

$$\frac{2}{v-1}\left\{\sum_i \mathrm{var}(\omega_i \mid D) - v\,\mathrm{var}(\omega \mid D)\right\}, \tag{A.2}$$

where ω denotes the mean of the ω_i. Also, if random vectors $\mathbf{R}$ and $\mathbf{S}$ are independent of each other and $\mathbf{R}$ is independent of a random variable T, then

$$\begin{aligned}
\mathrm{var}(\mathbf{R}^{\mathrm{T}}\mathbf{S}) &= \mathrm{tr}\left\{\mathrm{var}(\mathbf{R})\,\mathrm{var}(\mathbf{S})\right\} + \mathrm{E}(\mathbf{S})^{\mathrm{T}}\,\mathrm{var}(\mathbf{R})\,\mathrm{E}(\mathbf{S}) + \mathrm{E}(\mathbf{R})^{\mathrm{T}}\,\mathrm{var}(\mathbf{S})\,\mathrm{E}(\mathbf{R}),\\
\mathrm{cov}(\mathbf{R}^{\mathrm{T}}\mathbf{S}, T) &= \mathrm{E}(\mathbf{R})^{\mathrm{T}}\,\mathrm{cov}(\mathbf{S}, T).
\end{aligned}$$

So, using (A.1) and replacing $\mathbf{R}$, $\mathbf{S}$, and T first by $\mathbf{z}$, β_i, and α_i and then by $\mathbf{z}$, β, and α, we have

$$\begin{aligned}
\mathrm{var}(\omega_i \mid D) &= \mathrm{var}(\alpha_i \mid D) + 2\,\mathbf{m}_{\mathbf{z}}^{\mathrm{T}}\,\mathrm{cov}(\beta_i, \alpha_i \mid D) + \mathrm{E}(\beta_i \mid D)^{\mathrm{T}}\,\mathbf{S_x}\,\mathrm{E}(\beta_i \mid D)\\
&\quad + \mathrm{tr}\left\{\left(\mathbf{m_z}\mathbf{m}_{\mathbf{z}}^{\mathrm{T}} + \mathbf{S_x}\right)\mathrm{var}(\beta_i \mid D)\right\} + \mathrm{E}(\sigma^2_{VY} \mid D)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(\omega \mid D) &= \mathrm{var}(\alpha \mid D) + 2\,\mathbf{m}_{\mathbf{z}}^{\mathrm{T}}\,\mathrm{cov}(\beta, \alpha \mid D) + \mathrm{E}(\beta \mid D)^{\mathrm{T}}\,\mathbf{S_x}\,\mathrm{E}(\beta \mid D)\\
&\quad + \mathrm{tr}\left\{\left(\mathbf{m_z}\mathbf{m}_{\mathbf{z}}^{\mathrm{T}} + \mathbf{S_x}\right)\mathrm{var}(\beta \mid D)\right\} + v^{-1}\mathrm{E}(\sigma^2_{VY} \mid D).
\end{aligned}$$

Using (A.2), the average posterior variance of the differences between the expected yields is thus given by

$$\frac{2}{v-1}\Bigg(\sum_i \operatorname{var}(\alpha_i \mid D) - v\operatorname{var}(\alpha \mid D) + 2\,\mathbf{m_z}^{\mathrm{T}}\left\{\sum_i \operatorname{cov}(\beta_i, \alpha_i \mid D) - v\operatorname{cov}(\beta, \alpha \mid D)\right\} + \operatorname{tr}\left[\left(\mathbf{m_z}\mathbf{m_z}^{\mathrm{T}} + \mathbf{S_x}\right)\left\{\sum_i \operatorname{var}(\beta_i \mid D) - v\operatorname{var}(\beta \mid D)\right\}\right] + \operatorname{tr}\left[\mathbf{S_x}\left\{\sum_i \mathrm{E}(\beta_i \mid D)\mathrm{E}(\beta_i \mid D)^{\mathrm{T}} - v\,\mathrm{E}(\beta \mid D)\mathrm{E}(\beta \mid D)^{\mathrm{T}}\right\}\right] + (v-1)\mathrm{E}(\sigma^2_{VY} \mid D)\Bigg).$$

ACKNOWLEDGMENTS

The work was partially funded by the U.K. Home-Grown Cereals Authority and the Scottish Executive Rural Affairs Department. The British Council supported the third author for a short-term visit. We thank Agriculture Canada for permission to use the maize variety trials data analyzed in this article. Sally Ritchie helped in the initial development of the Bayesian models as part of an undergraduate project at the University of Edinburgh.

[Received June 2000. Accepted August 2001.]

REFERENCES

Brooks, S. P. (1998), "Markov Chain Monte Carlo Method and Its Application," *The Statistician*, 47, 69–100.

Digby, P. G. N. (1979), "Modified Joint Regression Analysis for Incomplete Variety × Environment Data," *Journal of Agricultural Science, Cambridge*, 93, 81–86.

Freeman, G. H., and Perkins, J. M. (1971), "Environmental and Genotype–Environmental Components of Variability. VIII. Relations Between Genotypes Grown in Different Environments and Measures of These Environments," *Heredity*, 27, 15–23.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Hardwick, R. C., and Wood, J. T. (1972), "Regression Methods for Studying Genotype–Environment Interactions," *Heredity*, 28, 209–222.

Kiniry, J. R., Ritchie, J. T., Musser, R. L., Flint, E. P., and Iwig, W. C. (1983), "The Photoperiod Sensitive Interval in Maize," *Agronomy Journal*, 75, 687–690.

Nabugoomu, F., Kempton, R. A., and Talbot, M. (1999), "Analysis of Series of Trials Where Varieties Differ in Sensitivity to Locations," *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 310–325.

Patterson, H. D. (1997), "Analysis of Series of Variety Trials," in *Statistical Methods for Plant Variety Evaluation*, eds. R. A. Kempton and P. N. Fox, London: Chapman and Hall, pp. 139–161.

Piepho, H. P., Denis, J. B., and van Eeuwijk, F. A. (1998),"Predicting Cultivar Differences Using Covariates," *Journal of Agricultural, Biological, and Environmental Statistics*, 3, 151–162.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. (1996), *BUGS: Bayesian Inference Using Gibbs Sampling* (Ver. 0.5, ii), Cambridge: MRC Biostatistics Unit.

——— (1997), *BUGS: Bayesian Inference Using Gibbs Sampling, Addendum to Manual* (Ver. 0.6), Cambridge: MRC Biostatistics Unit.

Talbot, M. (1984), "Yield Variability of Crop Varieties in the U.K.," *Journal of Agricultural Science, Cambridge*, 102, 315–321.

van Eeuwijk, F. A., Denis, J. B., and Kang, M. S. (1996), "Incorporating Additional Information on Genotypes and Environments in Models for Two-Way Genotype by Environment Tables," in *Genotype by Environment Interaction*, eds. M. S. Kang and H. G. Gauch, Jr., Boca Raton, FL: CRC Press, pp. 15–50.

Wood, J. T. (1976), "The Use of Environmental Variables in the Interpretation of Genotype–Environment Interaction," *Heredity*, 37, 1–7.

Yates, F., and Cochran, W. G. (1938), "The Analysis of Groups of Experiments," *Journal of Agricultural Science, Cambridge*, 28, 556–580.