## Description of the modelling approach

The analysis consisted of a Bayesian multiple linear regression model incorporating variable selection and fitted using Markov Chain Monte Carlo (MCMC). The approach is similar to that referred to as Bayesian Variable Selection Regression by Guan and Stephens (2011) and to Indicator Model Selection by O’Hara and Sillanpaa (2009). Bayesian variable selection is a relatively recent area of research offering a wide variety of possible methods but few guiding recommendations. We opted for an approach which emphasized ease of implementation and interpretation of results.

Our motivation for using a Bayesian approach in general, and Bayesian variable selection in particular, were:

* To recognize the sparsity of knowledge regarding soil carbon and its relationship to fire frequency which makes *a priori* choice of candidate models difficult.
* To avoid arbitrary model reduction and selection (e.g. AIC step-wise methods) which could discard possibly useful alternative models.
* To generate predictions on the basis of all candidate models identified by the fitting process, in a manner similar to model averaging or ensemble modelling.
* To report on effect sizes and probabilities in direct terms and without requiring corrections for multiple comparisons.

All data preparation, analyses and inspection of results were done in R (R Core Team 2015). Model fitting was done using the JAGS software via the rjags package for R (Plummer 2015).

For ease of description, the model is described here as a set of sub-models.

### Response value sub-model

For each field sample, there were five (occasionally three or four) values for percent recalcitrant carbon from replicate chemical digests. Values were log-transformed and treated as draws from a t-distribution, with an unknown mean specific to each field sample, and unknown standard deviation and shape parameters shared across all field samples. A small value for the shape parameter (e.g. 5 or less) depicts a heavy or thick tailed distribution under which outlier values are more probable, whereas with larger values (e.g. 10 or more) the distribution more closely resembles a Normal distribution. Allowing the common standard deviation and shape parameters to be driven by the data, as part of the MCMC process, means that isolated outlier values are effectively down-weighted (have less influence on the mean response value for a field sample) while avoiding either the need to identify outliers with arbitrary criteria, or ignore them by simply taking mean values (Kruschke 2013). The common parameters also express the assumption that all chemical digests were subject to the same sources of error. The fitted mean for each field sample became the response value for that MCMC iteration.

### Regression sub-model

The regression sub-model related the fitted means from the response value sub-model (log-transformed percent recalcitrant carbon) to predictors. Predictors consisted of log-transformed percent initial carbon (single value per field sample) and categorical variables for the three regions, two sample depths, three micro-site classes and three fire frequency classes.

The regression included the following three-way interaction terms:

* log10 percent initial carbon X region X sample depth
* region X sample depth X fire frequency
* region X micro-site X fire frequency

These were chosen on available knowledge and expert opinion (?). All single terms and pair-wise interactions making up these three-way interactions were also included in the matrix of predictors, together with an intercept term. With categorical variables represented as dummy variables, this resulted in a set of 41 predictors plus the intercept.

All regression coefficients had uninformative priors in the form of a vague Normal distribution with a common, unknown standard deviation based on a Uniform prior.

### Variable selection sub-model

For each iteration of the MCMC process, this sub-model chose a subset of predictors to include in the regression, with the coefficients for other predictors set to zero. Each predictor was associated with a binary indicator variable drawn from a Bernoulli distribution, with a common probability of inclusion for all predictors based on a Beta prior. We constrained variable selection so that if an interaction term was included, its component terms would also be included.

### Posterior prediction sub-model

Model predictions were generated for 54 cases representing all possible combinations of the categorical predictors (region, sample depth, micro-site class and fire frequency class). For each of these cases, the value of initial soil carbon was set to the median for the respective sample depth. At each iteration of the MCMC process, predictions of recalcitrant carbon were made for each case using the current regression (i.e. set of included predictors and their associated coefficients).

### Model fitting and outputs

The model was initialized for 50,000 MCMC iterations, after which 100,000 iterations were run and thinned to every 10th element, resulting in 10,000 samples of the posterior distribution of model terms (regression coefficients, indicators of predictor inclusion, t-distribution parameters and other general parameters) together with a matrix of posterior predictions.

Using these outputs, predictors were ranked according to their frequency of inclusion, akin to variable importance, and the posterior distribution of their coefficient values summarized by mean and central 95% interval. Predicted effects were examined by summarizing the matrix of posterior predictions along selected axes.

## References

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large scale problems. Annals of Applied Statistics 5: 1780-1815.

Kruschke, J.K. (2013). Bayesian estimation supersedes the t-test. Journal of Experimental Psychology: General 142: 573-603.

O’Hara, R.B. and Sillanpaa, M.J. (2009). A review of Bayesian variable selection methods: what, how and which. Bayesian Analysis 4: 85-118.

Plummer, Martyn (2015). rjags: Bayesian Graphical Models using MCMC. R package version 3-15. https://cran.r-project.org/web/packages/rjags/index.html

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/