TODO LIST

1) Update References to Bibtex throughout document
2) Finish Related Work section
3) Finish Data Processing sections: LANDSAT, DATA PROCESSING (Include plots of distribution of species and potentially the distribution of LFMC over the data if it is instructive), Also adding footnotes to indicate symbol representation of all predictors
4) Complete BMS and BMA sections (could potentially be retitled)
5) Add Abstract if we have room. Amend Discussion, Related Work and Introduction if necessary
6) Edit for typos, redundant information, coherence, consistency, continuity.
7) Generate summary tables to input to Appendix (Append Github link as well, assuming this is sufficient)
8) Reformat/typeset tables, figures, appendix
9) Make title page

1. INTRODUCTION - 0.5 Page

As the world has increasingly seen the rise of wildfires due to climate change, it is an important topic to try and improve predictive models for the ignition, spread, and severity of wildfires. Research has shown that the fuel moisture content of vegetation is an important determining factor in the ignition, behaviour and severity of wildfires. The fuel moisture content of vegetation can be split up into two categories: live fuel moisture content (LFMC) and dead fuel moisture content (DFMC). Both have been found to have crucial effects on the spread of wildfires. There are many models that predict DFMC well because of its static nature; on the other hand LFMC deals with living organisms and the moisture content of such vegetation is much more dynamic and harder to model. The available models that are used to predict LFMC are often found to be lacking. Furthermore, LFMC observations are painstakingly collected by hand. Not only does it require the researcher to travel into remote areas—often the most important areas in terms of wildfire risk—but they must also go through the long process of oven drying the samples. Moreover, this sampling cannot practically be used to create real-time predictions during fire season which rely on predictive models that in turn rely on being trained on these manual samples of LFMC.

The most prevalent models used for prediction of LFMC employ the use of meteorological indices and drought indices. It is found that these methods lead to somewhat useful results on a large scale, but when considered under a species or smaller scale spatial level, the models begin to fall apart with very high spatial variability. More recent approaches have begun to employ the use of remote sensing data with machine learning techniques to generate real-time predictions on a country level. These results are promising, but are once again susceptible to high spatial variability. Some studies have indicated that this spatial variability could be attributed to the fact that different species of vegetation react differently to the same weather conditions; some species can be considered high-responding, whose moisture content responds strongly to a change in weather, whereas others are low-responding. Research indicates that there is promise in using more mechanistic indices, especially as data on vegetation and vegetation distribution becomes better.

Our study investigates the viability of this mechanistic approach by combining meteorological factors, plant characteristics and remote sensing data. We focus our attention on the region of Southern California, an area of high wildfire risk, to prototype a model of this sort. Our model predicts the LFMC of nine different plant species scattered across Southern California using meteorological predictors (temperature, precipitation, relative humidity, solar radiation, and wind speed), NDVI, and various plant traits (specific leaf area, nitrogen per dry mass of leaf, phosphorous per dry mass of leaf, and plant height). A successful model would still depend on strong knowledge of the distribution of the vegetation within the area of concern, but this is outside the scope of this study. Apart from that, all of the predictors used are either readily available through automatic reporting or features of vegetation that are common domain knowledge a priori, which conforms to the constraints of creating a model for real-time applications.

2. RELATED WORK - 0.25 Page

In (Ruffault et al., 2018), the authors assess and show that the prevailing method of using drought indices to predict LFMC is plagued by its limited ability for spatial predictability and cannot provide reliable estimates on a local level. The main benefit of using drought indices is that the data can be automatically sensed and is relatively widely available. An improvement on this method, as suggested by the study is to use a more mechanistic approach. The authors in

(Castro et al., 2003) illustrates that when considering a specific plant species, meteorological predictors are sufficient to be able to produce strong predictions (approximately 0.8 R^2) which are able to generalize to different locations and maintain its predictive ability. The limitation here is that if one seeks to create a model that could be generalized for large areas, one would need to get accurate estimates of the distribution of vegetative species. A helpful approach to ameliorate the uncertainty in the distribution of species is to take advantage of latent traits amongst species that could give us information on species we have not even sampled. For example, (**?**) points out that plants can be largely categorized into high and low responding categories indicating their moisture content responsiveness to drought.

Another new approach to predicting LFMC is a more deterministic approach that uses NDVI as a proxy for the moisture content. Furthermore, this can be combined with land surface temperature measurements which can also be inferred from remote sensing satellite data. This approach is detailed in (Chuvieco et al., 2004) using a physics based approach involving microwave backscatter. They are quite successful in obtaining strong predictions (R^2 of about 0.8 using just multiple linear regression. The model was based off of a similar dataset of LFMC samples in the Mediterranean, but are only specifically successful with a small number of shrub and grassland species. Similarly, (McCandless et al., 2020) uses remote sensing data, but with more flexible models on the WFAS dataset for the entire United States. The goal of this study was to create a real-time remote-sensing predictive model at the national level. This study primarily uses land surface temperature and satellite bands (which are used to derive NDVI) as predictors, and multiple linear regression, neural networks, random forest, and gradient boosted regression as models. It achieves an overall mean squared absolute error of about 20% which is approximately 25% of the standard deviation of LFMC, but there is no evaluation of the accuracy of the model at a small scale. Furthermore, the computational costs involved were significant, involving over 4 TB of data and the use of the US NCAR Casper cluster supercomputer.

These approaches have either been able to create large scale predictions that suffer from spatial variability at a small scale, or have been able to create strong predictions locally that do not necessarily generalize across all dimensions of the problem.

Main articles:

1) Enhancing Wildfire Spread Modelling by Building a Gridded Fuel moisture Content product with machine Learning
Key points:
Data used: WFAS, MODIS Sattelite for bands 1-7, WRF-Hydro data for soil saturation, evapo, geographical characteristics
Models used: GBR, ANN, Random Forest, Multiple Linear Regression
Shortcomings: Large amounts of data 4TB. Elevation was the predictor of greatest importance. Mean absolute error of 21.92 on a mean vallue of 94.8 with SD 86.2. High spatial variability potentially not captured.

2) Combining NDVI and surface temperature for the estimation of live fuel moisture content in forest fire danger rating
methods: focused on areas of grasslands and shrubs
data: NDVI and LST, AVHRR
models: multiple linear regression
results: strong results of 0.85 +, but remember only for grasslands and shrubs. Also showed that LST was just a strong predictor for grasslands/shrubs (check on this)

3) How well do meteorological drought indices predict live fuel moisture content (LFMC)? An assessment for wildfire research and operations in Mediterranean ecosystems
4) Modelling moisture content in shrubs to predict wildfire risk in Catalonia (Spain)
Methods: Hand sampling over 2 years.
Models: Multiple linear backwards regression and GLM. used site as fixed factor
Shortcomings: Only applicable to their two specific species of shrubs.
Pros: good results with R^2 of about 0.8 and captures spatial variability for those specific species

4b) Critical live fuel moisture in chaparral ecosystems: A threshold for fire activity and its relationship to antecedent precipitation
Discusses the importance of LFMC predictions for Sourthern California, especially the 79 percent threshold

5) Continental-scale prediction of live fuel moisture content using soil moisture information
6) Evaluating remotely sensed live fuel moisture estimations for fire behaviour predictions in Georgia, USA

3. DATASET - 1 Page

In this study, we brought together four different datasets to predict the live fuel moisture content (LFMC). The LFMC observations come from surface observations sampled manually. The meteorological data comes from automatic surface weather stations. The satellite data comes from the Landsat7 satellite. The plant characteristics come from the TRY plant database which is a curated conglomeration of many databases that largely involve manual sampling.

3.1 WFAS

Observations of LFMC content are provided by the USFS Wildfire Assessment System. This contains a national database of LFMC as well as dead fuel moisture content (DFMC) samples. LFMC is given by

$$\frac{WaterWeight}{OvenDriedWeight} \cdot 100$$

so it is possible for these values to be higher than 100. This study focuses on the Southern California subset of this data which contains a total of 37,912 total observations spanning 1982-05-06 to 2021-11-02. Although there are strong guidelines for the sampling of observations, the sampling is largely carried out by citizen scientists or volunteers. This means the observations may not be completely reliable, as well as the information on site location is approximate at best. We used reverse geocoding to find the nearest coordinate values that the site names may indicate.

3.2 Meteorological Data

Meteorological data from automatic weather stations is often not consistent, and can contain many errors. Furthermore, databases such NOAA's Climate Data Online archives actually have large numbers of missing observations and many gaps and inconsistencies. One option was to use paid weather data services that generate interpolated data to fill in the gaps due to instrument errors, but this does not make sense for a study whose ultimate goal is real time prediction. Finally, we were able to use the Mesonet provided by Synoptic which as a third party gathers and quality controls weather stations across the USA. The most reliable network for our region was the California Irrigation Management Information System. From this network we were restricted to retrieving only the most common weather variables that would assuredly be available in all of our stations. Furthermore, we only used weather stations that were at most 30km away from our geocoded WFAS site locations. The combination of this reduced our dataset to 7628 observations that we would consider. From the weather stations we used precipitation, wind speed, solar radiation, relative humidity, and temperature. The quality controlled weather stations only had a small number of missing observations as well as observations that were clear measurement errors[1]. For these missing variables we made a local imputation by taking the average of the five preceding observations under the assumption that the weather is most similar to what had just passed. All of these readings were provided hourly. Using these readings we could extracted features pertaining to the three, seven and fifteen day rolling average, maximum, and minimums for each of these five initial parameters[2].

3.3 TRY Plant Database

Vegetation specific characteristics were retrieved from the TRY Plant Trait Database which is a conglomeration of many datasets worldwide curated by a network of vegetation scientists. From this database we requested a long list of traits for our species of concern and trimmed this list down to four traits for relevance and data availability: nitrogen content per dry leaf area, phosphorous content per dry leaf area, plant height, and specific leaf area (a ratio of leaf area to leaf biomass). These characteristics were chosen for their relevance in conjunction with NDVI data under the hypothesis that different plants have different relationships between NDVI and moisture content, and this can be characterized by the properties of their leaves as well as their overall plant height which is not easily inferred by a satellite scan of NDVI[3].

[1] Anything outside of ranges: $0 < SolarRadiation < 1100 w/m^2$; $0 < Precipitation < 100mm$; $0 < WindSpeed < 50km/hr$; $0 < RelativeHumidity < 100$; $-15 < Temp < 55C$ daily were considered as measurement errors based on the annual records of the area.

[2]

[3]

3.4 Landsat 7

The Landsat 7 is a satellite that was put into orbit in 1999 with a 16 day orbit cycle. It detects 8 spectral bands from earth's surface. We collected data at each site location from the satellite for the range of dates available to us from the WFAS dataset. We calculate the Normalized Vegetation Index (NDVI), which is a measure of the density of live vegetation on land—NDVI is a common predictor of LFMC in the literature. To keep things simple we took the mean NDVI in a 10km radius around the associated site of each LFMC observation.

3.5 Data Processing

From the meteorological data, we were able to generate drought indices that have been widely used to predict fuel moisture content and in wildfire predictive indices such as the Canadian Forest Fire Weather Index. These include the Duff Moisture Code (DMC), the Drought Code (DC), and the Build up Index (BUI)[4]. We also generated a growth cycle variable based on the day of the year to capture the intrinsic plant cycle throughout the seasons based on the formula

$$D = \cos\left(\frac{2\pi D}{365} - 0.59\right)$$

from (Castro et al., 2003). The initial meteorological features were further aggregated to produce for each variable (precipitation, humidity, temperature, wind speed, and solar radiation) their respective minimum, maximum and mean for the previous three, seven and fifteen days. Finally, we created interactions between the growth cycle and the meteorological variables to capture the different responses of plants to weather based on the phase of their growth cycle.

Due to the timing of satellite overflights from Landsat 7 (one observation every 16 days), NDVI data was not available on a daily basis. To adjust for NDVI not matching the target observation date we used an exponential decay function

$$ANDVI_t = \rho^{-h} \cdot NDVI_{t-h} \quad (1)$$

to downweight the importance of each observation. We set $\rho = 1.8$ to quickly decay in importance in the first 30 days. We then created interactions between our adjusted NDVI (ANDVI) and our vegetation characteristic variables to capture the hypothesis that ANDVI given the characteristics that generate ANDVI can be a proxy for fuel moisture content rather than just ANDVI alone. Furthermore, it is key to note that NDVI and NDVI related measures are included for their predictive value, but from an inference perspective it can only be viewed as a proxy of some latent unobservable characteristic of individual plant species that would also determine their LFMC.

[4]See appendix for relevant formulas: DMC(formulas 8-13), DC(formulas **??**-6), and BUI(formula 14)

Plant Type	**N = 6,498**[1]
buckwheat, eastern mojave	205 (3.2%)
ceanothus, bigpod	155 (2.4%)
ceanothus, hoaryleaf	130 (2.0%)
chamise	3,243 (50%)
chamise, new growth	1,071 (16%)
sage, black	557 (8.6%)
sage, purple	162 (2.5%)
sagebrush, black	421 (6.5%)
sagebrush, california	554 (8.5%)

Table 1. Sample distribution across Fuels.

We removed species Red Shank, Eastwoods Manzanita, and Brittlebrush from our study because these groups contained fewer than 25 observations each.

3.6 Data Exploration

Our data contains $n = 6498$ observations and $p = 122$ variables, starting in late 2003 to November 2021. Table 1 outlines the number of variables per Fuel and per stations.

The distribution of the LFMC variable is shown in figure X.

4. REGRESSION - 4 Pages AND SUPERVISED/UNSUPERVISED LEARNING 4 Pages

We estimate the following basic regression model.

$$y = \boldsymbol{\beta}^T \boldsymbol{X} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \rho I)$$

We will analyse this model mainly using Bayesian methods. We avoid using LASSO or other frequentist methods as we have many highly correlated variables.

Since we are estimating a Bayesian model, we must specify the priors we are placing on our parameters.

We let

$$\boldsymbol{\beta} \sim \mathcal{N}(0, ng\rho(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$
$$\rho \sim IG(0.1, 0.1)$$

which is a standard choice of priors for Bayesian regression.

We set g to a minimally reflect a minimally informative prior.

4.1 Bayesian Model Selection and Model Averaging

We run a Bayesian Model Selection procedure, which samples from a posterior distribution of a *spike and slab* prior $\boldsymbol{\gamma} \in \{0,1\}^p$. The idea is that the $\boldsymbol{\gamma}$ variable indicates which covariates are included in the model, i.e $\beta_j \neq 0$ if $\gamma_j = 1$. So if we set a prior on $\boldsymbol{\gamma}$ we can apply Bayes theorem and sample from $p(\boldsymbol{\gamma} \mid \mathbf{y})$, which allows us to find which models are most probable under the data. We set $\boldsymbol{\gamma} \sim$ BetaBinomial(1,1), which sets a uniform prior on model size.

First, we can examine the marginal posterior probability that a variable is included in the model $p(\gamma_j = 1 \mid y)$.

4.2 Bayesian Model Averaging

We use Bayesian Model Averaging (BMA) to compute point estimates for our β coefficients. This gives the expected value of a coefficient over all possible models.

$$E[\boldsymbol{\beta} \mid \boldsymbol{y}] = \sum_{\gamma_j} E\left[\beta_j \mid \gamma_j = 1, \boldsymbol{y}\right] p(\gamma_j = 1 \mid \boldsymbol{y}).$$

It is worth noting that we are not too interested in the actual point estimates of the model. Since our focus is mostly on prediction, and the problem we are studying is much too complex to assume a simple causal structure. To naively interpret the coefficients as the marginal effect of a covariate on LFMC could be misleading, since we have not accounted for the effect of confounding variables—see (Westreich and Greenland, 2013) for a discussion. We present the BMA point estimates for the whole unpooled model in table **??** for those variables with a high marginal probability, i.e those variables with a high probability of being contained in the *true* model.

The important weather variables like average solar radiation and minimum temperature are significant when interacted with the seasonality variable D. This reflects the non-linear (in this case sinusoidal) relationship in weather variables.

We can see that the interaction variables containing the plant traits and NDVI are likely to be in the true model. This highlights the relationship between plant traits and the vegetation denseness. For example, the interaction between plant height and NDVI is around -4. This indicates that the influence of NDVI on LFMC is dependent on plant height. For smaller plants, low NDVI (coarsely, low green-ness) influences LFMC less than for taller plants. This lends credence to our approach of using plant traits to model LFMC.

From our grouped fuel analysis, we analyse the difference in Bayesian estimates across plants. That is we run a model selection for each plant, the motivation is that we are trying to isolate common significant variables common to plants in order to motivate our hierarchical model design in section 4.4.

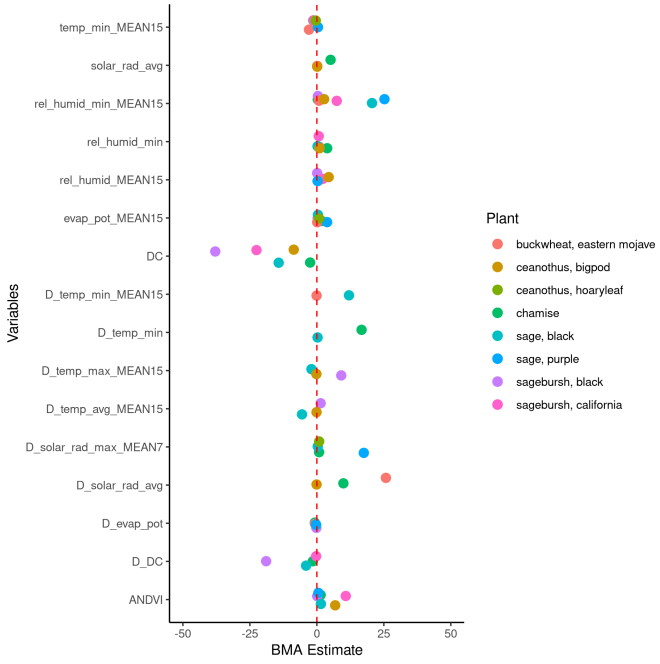

Fig. 1. Comparison of BMA for regressions subsetted by Plant. Variables shown are those for which the unpooled BMS placed a high probability on.

Figure 1 shows the BMA estimate for those variables who were probably in the unpooled model.

We can see that there is variation between the plants.. Of course it is not clear from this that the variables differ because of plant characteristics or because of location characteristics; since one plant is mostly located in one location.

4.3 Principal Components Analysis

Many of the meteorological variables are collinear with each other as is the nature of weather variables. We use PCA to reduce the dimensionality of the problem under the assumption that there are some smaller number of latent signals that drive weather patterns. Looking at the scree plot in figure 2 we can see that a significant amount of variation is described in the first component with the largest gap between component one and two. From this plot it appears that if there existed truly a latent variable signal structure, then it would consist of the first 4 components. Table 2 outline the makeup of the first six principal components. It is instructive to see that the principal components group our variables into intuitive clusters that could represent the true signals that are driving the variation in our data. These signals for the first six principal components can be largely categorized as seasonal temperature changes, relative humidity, wind speed, accumulated precipitation, NDVI's interaction with plant characteristics, and solar radiation.

PC1: Temperature	D_t, $D_t : T_{max15}$, $D_t : T_{mean15}$, $D_t : T_{max7}$, $D_t : T_{mean7}$, $D_t : T_{max3}$, $D_t : T_{mean3}$, $D_t : T$, $D_t : G_{max15}$
PC2: Humidity	H_{mean7}, H_{mean15}, H_{mean3}, H_{min7}, H_{min3}, H_{min15}, H, H_{min}, W_{max7}, W_{max15}
PC3: Wind	W_{mean7}, W_{mean3}, W_{mean15}, W_{max7}, W_{max3}, W_{max15}, W_{min7}, W_{min15}, W, W_{min3}, W_{max}
PC4: Precipitation	P_{mean7}, P_{Emean7}, P_{mean15}, $D_t : P_{Emean7}$, $P_{Emean15}$, $D_t : P_{mean7}$, P_{mean3}, P_{Emean3}, $D_t : P_{Emean15}$, $D_t : P_{Emean3}$
PC5: NDVI	$L_{phos} : A_{NDVI}$, A_{NDVI}, $L_{SLA} : A_{NDVI}$, $L_{nitro} : A_{NDVI}$, $V_{Height} : A_{NDVI}$, G_{mean3}, G_{max3}, T_{max3}, G_{max7}, G_{mean7}
PC6: Solar Radiation	G_{max3}, G_{max7}, G_{mean3}, G_{max}, G, G_{mean7}, D_{DC}, G_{max15}, G_{mean15}, B_{UI}

Table 2. The composition of the first six principal components including the ten most important variables for each component. We have given interpretations of the groupings on the left. See the appendix for a definition of each. Variables which are of the form X_{aggk} where $agg \in \{mean, min, max\}$ and $k \in \mathbb{N}$ indicates the value of X aggregated with it's k previous observations.

4.4 Mixed Effects Modelling

One strategy to combat the spatial variability that plagues many predictive models of LFMC would be to follow a more mechanistic approach that relies on characteristics of the vegetative species in question. These results could then be connected to the spatial dimension through estimating the distribution of the plants within each region; this portion is outside of the scope of this study. Without a strong model and data that can predict LFMC from some unknown characteristics, we will take these characteristics as latent variables and use a mixed effects model to account for and learn from these variables. We will generate a model both using principal components and using variables selected from Bayesian Model Selection from the general model given by equation: 2.

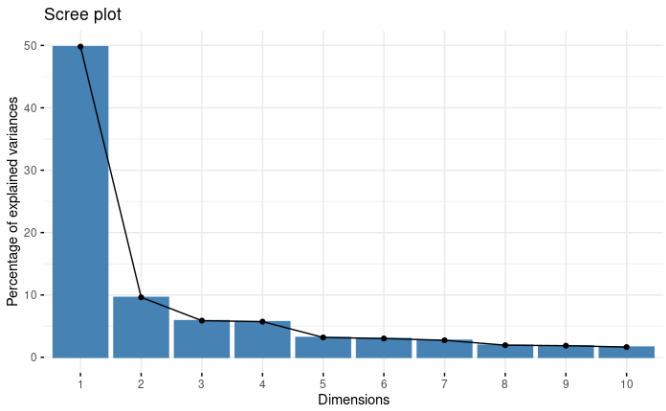

Fig. 2. The amount of explained variance for each of the first 10 principal components

$$\begin{aligned} \text{LFMC}_i &\sim \mathcal{N}\left(\mu, \sigma^2\right) \\ \mu &= \alpha_{j[i]} + \gamma X_0 + \beta_{j[i]} X_1 \\ \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} \mu_{\alpha_j} \\ \mu_{\beta_j} \end{bmatrix}, \ \Sigma\right), \text{for Species j} = 1, \ldots 9 \end{aligned} \tag{2}$$

LFMC is modelled under a normal distribution with mean μ and variance σ^2 whose mean parameter is estimated from a mixture of Gaussian distributions composed of three parts. The first part is the intercept for each group given by α_j. The second part is given by the fixed effects with coefficients γ and X_0 which represents relevant partition of the design matrix. The third part is given by the rest of the design matrix of the predictors used as random effects with group coefficients given by β_j. Finally, α_j's and β_j's are estimated by a normal distribution with mean $\mu_{\alpha_j - \beta_j}$ and covariance Σ.

For the PCA-based model we use the first twenty two principal components as predictors; these explain 95%

of the cumulative variance. We first fit two baseline: a completely pooled OLS and a completely unpooled OLS. We then fit a model varying the intercepts. In this case in reference to model 2 this would mean that X_0 is the entire design matrix of all predictors. Next we fit a model using all components as random effects; i.e. X_1 is the entire design matrix. We then fitted a model using just the first six principal components for X_1 and the other 16 in X_0; we choose these components through a combination of observing the scree plot (fig 2) and observing that the first six components still explains nearly 80% of the cumulative variance. After fitting this model, we run ANOVA tests to determine if all the components are truly significant as a random effect. Under its recommendation we fit a final model using only principal components 1, 3, 5, and 6[5].

For the BMS-based model we use the 54 predictors chosen and follow a similar approach. We make similar baseline plots as well as a varying intercept plot. Due to issues of collinearity amongst the random effects, we cannot use all the predictors that we heuristically chose from our analysis of the marginal posterior probabilities of inclusion. We further trim our choice of predictors by considering their covariances and setting a threshold of 0.9. Finally, we fit a model using eleven predictors: H_{mean7}, D_{DC}, T_{min15}, W_{mean15}, $D : D_{DMC}$, $D : W_{max}$, H, G_{max3}, G_{max}, and $L_{nitro} : A_{NDVI}$ to compose X_1. Using the same ANOVA tests, we remove H, G_{max3}, G_{max}, and $L_{nitro} : A_{NDVI}$ for a minimal model of 6 random effects[6].

Comparing the mixed effect models based on the deviance information criterion (DIC) in table 3, is the PCA-based model that uses all the principal components with a DIC of 49,607 and deviance 49,596, but BMS-based model that uses eleven varying slopes is not that much worse with a DIC of 49,783 and deviance 49,634. Although the BMS model uses fewer random effects, it has a higher number of effective parameters implied by the difference between the DIC and deviance. These models also perform similarly in prediction with out of sample R^2 of approximately 0.68.

We tested the predictive accuracy of our models on a held out sample and measure the overall R^2 as well as the species specific R^2. For our mixed effect models we make one thousand bootstrap simulated predictions[7]. We make simple predictions with 95% confidence intervals with our OLS models. Figures 3 and 4 show that our models can both successfully improve model predictive accuracy as well as reduce the variation between groups. These figures outline the different predictive abilities of the different models using the same base predictors as specified above. By grouping the predictions in the same way for each model, stratified by each individual species group, we can see the ability of the mixed effects models to learn information across groups. In figure 3 it is evident that the baseline OLS models have a relatively high dispersion of predictive abilities across species, and the mixed effects models are able to learn information across groups and significantly increase the predictions of species such as Black Sage, Eastern Mojave Buckwheat, Chamise, and Chamise New Growth. These species would likely have been characterized as "low-responding" species by (reference drought study here). There is a large disparity in the number of observations for each species ranging from a hundred to a few thousand. In this case, we might infer that Eastern Mojave Buckwheat and Black Sage, which had only 206 and 421 observations, benefited in this regard by taking advantage of the grand mean of all groups and pooling towards it. But, the same cannot be said of Chamise, and Chamise New Growth, which have 3,251 and 1,073 observations respectively. Together they make up for nearly two thirds of all observations in this sample. For these two groups we can observe a large improvement from the completely pooled OLS and the completely unpooled OLS predictions which indicates that there are species specific traits that are not accounted for by the predictors. Further improvement is only then achieved significantly when random slopes are incorporated, showing that different species react differently to different predictors, and by characterizing this we can more accurately describe the data. This regularization is effective and useful for both large and small group sizes meaning that the improvements in the global predictive ability is not solely derived from regularizing the groups with small sample sizes[8]. In the BMS-based models

[5] This model is referred to as the PC min model in the rest of this paper.

[6]

[7] The predictInterval function was used from merTools which differs from the standard arm::sim by incorporating the uncertainty in the variance of the group parameters by making a few draws of these variances while still treating them as 'fixed'. This leads to a higher prediction interval than `arm::sim` which only incorporates the uncertainty of the fixed effects and the observation level variances and of course is higher than expected confidence intervals of OLS predictions by nature.

[8] The appendix contains summary tables of the random effects and estimated coefficients to see the precise changes in estimated

Models	Deviance	DIC	OOS R^2
PCA - Varying Intercept	51438	52437.8	0.5463
PCA - Varying Intercept + 22 PCs	49595.5	49606.9	0.6846
PCA - Varying Intercept + 6 PCs	50088.5	50089.7	0.6696
PCA - Varying Intercept + 4 PCs	50160.3	510158.4	0.6582
BMS - Varying Intercept	50891.5	51808.8	0.5961
BMS - Varying Intercept + 11 Slopes	49634.3	49783.2	0.6880
BMS - Varying Intercept + 6 Slopes	49690.2	49850	0.6802

Table 3. A comparison of mixed effect models' deviance, DIC and out of sample R^2. The PCA models indicate the number of first principal components used as random effects, with the exception that '4 PCs' uses principal components 1, 3, 5, and 6 chosen based on ANOVA tests. The simplest BMS model uses: H_{mean7}, D_{DC}, T_{min15}, W_{mean15}, $D : D_{DMC}$, $D : W_{max}$; and the larger model adds on: H, G_{max3}, G_{max}, $L_{nitro} : A_{NDVI}$.

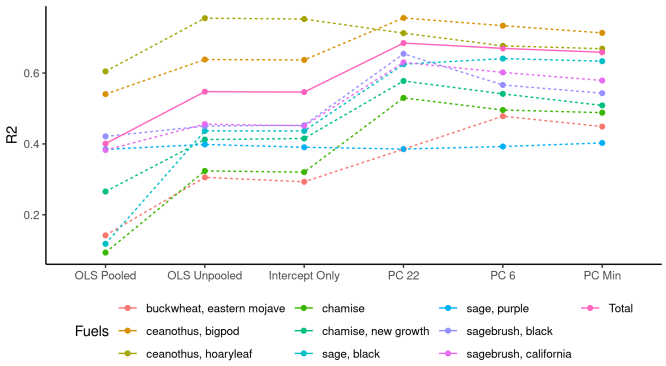

Fig. 3. Predictive accuracy of different models using the first twenty two principal components as predictors.

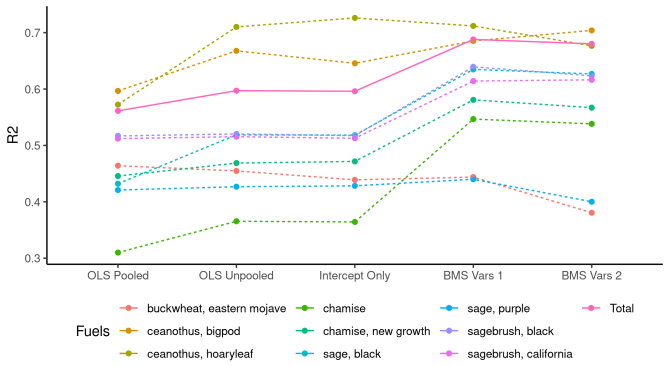

Fig. 4. Predictive accuracy of different models using predictors selected using Bayesian model selection.

(fig 4) a similar story holds for Chamise and Chamise, New Growth, albeit less drastic.

Figure 6 compares the OLS unpooled estimates and our PCA-based model for Black Sage, one of the species that saw the most drastic improvement. We can observe that the prediction intervals are much greater than the OLS confidence intervals. But, in the end, the bootstrapped mean predictions can be seen to be better than the OLS estimates, especially for the values that are below the 79% threshold. We notice that some of our prediction intervals fall below zero, which indicates an area of improvement on this model since it is impossible for a plant's LFMC to be below zero.

Comparing the predictions using the best BMS and PCA based models for Chamise in figure 5, there is not a huge difference in their predictions nor their predictive intervals. This is further evident in the residual unexplained variance of the observation level of these models; for the PCA model it is 795.2, and for the BMS model it is 809.8. Although the BMS model's predictors could be a bit easier to interpret. Intuitively the view that a small number of signals are responsible for the large number of predictors that forms the basis of principal components analysis aligns with the types of predictors and the way they are generated in this model. It also then makes sense that the predictors chosen through BMS include predictors in all categories (temperature, precipitation, humidity, solar radiation, NDVI, drought indices), indicating that BMS is able to select these predictors from amongst an environment of collinearity, but these might not necessarily be the true determinants of LFMC.

coefficients between such models as OLS and the different random effects models

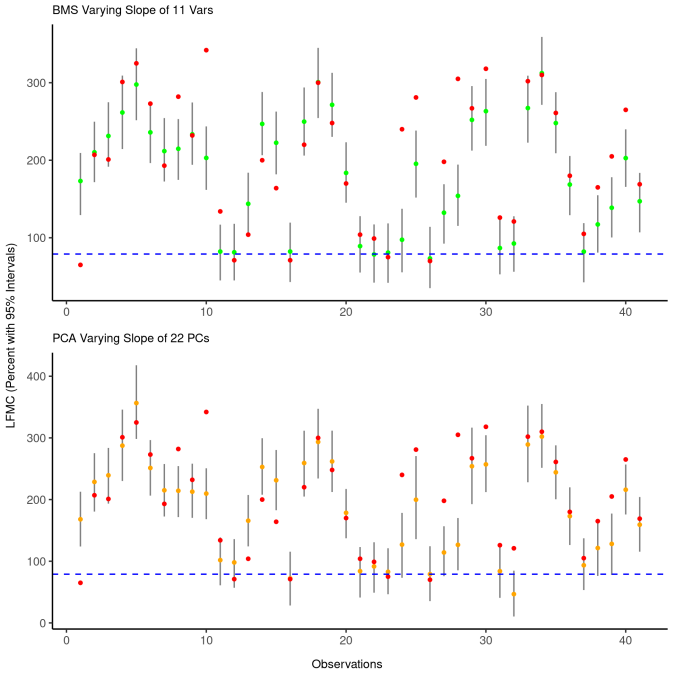

Fig. 5. Out of sample predictions of Black Sage using unpooled OLS and mixed effects model using all 22 principal components as predictors. The red points indicate the true values and the blue dotted line indicates LFMC=79%. Prediction intervals for the mixed effects model and confidence intervals for OLS are shown for the 95% level.

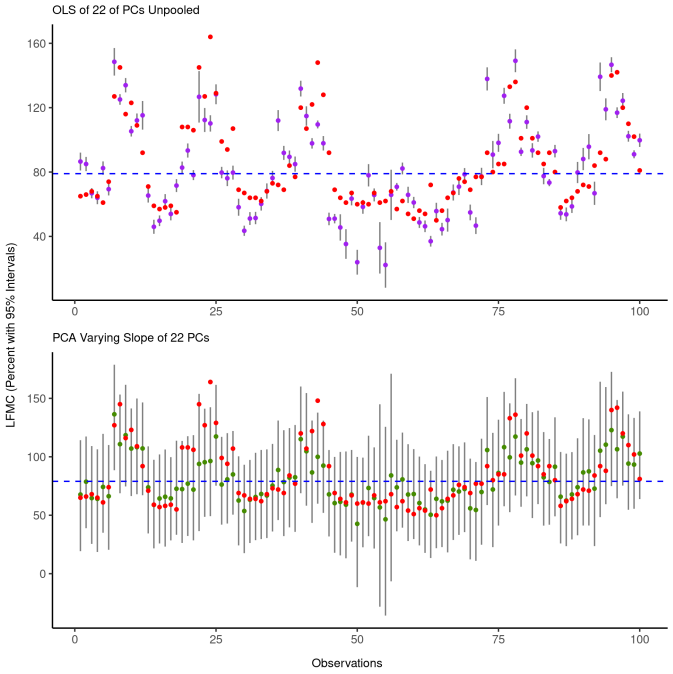

Fig. 6. Out of sample predictions for the first 100 observations of Chamise using a PCA-based mixed effects model and a BMS-based mixed effects model. The red points indicate the true values and the blue dotted line indicates LFMC=79%. Prediction intervals for the mixed effects model and confidence intervals for OLS are shown for the 95% level.

5. DISCUSSION - 0.25 Page

We learned the following from the Bayesian Model Selection ...

Mixed effects models here are an effective way to deal with the dispersion of predictive ability amongst groups of plant species. They provide stronger predictive results and also insights into the different responses of different species that determine their LFMC. There are a few ways in which these models can be improved and extended to provide a more complete picture. Firstly, given more computing resources, it would be possible to perform a more rigorous search of our predictor space for the predictors that can capture the most variance between groups. Moreover, it would be an improvement to run hierarchical models to encode prior domain knowledge. This would include a sensible lower bound to LFMC since living plants have a definite lower bound. As seen in the predictions in figure 6, this lower bound was violated. Another way to encode prior beliefs would be to employ quantile regression within a mixed effects model. Since this model, is motivated to improve wildfire prediction models and we know that the 79% threshold of LFMC is a critical threshold (insert reference here to critical levels of LFMC paper), we could use quantile regression to upweight the importance of these lower observations and downweight the heavy tails in the other direction; we are more interested that vegetation is too dry rather than over saturated. This would be a cleaner way than the use of logistic regression in (reference to drought paper).

Further work can also be done by gathering more data. Gathering more LFMC data in a significant way may not be feasible, but gathering more accurately matched meteorological and NDVI data is. This would allow the model to be expanded to find other groupings by including more groups of species as well as to have enough observations to consider the time and spatial dimensions. Firstly, if more species could be included, then there are possibilities of using clustering techniques to discover latent clusterings of the species that could be distinct from the current prevalent groupings: by species, as low and high responding species, or by type (i.e. shrub or tree). As we saw in figure 3, the predictability of certain species was not solely due to a low sample size. Finally, with more observations it would be prudent to find ways to learn more from the data via time and spatial dimensions such as through time series techniques and spatial regression.

6. APPENDIX

6.1 Code

All source code can be found at https://github.com/antotocar34/fmc_prediction

The main code for analysis is in https://github.com/antotocar34/fmc_prediction/tree/master/code/analysis/main

6.2 Formulas

Drought Code

$$DC = \begin{cases} DC_{t-1} + 0.5 \cdot V & \text{for } P \leq 2.8 \\ DC_{r_t} + -.5 \cdot V & \text{for } P > 2.8 \end{cases} \tag{3}$$

Where DC_{t-1} is the previous day's value, and if unavailable is set to 15 and V is the potential evapotranspiration calculated by:

$$V = 0.36 \cdot (T_{12} + 2.8) + L_f \tag{4}$$

Where T_{12} is the temperature recorded at noon if $T_{12} \geq -2.8$ and otherwise $T_{12} = -2.8$ and L_f is a day-length factor given by Table 4

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
-1.6	-1.6	-1.6	0.9	3.8	5.8	6.4	5.0	2.4	0.4	-1.6	-1.6

Table 4. Day length factor for the drought code

And DC_{r_t} is calculated by first calculating effective precipitation P_{eff} using precipitation P:

$$P_{eff} = 0.83 \cdot P - 1.27 \tag{5}$$

Then calculating the day's moisture equivalent after rain Q_{r_t}:

$$Q_{r_t} = 800 \cdot e^{\frac{-DC_{t-1}}{400}} + 3.937 \cdot P_d \tag{6}$$

And finally calculating DC_{r_t}. Note if $DC_{r_t} < 0$ then $DC_{r_t} = 0$:

$$DC_{r_t} = 400 \cdot \ln \frac{800}{Q_{r_t}} \tag{7}$$

Duff Moisture Code

$$DMC_t = \begin{cases} DMC_{t-1} + 100 \cdot K & \text{for} P \leq 1.5 \\ DMC_{r_t} + 100 \cdot K & \text{for} P > 1,5 \end{cases} \tag{8}$$

Where DMC_{t-1} is the previous day's DMC or 6 if unavailable, K is the log drying rate calculated by using the temperature recorded at noon, T_{12} if $T_{12} \geq -1.1$ otherwise $T_{12} = -1.1$ and the relative humidity in percent recorded at noon H_{12} and the effective day length L_e given by 5

$$K = 1.894 \cdot (T_{12} + 1.1) \cdot (100 - H_{12}) \cdot L_e \cdot 10^{-6} \tag{9}$$

And DMC_{r_t} is calculated by first calculating effective rainfall P_e using precipitation P:

$$P_e = 0.92 \cdot P - 1.27 \tag{10}$$

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
6.5	7.5	9.0	12.8	13.9	13.9	12.4	10.9	9.4	8.0	7.0	6.0

Table 5. Effective day length for duff moisture code

Then calculating the duff moisture content M_{r_t}:

$$M_{r_t} = 20 + e^{5.6348 - \frac{DMC_{t-1}}{43.43}} + \frac{1000 \cdot Pe}{48.77 + b \cdot P_e} \tag{11}$$

Where b is calculated by:

$$b = \begin{cases} \frac{100}{0.5 + 0.3 \cdot DMC_{t-1}} & , \text{for} DMC_{t-1} \leq 33 \\ 14 - 1.3 \ln(DMC_{t-1}) & , \text{for} 33 < DMC_{t-1} \leq 65 \\ 6.2 \cdot \ln(DMC_{t-1}) - 17.2 & . \text{for} DMC_{t-1} > 65 \end{cases} \tag{12}$$

And finally:

$$DMC_{r_t} = 244.72 - 43.43 \cdot \ln(M_{r_t} - 20), \text{if} DMC_{r_t} < 0 \text{then} DMC_{r_t} = 0 \tag{13}$$

Build Up Index

$$BUI = \begin{cases} 0.8 \cdot \frac{DMC \cdot DC}{DMC + 0.4 \cdot DC} & \text{for } DMC \leq 0.4 \cdot DC \\ DMC - (1 - \frac{0.8 \cdot DC}{DMC + 0.4 \cdot DC}) \cdot [0.92 + (0.0114 \cdot DC)^{1.7}] & \text{for } DMC > 0.4 \cdot DC \end{cases} \tag{14}$$

6.2.1 NDVI

6.3 Figures

1) Extra figures showing other models we tried?
2) Figure of selecting rho for data preprocessing
3) Figure of NDVI sat image for coolness?

6.4 Tables

1) table of results from drought code literature paper SEE DROUGHT INDICES R2 RESULTS.DOCX

		DC			DMC			KBDI		
Species	Site	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Cistus albidus	D06S3	0.52	17.2	13.3	0.57	16.2	12.9	0.46	18.2	13.9
Cistus albidus	D30S2	0.47	18.1	13.9	0.42	18.9	15.4	0.36	19.9	15.2
Cistus albidus	D34S2	0.50	14.4	11.3	0.47	14.8	11.8	0.37	16.1	12.4
Cistus albidus	D83S2	0.64	12.8	9.5	0.25	18.3	15.4	0.62	13.1	9.8
Cistus albidus	D83S3	0.39	17.7	13.9	0.18	20.4	16.7	0.41	17.4	13.6
Cistus mon	D06S2	0.45	20.2	15.8	0.50	19.4	15.3	0.41	21.0	15.7
Cistus mon	D11S2	0.32	16.0	13.0	0.29	16.3	13.1	0.24	16.8	13.5
Cistus mon	D2AS1	0.47	18.5	14.9	0.38	20.1	16.3	0.34	20.7	16.1
Cistus mon	D2BS2	0.08	23.4	18.9	0.23	21.4	17.1	0.06	23.7	19.1
Cistus mon	D2BS3	0.18	21.1	17.2	0.42	17.7	14.4	0.19	21.0	17.0
Cistus mon	D66S1	0.15	23.8	18.7	0.47	18.8	14.7	0.04	25.4	20.3
Cistus mon	D66S2	0.29	22.5	18.8	0.38	21.1	17.5	0.22	23.6	19.4
Cistus mon	D83S1	0.48	17.4	12.8	0.27	20.5	16.4	0.44	18.0	13.1
Cistus mon	D83S3	0.42	20.0	16.0	0.27	22.5	18.6	0.43	19.9	15.7
Erica arborea	D2AS1	0.48	13.7	10.2	0.18	17.1	12.5	0.37	15.0	11.1
Erica arborea	D2AS2	0.51	11.1	8.8	0.45	11.7	8.9	0.47	11.6	9.2
Erica arborea	D2BS2	0.34	12.1	9.0	0.17	13.6	10.0	0.33	12.2	9.2
Erica arborea	D2BS3	0.49	10.5	8.6	0.32	12.1	9.9	0.45	10.9	8.7
Erica arborea	D66S1	0.43	11.0	8.8	0.25	12.6	9.9	0.30	12.1	9.6
Erica arborea	D83S1	0.61	11.2	8.8	0.10	17.1	13.6	0.54	12.2	9.8
Quercus coccifera	D11S1	0.29	5.8	4.6	0.05	6.7	5.3	0.24	6.0	4.7
Quercus coccifera	D13S1	0.40	6.5	4.7	0.10	8.0	6.1	0.36	6.7	4.9
Quercus coccifera	D13S2	0.36	6.2	4.5	0.09	7.4	5.8	0.37	6.2	4.5
Quercus coccifera	D34S2	0.42	5.7	4.3	0.06	7.3	5.8	0.47	5.5	4.2
Quercus coccifera	D84S1	0.54	4.4	3.4	0.10	6.2	4.8	0.56	4.3	3.3
Quercus ilex	D30S2	0.40	6.2	4.8	0.12	7.5	6.0	0.37	6.3	5.0
Quercus ilex	D83S2	0.43	4.2	3.3	0.06	5.4	4.3	0.45	4.1	3.3
Quercus ilex	D84S1	0.56	4.5	3.6	0.06	6.6	5.4	0.61	4.3	3.4
Quercus ilex	D84S2	0.51	6.6	4.7	0.23	8.2	6.6	0.42	7.1	5.1
Rosmarinus off	D06S1	0.67	16.5	13.2	0.58	18.4	14.8	0.61	17.8	14.0
Rosmarinus off	D11S1	0.46	16.8	13.9	0.39	17.9	14.5	0.39	17.9	14.9
Rosmarinus off	D13S1	0.41	18.7	15.3	0.50	17.1	13.6	0.26	20.9	17.0
Rosmarinus off	D13S2	0.47	14.7	11.2	0.45	15.0	12.1	0.32	16.7	12.8
Rosmarinus off	D84S2	0.45	24.4	20.2	0.60	20.7	16.1	0.31	27.2	23.2

2) All results from Hierarchical modelling
SUMMARY TABLE OF OLS POOLED BMS
SUMMARY TABLE OF OLS UNPOOLED BMS
SUMMARY TABLEL OF VAR INTERCEPT MODEL BMS
SUMMARY TABLE OF VAR SLOPE BMS 11 MODEL
TABLE SHOWING RANEF COEFFCIENTS of BMS 11

SUMMARY TABLE OF OLS POOLED PCA
SUMMARY TABLE OF OLS UNPOOLED PCA
SUMMARY TABLE OF VAR INTERCEPT MODEL PCA
SUMMARY TABLE OF VAR SLOPE BMS PC 22 MODEL
TABLE SHOWING RANEF COEFFICIENTS of PC 22

	BMA Estimate	2.5%	97.5%	$p(\cdot \mid \gamma)$
Station Elevation	-3.370478	-5.127001	-1.613143	0.9999536
G	6.350300	4.102470	8.593709	0.9807483
H_{min}	4.750553	2.585580	7.093371	0.9960038
Drought COde	-8.478542	-11.477374	-6.585258	1.0000000
Solar_Rad_Max_MEAN7	8.338130	0.000000	12.918999	0.8532256
rel_humid_min_MEAN15	16.044091	9.492875	22.604520	0.9999944
rel_humid_MEAN15	-9.780191	-16.890643	-2.822964	0.9954583
temp_min_MEAN15	-8.613416	-14.058442	-3.061096	0.9930988
evap_pot_MEAN15	19.613844	14.465487	24.875580	0.9941110
D_solar_rad_avg	13.535149	8.116125	19.126382	0.9785908
D_temp_min	20.573545	13.743946	33.975686	0.9871351
D_evap_pot	-18.391673	-25.613727	-10.672422	0.9848089
D_DC	-4.055092	-6.317401	-1.682868	0.9774863
D_solar_rad_max_MEAN7	32.617237	22.700637	41.218592	0.9597782
D_wind_speed_max_MEAN15	-26.134160	-40.647366	0.000000	0.8690481
D_temp_max_MEAN15	-152.708005	-204.046987	-106.004539	1.0000000
D_temp_min_MEAN15	-54.003196	-69.996789	-38.484842	0.9678281
D_temp_avg_MEAN15	183.904898	168.929414	209.749593	0.9999964
ANDVI	-28.178194	-30.559631	-25.766596	1.0000000
leaf_nitr_mass_ANDVI	16.067462	14.192315	17.920441	1.0000000
leaf_phos_ANDVI	36.670841	34.630477	38.768370	1.0000000
plant_height_ANDVI	-4.468335	-5.291190	-3.646047	0.9998859
SLA_ANDVI	-10.777184	-11.517194	-10.067203	1.0000000

Table 6. BMA estimates of variables with high probability.

Summary tables SEE RESULTS R.ODT =¿ NEED SUMMARY TABLES
3) PCA summary results

Table 7.

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	7.633	0.498	0.498
PC2	3.354	0.096	0.594
PC3	2.625	0.059	0.653
PC4	2.588	0.057	0.710
PC5	1.932	0.032	0.742
PC6	1.881	0.030	0.772
PC7	1.790	0.027	0.800
PC8	1.510	0.020	0.819
PC9	1.480	0.019	0.838
PC10	1.389	0.016	0.855
PC11	1.261	0.014	0.868
PC12	1.247	0.013	0.881
PC13	1.206	0.012	0.894
PC14	1.064	0.010	0.904
PC15	1.031	0.009	0.913
PC16	0.986	0.008	0.921
PC17	0.909	0.007	0.928
PC18	0.860	0.006	0.934
PC19	0.821	0.006	0.940
PC20	0.757	0.005	0.945
PC21	0.739	0.005	0.950
PC22	0.721	0.004	0.954

4) Coefficient estimates of best tables

Table 8.

	Fuels	OLS Pooled	OLS Unpooled	Intercept Only	BMS Vars 1	BMS Vars 2
1	buckwheat, eastern mojave	0.464	0.455	0.439	0.444	0.381
2	ceanothus, bigpod	0.597	0.668	0.646	0.685	0.704
3	ceanothus, hoaryleaf	0.572	0.710	0.726	0.712	0.677
4	chamise	0.310	0.365	0.364	0.547	0.538
5	chamise, new growth	0.446	0.469	0.472	0.581	0.567
6	sage, black	0.432	0.519	0.518	0.635	0.627
7	sage, purple	0.421	0.427	0.428	0.440	0.400
8	sagebrush, black	0.517	0.520	0.517	0.639	0.623
9	sagebrush, california	0.512	0.516	0.513	0.614	0.616
10	Total	0.561	0.597	0.596	0.688	0.680

Table 9.

	Fuels	OLS Pooled	OLS Unpooled	Intercept Only	PC 22	PC 6	PC Min
1	buckwheat, eastern mojave	0.142	0.306	0.293	0.385	0.478	0.449
2	ceanothus, bigpod	0.541	0.638	0.637	0.756	0.734	0.713
3	ceanothus, hoaryleaf	0.605	0.755	0.752	0.713	0.677	0.669
4	chamise	0.093	0.324	0.320	0.530	0.496	0.488
5	chamise, new growth	0.266	0.412	0.415	0.578	0.541	0.509
6	sage, black	0.118	0.437	0.437	0.625	0.641	0.633
7	sage, purple	0.385	0.399	0.391	0.386	0.393	0.403
8	sagebrush, black	0.421	0.450	0.453	0.654	0.566	0.543
9	sagebrush, california	0.383	0.456	0.451	0.630	0.602	0.579
10	Total	0.401	0.547	0.546	0.685	0.670	0.659

5) Anova tables

Table 10.

	npar	$\log p(\mathbf{y} \mid \mathbf{x})$	AIC	LRT	Df	Pr(>Chisq)
<none>	134	−$24,711.260	49,690.530			
H	122	−24,717.170	49,678.340	11.811	12	0.461
H_{mean7}	122	−24,736.930	49,717.860	51.330	12	0.00000
solar_rad_max_MEAN3	122	−24,713.740	49,671.470	4.948	12	0.960
DC	122	24,848.630	−49,941.270	274.739	12	0
temp_min_MEAN15	122	−24,733.510	49,711.030	44.500	12	0.00001
wind_speed_avg_MEAN15	122	−24,731.650	49,707.290	40.763	12	0.0001
$D \cdot DMC$	122	−24,726.060	−49,696.130	29.604	12	0.003
$L_{nitro} \cdot A_{NDVI}$	122	−24,715.640	49,675.280	8.748	12	0.724
D_wind_speed_max	122	−24,755.600	49,755.210	88.682	12	0
G_{max}	122	−24,714.980	49,673.960	7.428	12	0.828
$D \cdot H$	122	−24,718.760	49,681.520	14.994	12	0.242

Table 11.

Principal Component	npar	Likelihood	AIC	LRT	df	$P(>\chi^2_{df})$
	52	-25,043.690	50,191.380			
1	45	-25,537.270	51,164.540	987.165	7	0
2	45	-25,078.580	50,247.160	69.785	7	0
3	45	-25,128.690	50,347.380	170.005	7	0
4	45	-25,047.710	50,185.420	8.041	7	0.329
5	45	-25,069.010	50,228.010	50.636	7	0
6	45	-25,091.280	50,272.560	95.180	7	0

6.5 Variable Name definitions

1. D_t - Day of year
2. P - Accumulated Precipitation

3. DC - Drought Code
4. DMC - Duff Moisture Code
5. G - Solar Radiation
6. T - Temperature
7. H - Relative Humidity
8. W - Wind Speed
9. V_{height} - Plant Height
10. BUI - Build up Index
11. L_{phos} - Phosphorous per dry mass of leaf.
12. L_{nitro} - Nitrogen per dry mass of leaf.
13. SLA - Specific Leaf Area.

References

Castro, F., Tudela, A., and Sebastià, M. T. (2003). Modeling moisture content in shrubs to predict fire risk in catalonia (spain). *Agricultural and Forest Meteorology*, 116(1-2):49–59.

Chuvieco, E., Cocero, D., Riano, D., Martin, P., Martınez-Vega, J., De La Riva, J., and Pérez, F. (2004). Combining ndvi and surface temperature for the estimation of live fuel moisture content in forest fire danger rating. *Remote Sensing of Environment*, 92(3):322–331.

McCandless, T. C., Kosovic, B., and Petzke, W. (2020). Enhancing wildfire spread modelling by building a gridded fuel moisture content product with machine learning. *Machine Learning: Science and Technology*, 1(3):035010.

Ruffault, J., Martin-StPaul, N., Pimont, F., and Dupuy, J.-L. (2018). How well do meteorological drought indices predict live fuel moisture content (lfmc)? an assessment for wildfire research and operations in mediterranean ecosystems. *Agricultural and Forest Meteorology*, 262:391–401.

Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298.