January Update 2021

07/01/2021

## Topic for discussion:

1. Modeling methods/process
2. Which accuracy measure to use? standard metrics, aspatial metrics
3. Overall accuracy measures
4. Fuzzy/alternate call metrics
5. Imbalanced datasets
6. MapUnit (class) accuracy:

### 1) Modelling methods/process

Over the last month, we have been working through the model development (converting modeling to tidymodel format/package) and ensuring modeling follows both best practices and accounts for sampling structure (i.e paired transects).

Thanks to input and advice from Brandon we are using a 10 x 5 cross validation to assess fit after an intial training and testing split. This process is bootstrapped for each “slice” of input data derived from the sliced latin hyper cube sampling. Each slice represents similar hypercube “space” and contains approximately 5 sites.For example stage 1 SBSmc2 at Deception includes 5 slices each with 5 sites = 25 sites or 50 triangles in total. This provides 1) confidence intervals around accuracy measures (based on 5 slices), 2) minimises “data leakage” between the test and training sets, 3) enables sites to be “blocked” to training and testing sets to prevent the model from training and testing on data from the same site (ie pair of transects). Cross validation is stratified by mapunit and grouped by site id. Test results are reported below.

For rapid comparisons an initial tuning of hyperparameters was completed with these parameters applied to all model comparisons shown below. For final comparisons each model will be indivdually tuned.

In addition to boostrapping by slices, We also tested bootstrapping via sites, however this produced a very wide range of confidence intervals due to diverse sites (ie high proportion on non-forest sites).

Modelling code has been developed for the following packages (mlr, tidymodels, caret). To date tidymodels provides best function and documentation to support multiple machine learning models and results collection. Very similar results were obtained when testing the same model in two packages (tidymodels, mlr).

We used a small subset of data (SBSmc2) from Deception stage 1 to investigate a number of preliminary questions which arose from results to date.

### 2) Which accuracy measure to use?

### **Standard metrics**:

* A number of metrics can be used to assess machine learning model accuracy: **accuracy (spatial accuracy), precision, kappa, recall, sensitivity, specificity**.
* **Accuracy** is a common metric, however, it is sensitive to imbalanced data, as it overpredicts common class at the expense of minor classes.
* We explored the literature to assess other metrics which we could be used to assess success across classes (ie **f-measure, j index, mcc**). See additional notes for definitions.

### **Aspatial metrics**:

We explored aspatial assessment metrics, which are based on the proportion of mapunit which exists rather than exact spatial match. This is equivalent to existing AA measures.

* **Aspatial\_acc** is the accuracy per slice (i.e total accuracy over the site).
* **Aspatial\_meanacc** is the accuracy based on the average of map units (ie: 100% correct = 0% correct).

### 3) Overall Accuracy Metrics:

The following figures provides an example of a single model run (5m all training points) with the three types of overall accuracy. Results from the cross validation and final test are included for comparison.

**Aspatial**: - Aspat\_p\_acc (aspatial prime accuracy per slice) - aspat\_p\_meanacc (aspatial prime average accuracy per mapunit)

**MultiClass**: - f\_means - mcc(matthews correlation coefficient)

**Spatial**: - acc\_spatial (accuracy) - kap(kappa) - precision - recall - sens(sensitivity) - spec(specificity)

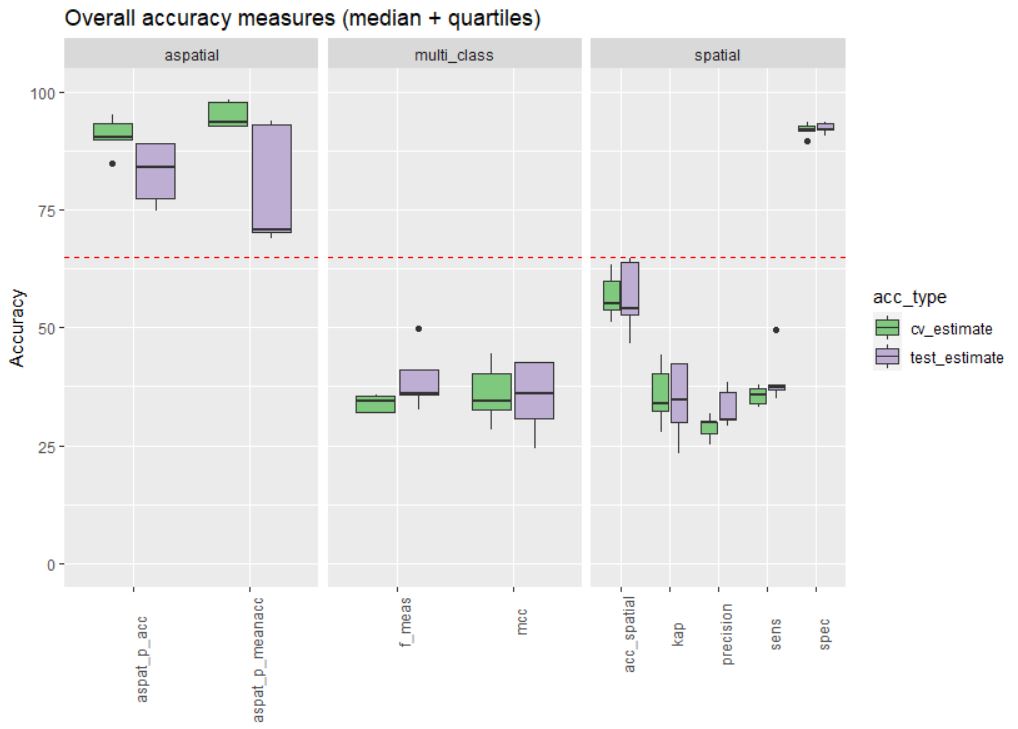


Fig.1: Overall accuracy metrics options (test and cv) for a single model output.

### **Take aways**:

* Wide range in accuracy measures depending on method chosen.
* Number of metrics similar (f\_means, mcc, kappa, sensitivity)
* According to aspatial accuracy measures we are succeeding at reaching > 0.65 accuracy.
* Similar results for test and CV can be used to assess overfitting (or lack of).

**Proposed Metrics** :

* aspatial average (aspat\_p\_acc) - equivalent to current AA measure
* aspatial mapunit accuracy (aspat\_p\_meanacc) - map unit specific measure of AA
* spatial accuracy (acc\_spatial) - accuracy equivalent
* mcc - measure of multi-class accuracy

### 4) **Alt/Fuzzy calls**:

In reality we have multiple forms of fuzziness. These represent fuzzyiness on the ground, ie non-central concepts calls, or conceptual fuzziness between units. We tested a number of options to incorporate this into the accuracy measures:

**Alternate calls** - (fuzziness in reality) can be included in accuracy measure similar to AA method in which the secondary call is included if it matches the predicted call.

**Fuzzy** (conception) values assign partial correct values for mapunit calls that are similar (on the edatopic position) to the correct calls. In this case, scores could be awarded for on a sliding scale from 1 (Correct) to 0 (no where close) with partial credit assigned to closely related mapunits. Note this requires a matrix which specifies the similarity between all combinations of possible calls. This was also calculated for primary and alternate calls (spat\_fpa).

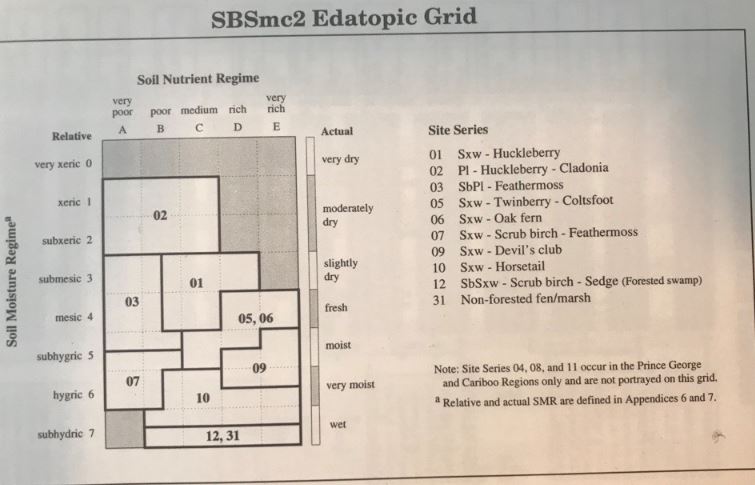
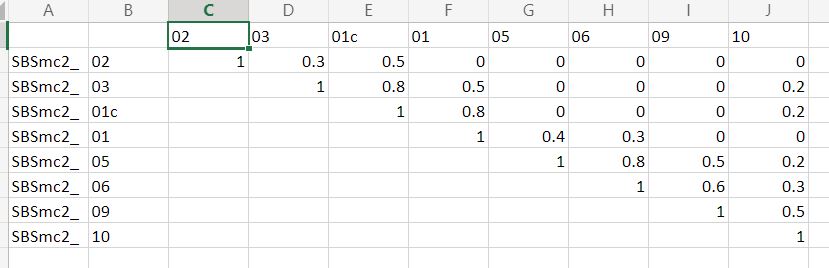


Fig.2: Edatopic grid and associated values



An example of the proposed metrics including fuzzy and prime/alt calls includes;

-aspat\_p\_acc (overall apatial accuracy) -aspat\_p\_meanacc (average mapunit aspatial accuracy) -mcc -spat\_p\_acc (accuracy) -spat\_pa\_acc (prime / alt call accuracy) -spat\_fp\_tot (fuzzy prime) -spat\_fpa\_tot (fuzzy prime/alt accuracy)

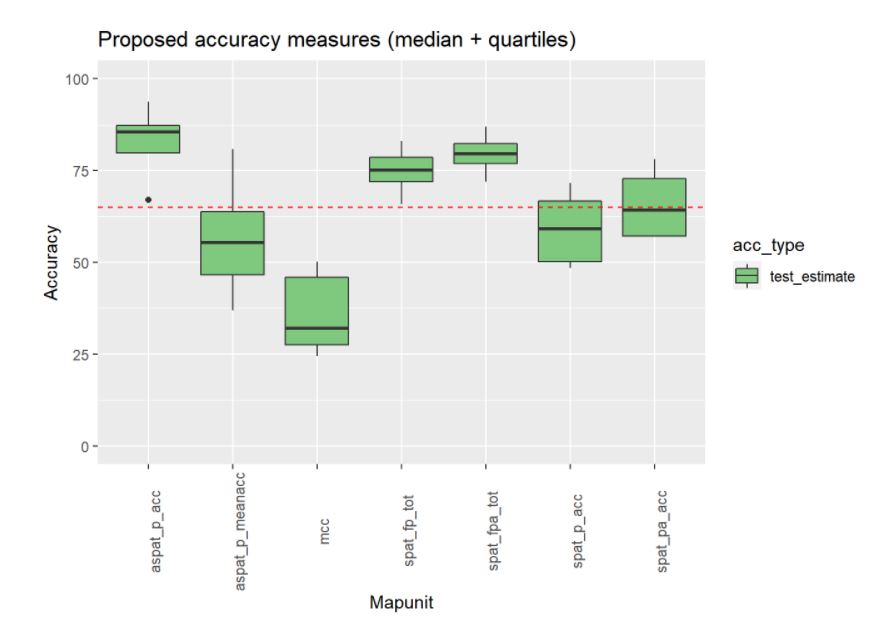


Fig.3: Proposed accuracy measures).

### 5) Imbalanced datasets

As accuracy is impacted by imbalance of training data we also assessed training point balance on model accuracy. We tested a variety of balancing methods; raw training pts, raw pure training points, smoting (upsample uncommon sites at various ratios), downsample common sites.

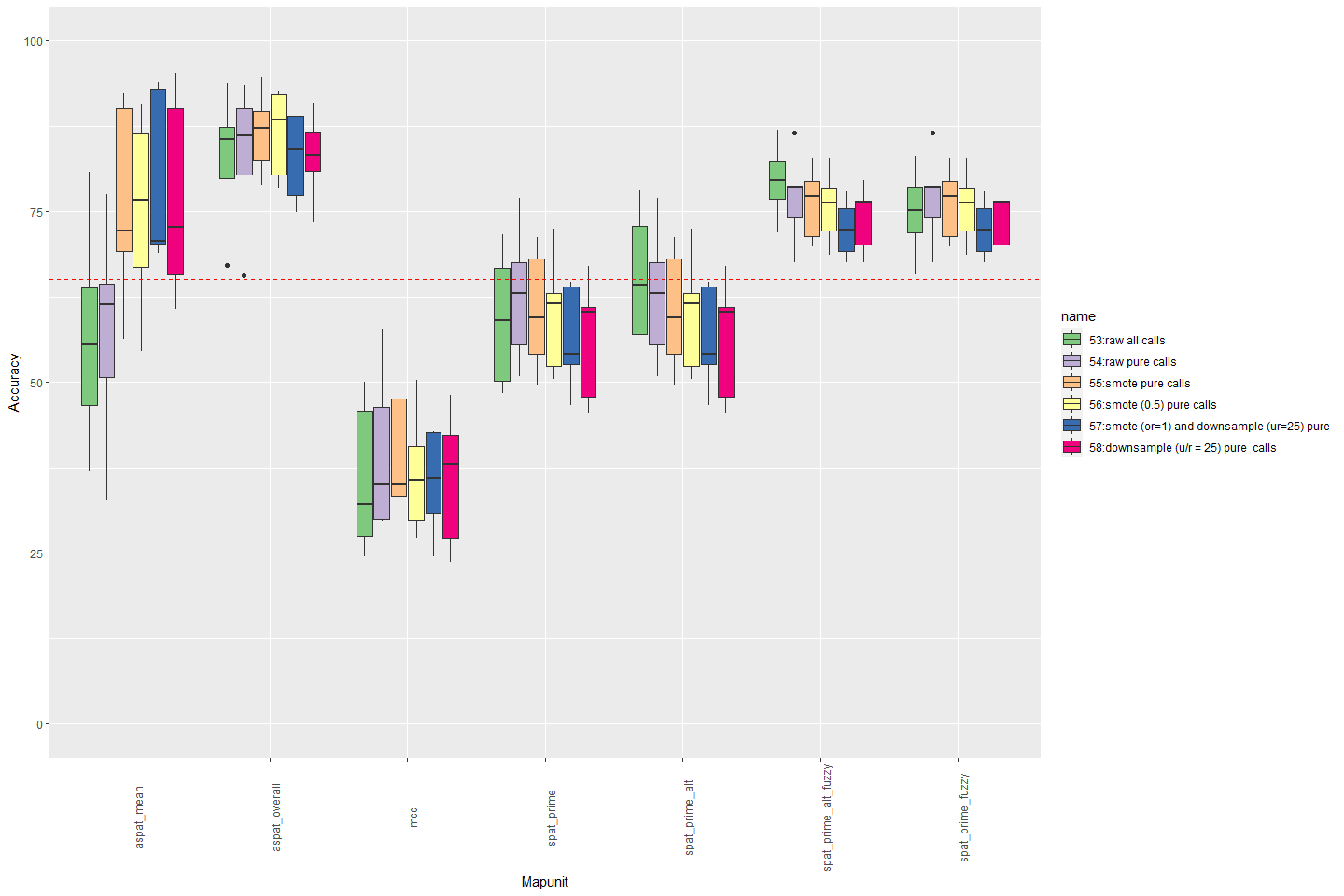


Fig.4:Test accuracy with balancing options (median + quartiles)

**Takeaways**:

* balancing (smote/downsample) impacts the accuracy measures particularly aspatial mapunit measures.
* acccounting for fuzziness (prime or prime/alt calls) greatly increases the overall accuracy measures.
* Balancing requires more careful consideration given the changes in accuracy.
* surprisingly mcc not very different with balancing.

### 6) Accuracy per MapUnit

In addition to overall accuracy, we considered how accuracy measures varied in assessing accuracy per mapunit. An example of model outputs is below.

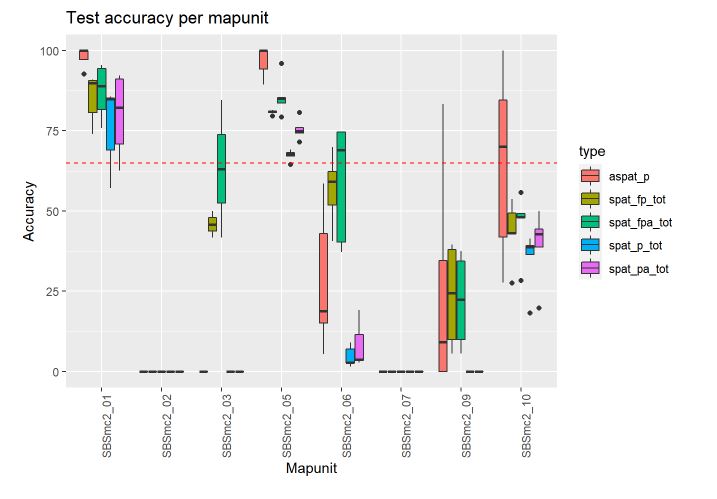


Fig.5: Example of map unit accuracies for model 53 (raw all calls)

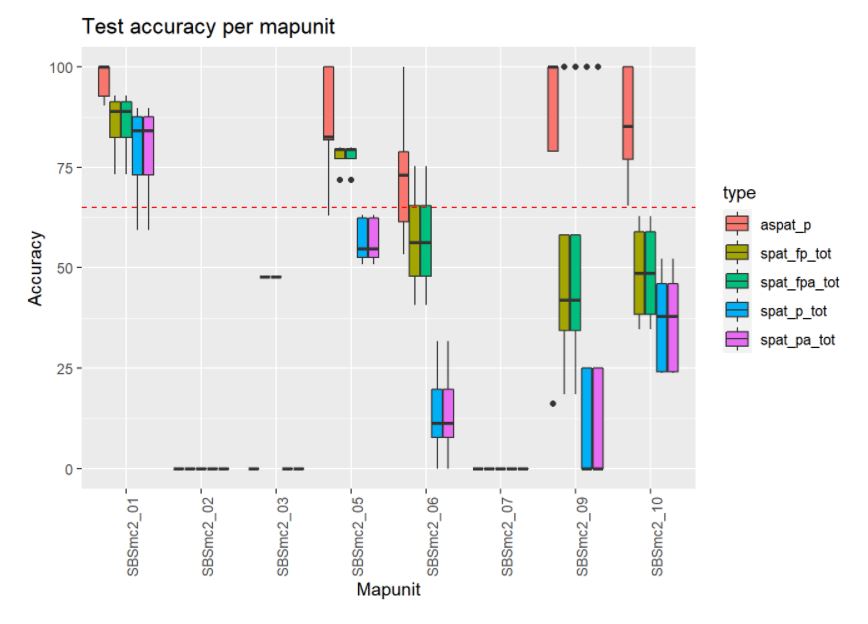
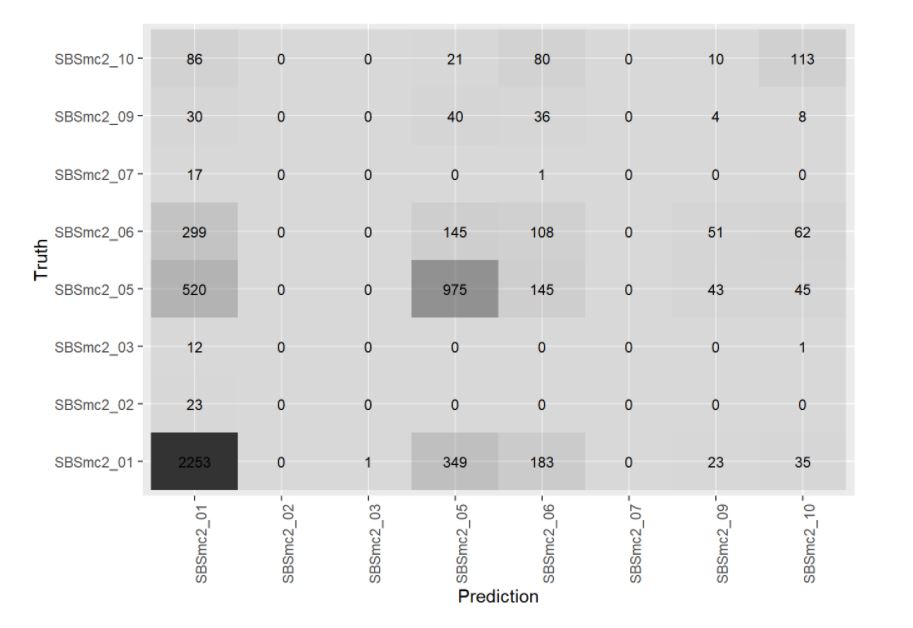
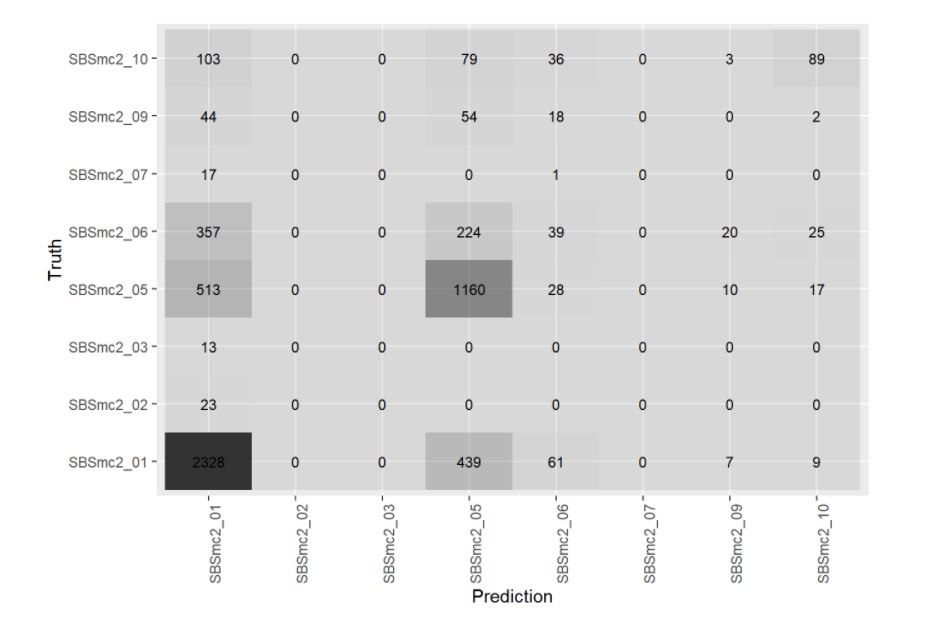


Fig.6: Example of map unit accuracies for model 56 (smoted (0.5) pure call

Comparing the raw model of pure calls (54) to the partial smoted model (56). When looking at the confusion matrix smoting model makes high impact on uncommon sites.



### Take aways:

* aspatial performs much better than spatial.
* sites with very few training pt (SBSmc2\_02/ 03/ 07) had little difference with balancing
* sites with moderate amount of data (SBSmc2\_09) showed improved metrics in confusion table.
* uncommon mapunits require more data
* pure calls remove before training (stage 1 entire training data)

### General comments/notes

Why does aspatial score higher than spatial?

Discuss with Bob as to why?

1. Field calls variable among the data collection.
2. Pattern becomes closer to the truth as classes are rolled up.
3. Importance of landscape scale (multi-scale) and some spatial displacement of classes on the ground.
4. Issue of fuzzy boundary (again influence of individual samplers)

Solutions/Possible options: - View as a population level sample rather than explicit - Potential to use individual crew as a measurement error? or account for difference in prime and alt calls. - Use a fuzzy matrix and or alt calls.