Conducting small area estimation with the UNFPA's SAEApp

Introduction

This document provides an introduction the the UNFPA small area estimation application. Small area estimation is a statistical method often used to combine high-resolution census data with subject-specific but lower-resolution household surveys. This application provides a user-friendlty interface for conducting a small-area estimation analysis. It includes an interface for loading numerical census and survey data, as well as shapefiles for display of results. Users are directed to choose the relevant indicatorsfrom survey data, as well as variables that are common to both survey and census data as a basis for building a linear model. The application presents methods for checking the suitability of variables for modeling. When the user has constructed a suitable model, the app allows for diagnostic testing of model results and finally prediction of the results into relevant census units. Users can export analytic steps as tabular or figure results.

Background

The methods used in this application were described by the UNFPA's Small Area Estimation Guidance.

Analytic Approach

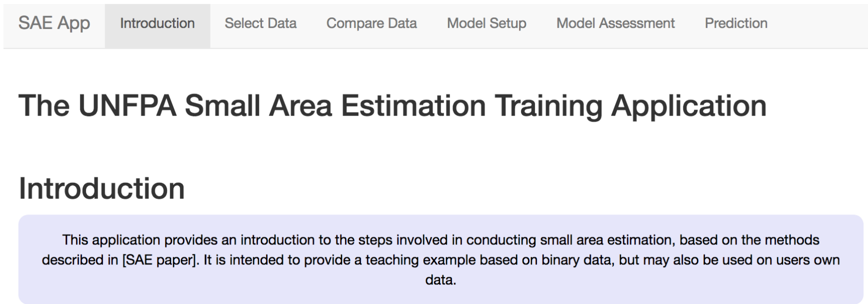
Population demographic data found in international household surveys often use binary responses to estimate prevalence rates in populations. Because of this, the application uses a logistic regression modeling approach. In addition, the distribution of effort in population health surveys is not equal within the regions, meaning that over-representation of sampling effort in some areas may bias results in others. For this reason, the application allows for use of a random effects model.

At present, the app is not suitable for continuous response variables. Future iterations may provide this functionality.

How to use the application The application consists of different tabs for each stage of the analysis. These include selecting and comparing data, Setup of the models, model assessment and finally prediction. Under each tab, simple directions allow the user to control outputs and download results. Note that the contents of each tab depend on the previous tab results, so errors in data setup will propogate through the application and may result in error messages instead of expected results.

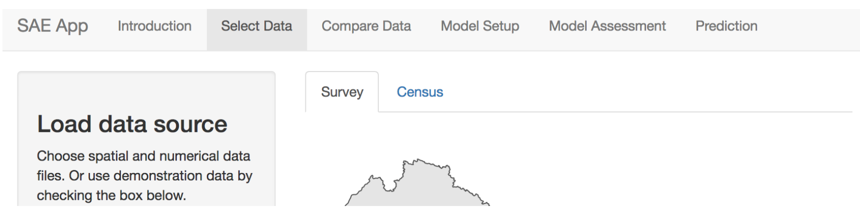
Landing Page

The landing page provides the gateway to the application features. Use the tabs at the top of the page to access different parts of the application.



Select Data

The Data selection tab allows users to either use sample DHS and survey datasets from Nepal, or to load the their own data. Currently the application does not allow for formatting of data, so all numerical data must be in a harmonized prior to being entered into the application.

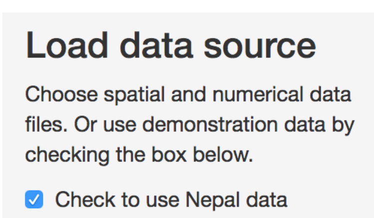


For a detailed description of the harmonization process, see ______.: At minimum,

1. Item. Survey and census data must share exactly the same variable names
2. Item. Variables that are common between survey and census must have the same definitions. For example, categorical variables must have the same number of classes, and these classes must be identical between the two sources
3. Item. Variables must have roughly similar distributions between census and survey datasets.

Code used to conduct harmonization are included in the appendix.

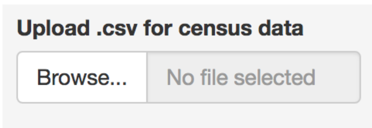
Survey and census data each have their own tabs to assist with data loading. By default, data from Nepal will load, as the 'Use Nepal data' checkbox will be checked. To use other data, uncheck this box.



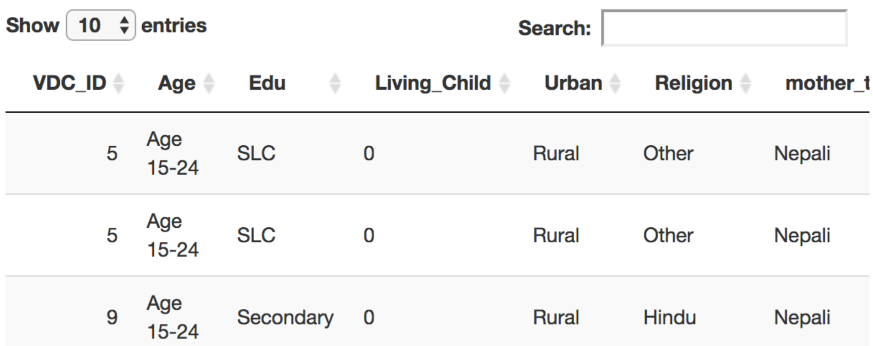
Survey tab

Survey data should include three different types of data: the indicator of interest, a column for the names of the survey regions used and additional columns for variables that will be used in the predictive model. If loading an external file, uncheck the 'Use Nepal Data' box and click on the 'choose a .csv file' box to select a file.

Figure 1: Survey Tab

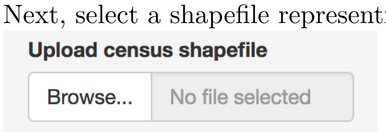


Once you have selected a file, it should appear at the bottom of the page.



VDC_ID	Age	Edu	Living_Child	Urban	Religion	mother_t
5	Age 15-24	SLC	0	Rural	Other	Nepali
5	Age 15-24	SLC	0	Rural	Other	Nepali
9	Age 15-24	Secondary	0	Rural	Hindu	Nepali

Next, select a shapefile representing the survey regions.



At present, only ESRI shapefiles are accepted. it is also important that when selecting the shapefile, all associated files are selected. Once a shapefile has been loaded it should appear at the top of the page.

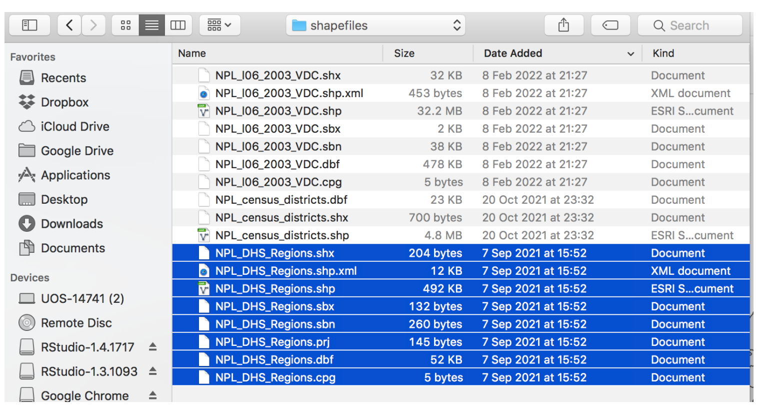


Figure 2: Choosing all files for a shapefile

Once the shapefiles are loaded, a map will appear, unless there has been an error.

When the survey data have been loaded, you should select which variable will be the indicator to model, using the drop-down menu.

Note that only binary variables are currently supported. Variables should be formatted as ‘0’ or ‘1’ only.

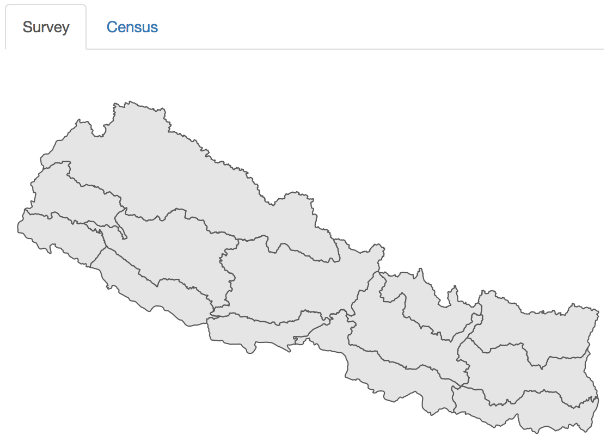


Figure 3: Map of survey areas

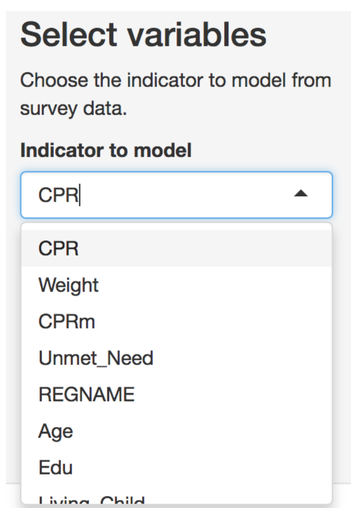


Figure 4: Indicator selection tab

Users should then select which variable represents the region or strata names in the data, as these will be used both in mapping and in creation of fixed or random effects. These region names should also be present and identically named in the census data to ensure accurate matching between the datasets.

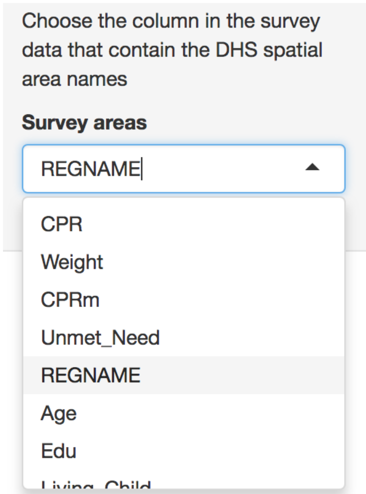


Figure 5: Survey area names

Census

When selecting the census tab, default the map and table information will take some time to load as the datasets are larger.

Figure 6: The Census tab controls upload and structuring of census data

After the survey data are loaded, Census data should be loaded in a similar way. Note that the census data should include both the same survey stratum names as the survey data, as well as an additional spatial stratum representing the area at which predictions will be made.

Users then can select the predictor variables that will be. Note that only predictors that appear in both survey and census data can be selected. To reduce confusion, it is best that the loaded data only include variables that will be relevant to the analysis. However, these variables can be removed later if they are not needed.

Compare Data

Predictive modeling of small areas based on survey data relies on the assumption that the populations sampled in the survey are a random sample of the census. As such, the types of respondents in census and survey data should be present in similar ratios for each. Therefore, before modeling the indicator of interest, it is important to compare the distribution in each.

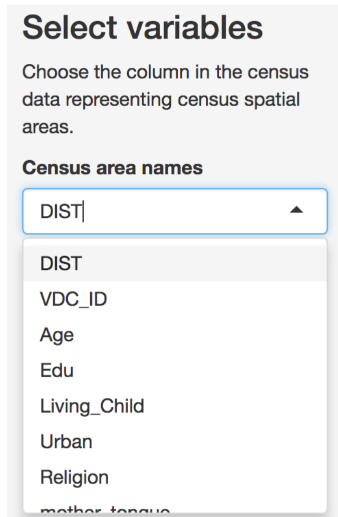


Figure 7: Census spatial area selection

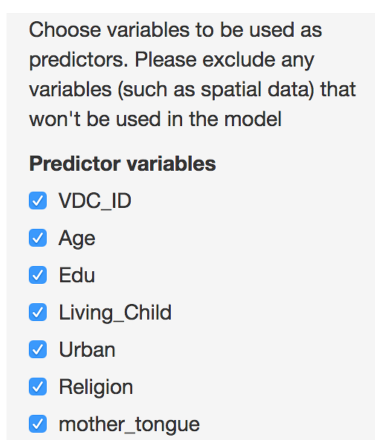


Figure 8: Select predictor variables

Correlations

A simple way to check whether the two data types are related is to look at how closely they are correlated. This can be difficult to do with binary data, so we aggregate these data at the level of survey regions and visualize as a scatterplot. By default these correlations are only shown for variables checked in the 'Data Selection' tab.

When opening the **Compare data** Tab, the first tab displays a scatterplot of the correlations between data variables. These plots are based on data aggregated at the survey region level. Variables that appear linearly correlated between survey and census shold be considered for inclusion in the model. Those that are not should be excluded, unless there is some reason to explicitly include them.

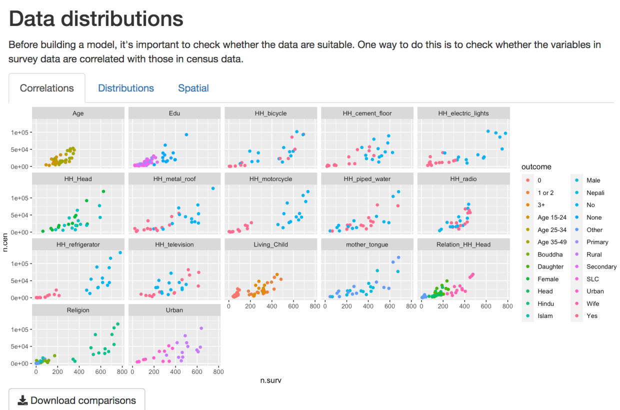


Figure 9: Scatterplot of distributions of variables at survey region level

Distributions.

For a closer look at individual predictors, in the distributions tab we look more closely at the distributions of individual variables to see how different levels of each variable correlate. Use the drop-down menu to select which variable to view. Here we can look more closely at the different levels of each variable.

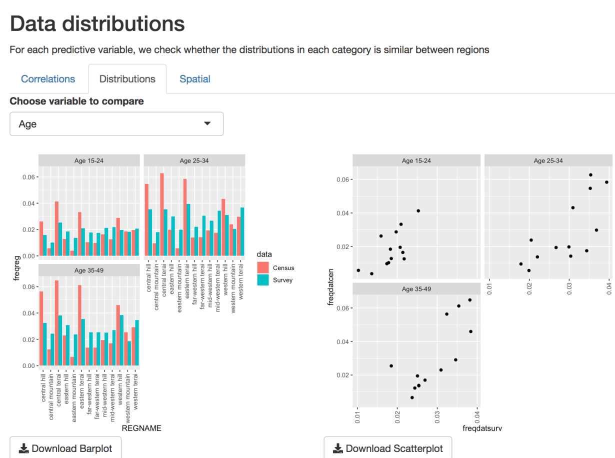


Figure 10: Barplots and scatterplots of individual variables

Spatial

Finally, it's important that the spatial distribution of effort is similar in survey and census data. The map in the spatial section shows the relative effort applied in each survey region.

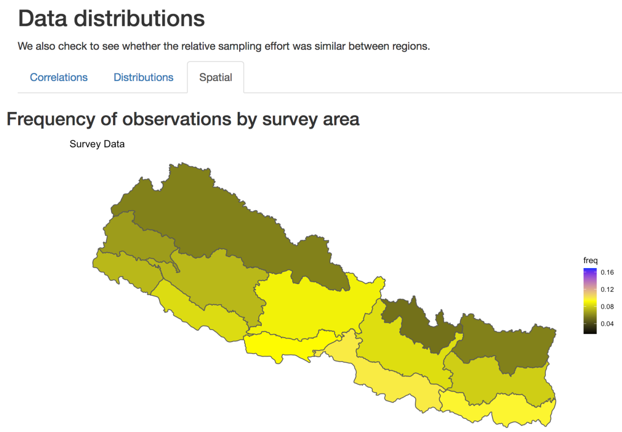


Figure 11: Spatial distribution of sampling effort

Model Setup

The model setup tab provides tools to further check the suitability of variables for prediction before building the model. The model building component then allows you to see how inclusion of different parameters influences the model outcome.

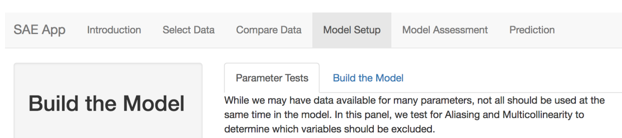


Figure 12: Model building tab

Parameter Tests The parameter tests component uses alias testing and variance inflation factors to look for variables that are highly correlated with one another. The Alias test button looks for variables that are perfectly correlated, and must be run before the Collinearity or Variance Inflation Factors (VIF) test. The VIF test checks each variable for variance inflation. Run the Alias test first, as the VIF test cannot run on variables that are nearly identical.

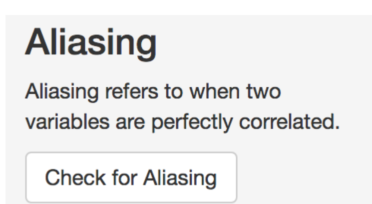
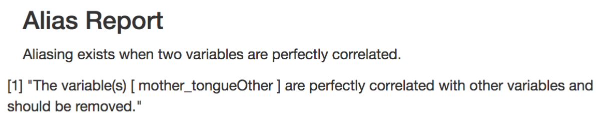


Figure 13: Alias test button

The results of the Alias test will indicate whether there are variables that need removing from the candidate variable set.



When variables are identified in the Alias test, you can remove them from the list of parameters

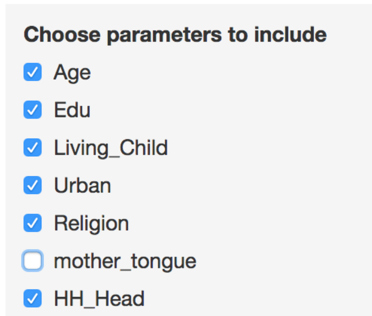


Figure 14: Removal of 'mother tongue' as a predictor

After removing the indicated variables from the list of predictors, you should re-run the alias test, then you can run the VIF test to test for collinearity.

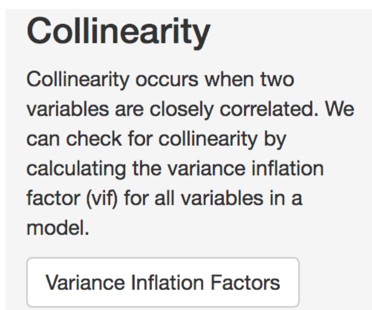


Figure 15: Collinearity button

In the VIF test, results show a GVIF score and a degrees of freedom (DF). Variables that have a GVIF score of higher than 2 should be removed also, as including them may lead to unexpected predictions.

Variance Inflation Table

Collinearity occurs when multiple variables are highly correlated with one another. Such correlations can make it difficult to estimate the coefficients for those variables. Variance inflation factors help estimate the severity of multicollinearity in individual variables. Variables with a VIF score of greater than 2 should be excluded.

Show 10 entries Search:

	GVIF	Df
Age	1.7	2.0
Edu	1.9	3.0
Living_Child	1.7	2.0
Urban	1.3	1.0
Religion	1.2	3.0
HH_Head	1.2	1.0
Relation_HH_Head	1.4	3.0
HH_radio	1.1	1.0
HH_television	1.7	1.0
HH_motorcycle	1.4	1.0

Showing 1 to 10 of 16 entries Previous 1 2 Next

Build the Model

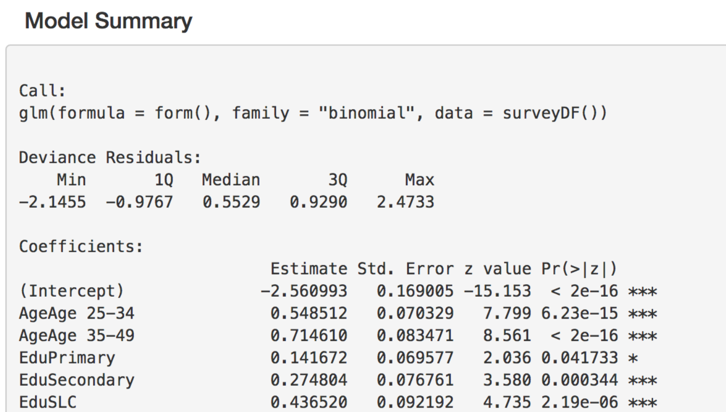
Once the Variance inflation test has been done, you can proceed to building the model by selecting the 'Build the Model'. By default, the model consists of all variables that were not deselected in previous steps. As well, the spatial regions are by default included as a fixed effect.

Parameter Tests | Build the Model

Model Formula

CPR ~ Age + Edu + Living_Child + Urban + Religion + HH_Head + Relation_HH_Head + HH_radio + HH_television + HH_motorcycle + HH_bicycle + HH_refrigerator + HH_cement_floor + HH_metal_roof + HH_piped_water + HH_electric_lights + REGNAME

The model results are displayed as a summary table. In this summary, the estimates of the model coefficients are displayed on the logit scale, along with other statistics. Significance of each coefficient is indicated with stars, where * = significance at p=0.05, ** = significance at p=0.01, *** = significance at p=0.001.



Model Summary

```
Call:
glm(formula = form(), family = "binomial", data = surveyDF())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1455  -0.9767   0.5529   0.9290   2.4733

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.560993   0.169005 -15.153  < 2e-16 ***
AgeAge 25-34              0.548512   0.070329   7.799 6.23e-15 ***
AgeAge 35-49              0.714610   0.083471   8.561  < 2e-16 ***
EduPrimary                0.141672   0.069577   2.036 0.041733 *
EduSecondary              0.274804   0.076761   3.580 0.000344 ***
EduSLC                    0.436520   0.092192   4.735 2.19e-06 ***
```

Every time a variable is added, the model is re-run, and a new summary table is built. If you wish to save the results of a particular model, you can save them using the download button.

Stepwise Regression.

While you may have predictors that you wish to include, but it may be difficult to decide which combination of variables is best to use. In this case, you may use the stepwise variable selection button to choose the most parsimonious set of variables. The tool uses both forward and backwards stepwise selection, and retains the model with the lowest AIC score. The selected variables will remain checked in the list.

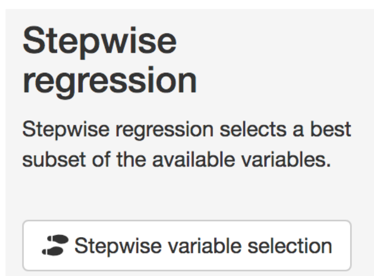
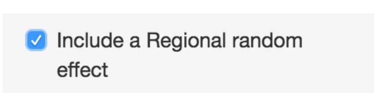


Figure 16: Stepwise selection button

Random effects

Because the sampling effort in each of the survey regions can often be different, treating thema s a fixed effect may not be the most reliable way to account for regional variations. Instead, it is possible to include spatial regions as a random effect. In this way, oversampling in some areas will not skew results, and the uncertainty in estimates will be properly accounted for. Checking the ‘include random effects’ box will change the fixed effects to random effects in the model.



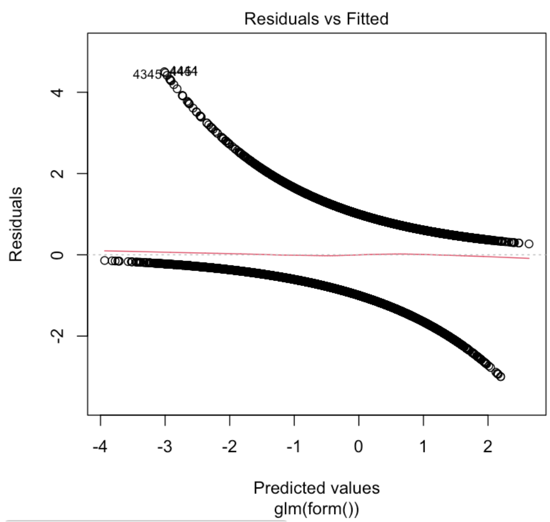
Note that including random effects in the model will make it much slower to run. It is best to first run the variable selection before including the random effect term.

Assessment of model fit

Following fitting of the model, it is important to check whether the model fit is suitable

Assessment

The 'Assess' tab two a few ways to check the fit of the model. The first is the residual plot, which shows each of the predicted response values used in the model against the difference between the observed and predicted values. A model that has equal dispersion of points above and below the line indicates an unbiased fit.



We can also look at howthe observed values compare to the fitted values. With binary variables this can be challenging to intepret, se we can look at observed and predicted values aggregated in survey regions. Note that because we have included survey regions as a fixed or random effect, we expect that the model will usually perform very well at this aggregated level.

Validation

Another way to evaluate the predictive power for binomial random variables in a logistic model is to look at the percentage of times the model correctly classifies the response variable. A Confusion matrix describes the percentage of time the model correctly predicts a true positive or true negative. A better-performing model should have higher true positive and negative rates than false rates.

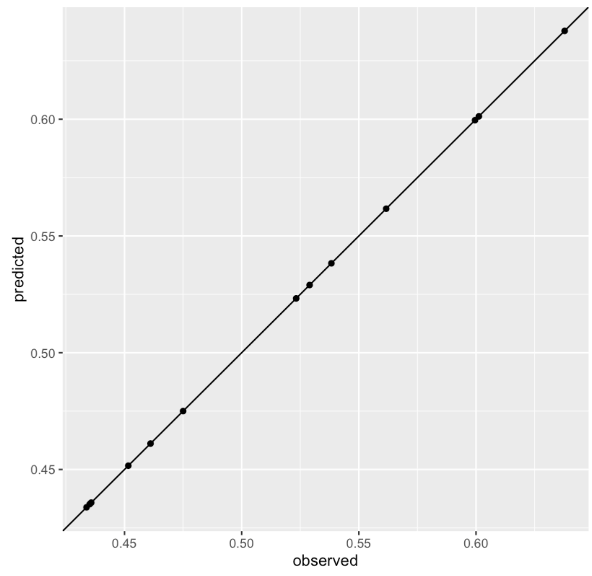
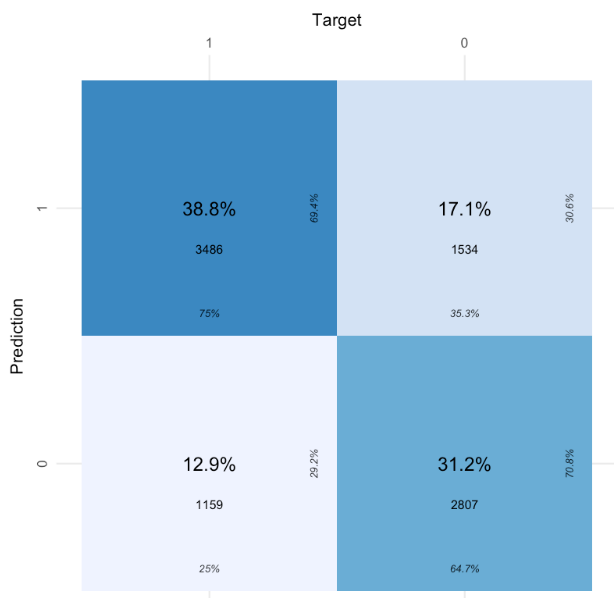
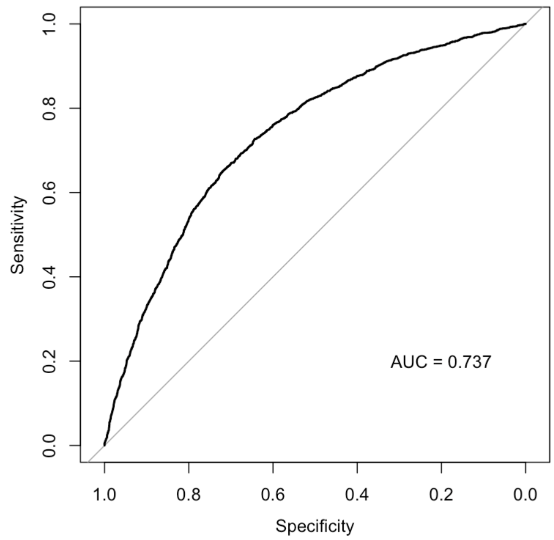


Figure 17: Observed verus predicted outcomes aggregated at survey region level



Another way to look at model performance is to examine the receiver-operator curve or ROC. This curve plots 'Sensitivity' against 'Specificity', which can be described as the rate of false positives (y axis) at each level of true positivity (x axis). An ROC curve that is more convex (or farther from the diagonal) suggests better model performance. The area under the ROC curve (or AUC) provides a measure of the model accuracy; an AUC of 0.5 means the model cannot predict with any accuracy, while an AUC means it is perfectly accurate. Values of greater than 0.75 suggest relatively good model performance.



Cross Validation

Finally, the goal of small area prediction is to predict data into areas that have not been sampled. The most effective way to tell whether a model is reliable in this case is to perform croass-validation testing. Here, we randonly split the data into sections called ‘Folds’. In each iteration of the cross-validation, one of these folds is excluded from the main data (the testing data), then the remaining data are used to build the model (the training data). The resulting model is then used to predict the indicator of interest in both sets of data with similar accuracy.

Number of Folds

24

Show 10 entries Search:

	model.AUC	test.AUC
	0.57	0.52
	0.57	0.58
	0.57	0.59
	0.57	0.57
	0.57	0.56
	0.57	0.59
	0.57	0.55

Figure 18: Results of the cross validation. Similar AUC scores in the training and test data indicate the model is relatively stable.

Prediction

Now that we have assessed the model, we can use it to predict into the census data using the same predictor variables. The predictions tab does this, plotting the predictions at census scale, then survey region scale, then finally with direct estimates. In the tabular data section, the predictions are aggregated at both survey and census scales. A bootstrapped 95% confidence interval is also calculated for each region.

Results within survey regions

Show 10 entries

Search:

	REGNAME	mn	lwr	upr
1	central hill	0.63	0.60	0.66
2	central mountain	0.58	0.55	0.61
3	central terai	0.48	0.46	0.51
4	eastern hill	0.43	0.40	0.46
5	eastern mountain	0.42	0.39	0.45
6	eastern terai	0.49	0.46	0.52
7	far-western hill	0.43	0.40	0.45
8	far-western terai	0.57	0.54	0.60
9	mid-western hill	0.42	0.39	0.44
10	mid-western terai	0.52	0.49	0.55

Figure 19: Tabular results