Results — overview and interpretation

2023-09-22

Table of contents

1 Overview

The folder `var/` contains information on all variables:

- `variables.csv`: Description of the variables, including codes, categories, reference levels, question text, etc.;
- `cat.levels`: The levels for each categorical variable.

The folder `willingness/` contains the main results for the "willingness to adapt" (i.e. whether a respondent answered `Yes, within the next 5 years` or `Yes, in 6 to 10 years` for an adaptation action).

- `willingness/varsel/var.sel.pdf`: graphical overview of the variable selection results;
- `willingness/varsel/var.sel.csv`: quantitative results of the variable selection;
- `willingness/irt/prob.cf.pdf`: visualization of marginal effects and comparisons based on the item-response model;
- `willingness/irt/predictions.csv`: marginals corresponding to the visualizations in `prob.cf.pdf`;
- `willingness/irt/comparisons.csv`: comparisons and slopes corresponding to the visualizations in `prob.cf.pdf`.

Further simplification / aggregation is possible (explained in Section 4.3), and I have included simple examples for variables `C01` and `F15`. The corresponding files are:

- `willingness/irt/prob.cf.agg.pdf`: visualization of the aggregated marginal effects and comparisons from the item-response model;
- `willingness/irt/predictions.agg.csv`: aggregated marginals corresponding to the visualizations in `prob.cf.agg.pdf`;
- `willingness/irt/comparisons.csv`: aggregated comparisons corresponding to the visualizations in `prob.cf.agg.pdf`.

The file `metadata.pdf` contains detailed column descriptions for all tables in `.csv` format. Suggestions as to how to use the information in the tables to describe the results from the item-response model is discussed below (Section 4).

In case you're interested, I have included the same kind of results for what can be called "urgency to adapt" (i.e. whether a respondent answered `Yes, within the next 5 years`) – which can be found in the folder `urgency/`. I'm not sure whether they are useful for the paper we want to write, so I'll let you decide whether we should even use them. For starters, the variable selection results indicate that most of the variables deemed important for willingness are also relevant for urgency.

2 Variable selection

2.1 Variable importance in random forests

In the case of random forest models, we have used the permutation variable importance (Breiman, 2001) to rank explanatory variables. Permutation importance measures how much the prediction error of a random forest model increases if that specific variable is rendered meaningless. More specifically: The performance of a random forest can be expressed by how far, on average, the model predictions are off from the observed truth. For example, if we observed 1, and the model predicts 0.8, the error for this observation will be 0.2. Averaging all these errors for the random forest gives the *prediction error*. Now we can take a variable we are interested in and shuffle its values (also called *permuting* or *noising* a variable). The permuted variable does not carry any information, and we can feed it to the model to generate predictions. As a result of the permutation, the prediction error may change, and this change can be expressed as a percentage of the original prediction error. For example, if permuting a variable increases the prediction error from 0.2 to 0.3, it constitutes a 50% increase in prediction error. This percentage increase in prediction error is the classic *variable importance* measure of Breiman (2001). If a variable has a variable importance score of 10, it means that permuting the variable (i.e. removing its information content) results in a 10% increase in prediction error. Negative variable importance means that permuting the variable actually reduces the prediction error, that is it improves the model.

A problem with variable importance measures arises if two (or more) strongly correlated variables are present in the random forest. If variables are strongly correlated, they present partly duplicated information: once one of the variables is included in the model, the second adds little additional information. This has consequences for variable importance measures: We can imagine two variables that carry very important information, but that are also highly correlated. If one of the correlated variables is permuted, the predictions of the model are not affected much, because the same information is still present in the form of the other, correlated variable. This leads, for both variables, to a low importance score. Only once both variables are removed together does the prediction error increase substantially. Now if the goal is to select a small number of variables from all the variables in the random forest, this creates a problem: if there is a group of correlated variables, they have little chance of being selected even though the informa-

tion they carry may be more important than the information contained in the variables with very high importance scores.

They only way to work around this problem is finding a random forest model with a reduced number of variables, obtained by progressively including new variables into an "empty" model, or by excluding them from a full model. However there are no established thresholds at what point a smaller model is "close enough" in performance to the best model, or at what point adding additional variables merely leads to overfitting (i.e. fitting to random "noise" patterns in the data). It is difficult to find any example for this kind of variables selection in the literature (and especially not with the idea to feed the variables to a regression model later), since machine learning models most often work with the full set of variables (i.e. these models are usually aimed at predicting well, not at being parsimonious). All that to say we could end up in a world of pain doing step-wise selection with random forests.

2.2 Alternative approach: Bayesian model selection

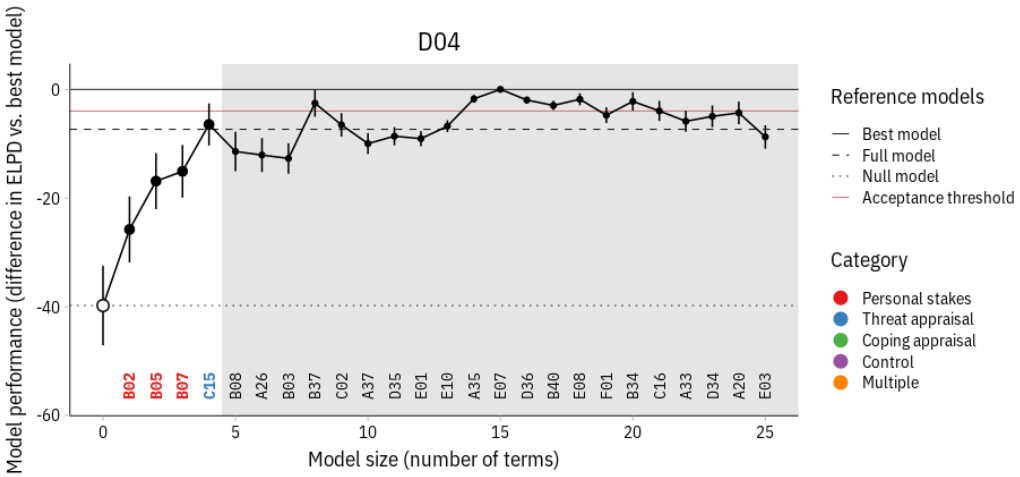
From a Bayesian point of view, selecting a subset of all explanatory variables actually falls within approaches for *model selection*, that is one wants to find – among many possible models – a model that strikes a compromise between predictive performance and sparsity (i.e. the number of explanatory variables). Rather than using random forests for variable selection, I suggest adopting a Bayesian variable selection method called *projection predictive feature selection*, presented by Piironen & Vehtari (2017a). This method makes use of Bayesian leave-one-out cross-validation (LOO-CV) to evaluate model performance (Vehtari, Gelman, & Gabry, 2017), which has been shown to be robust and currently constitutes the quasi-standard. This selection approach also works well with horseshoe priors (Carvalho, Polson, & Scott, 2009; Piironen & Vehtari, 2017b).

In our case, applying this method looks as follows:

- For each adaptation action, formulate a *full model* – in this case a model containing the effects of all possible explanatory variables. A regularized horseshoe prior is employed for this model to shrink small (read: relatively unimportant) effects towards zero.

- Once the full model is estimated, it is used to perform projection predictive feature selection. The procedure starts with an intercept-only *null model*. Then a search in the parameter space is conducted to suggest a ranking of the explanatory variables. According to this ranking, explanatory variables are then added one by one to create new submodels – so that each submodel has a larger size (i.e. number of variables) than the previous one. Then the performance of all submodels is evaluated, and subsequently compared against the *best model*. The best model is the one with the highest performance among all estimated models (i.e. the best model can be either the full model, or any of the smaller submodels).
- A decision rule is then used to select the smallest model with still acceptable performance.

As a measure to evaluate (and compare) model performance, we use the difference in ELPD. The ELPD (expected log pointwise predictive density) quantifies how well the model can be expected to predict a single data point it has not yet seen. For ELPD, higher is better (as opposed to AIC, where lower is better). Since we are comparing all models against a reference model, we use the difference in ELPD against the best model. This means the best model always has a difference in ELPD of 0, and all other models have negative values. This is what the output looks like:



The y-axis shows the difference in ELPD against the best model, and along the x-axis, the size of the model (i.e. the number of explanatory variables in the model) increases. The vertical lines are the standard errors for the estimates of the difference in ELPD. Along the x-axis there is additional information on the variable that is newly included in a model of that size. For example, at model size 3, we can see that the model contains the variables `B02` and `B05` (because they have been included earlier) as well as variable `B07`, which is newly included. Several (horizontal) reference lines are also shown: The performance differences for the null model (dotted), the full model (dashed), and the best model (solid, always at 0).

I settled on the following decision rule: We establish a threshold equivalent to 90% of the performance improvement that the best model provides over the null model. We then select the smallest model that is not more than one standard error below this threshold (i.e. whose performance is "indistinguishable" from this threshold). In other words, we are willing to sacrifice 10% of relative performance in order to obtain a substantially smaller model.

Visually, the variable selection is straightforward. In the figure, the performance threshold is indicated by a solid red line. Among all the submodels whose vertical black line (i.e. its standard error) either overlaps with the threshold or is fully above the threshold, we choose the smallest one. In this case it's the model of size 4, which contains the variables `B02`, `B05`, `B07`, `C15`.

For convenience, the selected variables for each adaptation action are coloured according to the variable category they belong to. In addition, if a variable has been selected for *any* of the ten adaptation actions, it is printed in bold. For example, the sixth-highest ranking variable, `A26`, is not selected for adaptation action `D04`, but it is selected for one (or several) of the other adaptation actions and thus shown in bold. The seventh-highest ranking variable, `B03`, is neither selected for `D04`, nor for any other adaptation action.

In terms of interpretation, the most important information is contained in the (colored) ranking of predictor variables that have been selected. This can also give some lines of inquiry for the combined item-response model later: variable `B02` appears to be important for adaptation `D04`, so it is probably worth taking a look at the quantitative effect that this variable has on `D04` in the item response model. The full ranking for the 25 most important variables can be found in the accompanying table.

Overall this Bayesian variable selection approach has several advantages:

- It works with correlated variables (which was the motivation behind using it);
- It is based on a well-established measure of model performance;
- The performance estimates have uncertainty intervals, which means we can define at which point two performance estimates cannot be distinguished;
- We can formulate how much predictive performance we are willing to sacrifice to obtain a smaller model, and thus a smaller subset of variables;
- As a result of the points above, different numbers of variables can be selected for different adaptation actions. This is more realistic than our previous rule (which was "three variables per adaptation action"), as some actions may be sufficiently explained by just one or two variables, while others are more complex and require a handful of variables to be explained.
- And finally, we stay within the framework of Bayesian regression (which is also used for the item-response framework), instead of evoking (and having to justify) a completely different machine learning method, such as random forests.

The drawback: These things take quite some computational time (more than 500 CPU hours in this case) and require additional model checking. But with access to a high performance computing cluster (thank you, *Calcul Québec*) this is not an issue.

3 Terms to describe outcomes, effects, and uncertainty

When fitting Bayesian models to explain binary outcomes, one is confronted with different types of probabilities. To avoid talking about hardly intelligible things – such as talking about the probability that a probability is larger than another probability – I propose we follow a few conventions. Since this problem is not new in Bayesian inference, there are a few terms that can be useful to work with probabilities of different kinds:

- *Willingness to adapt*: We could use this term to talk about the probability that a respondent chooses either "Yes, in 6 to 10 years" or "Yes, within the next 5 years" for an adaptation action.
- *Urgency to adapt*: This could be a term to refer to the probability that a respondent chooses "Yes, within the next 5 years" for an adaptation action.
- *Certainty*: expresses the probability of an hypothesis and is usually called *degree of belief* by Bayesian (although that term is a mouthful). That is, if we compare two levels of a variable, the *certainty that the willingness increases* quantifies how strongly

we believe that one level increases the willingness, when compared to the other level. The certainty (or degree of belief) is thus a kind of probability that is entirely different from the probability used to model an outcome (e.g. “willingness”).

Below (Section 4) you will find some examples for how to use these terms when presenting and discussing results of the item-response models.

To make writing (and reading) the manuscript easier, we could also use pre-defined levels of certainty (which would also have to be stated in the methods section):

- *Low certainty*: a certainty between 50% and 75%;
- *moderate certainty*: a certainty between 75% and 90%;
- *high certainty*: a certainty above 90%.

Certainty levels below 50% are not needed as a certainty of 45% that an effect is positive means that there is a 55% certainty that the effect is *negative* (at least in our case).

Of course, this is just an idea that we can discuss further.

4 Item-response models

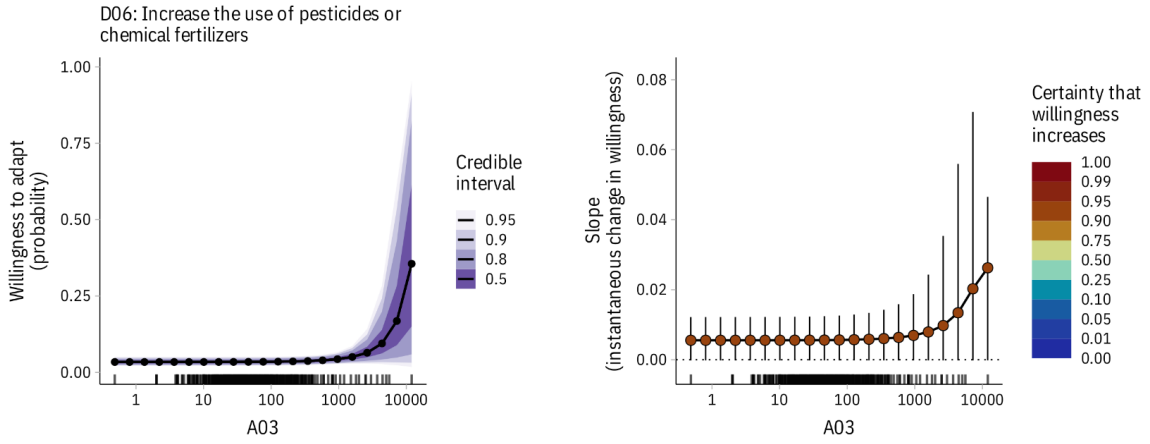
All variables that have been selected for at least one adaption action are included in the item-response model. Each adaptation action is seen as an *item* and the model aims to explain the response given to this item as a function of the selected variables (which are called *person covariates*), the identity of the item, and the identity of the respondent. We use a 2-parameter logistic model with differential item functioning (Bürkner, 2021) to account for the possibility that a given person covariate can affect each item in a different way.

To interpret the effect of a single covariate, we use *counterfactual marginal effects*. They are calculated by taking the entire data set, and then setting the covariate in question to a specific value, while leaving all other variables at their original value. Model predictions for all these observations are then averaged in order to calculate a marginal value. This can be repeated for a range of values of the covariate. For example, we can take the entire data set, set the value of `F14` (self-identification as indigenous) to `Yes` throughout

(i.e. for all observations), generate model predictions for this data, and average the predictions to obtain a marginal value for `F14 = Yes`. The we can repeat this for `F14 = No`. And since we are using a Bayesian model, we obtain the entire posterior distribution for these marginals. Why the trouble? Counterfactual marginal effects are not conditional on reference values for the other covariates (which have to be chosen, essentially arbitrarily), they are only conditional on the entire sample (as is everything else we estimate). This means that, for each item, the probabilities (e.g. for the willingness to adapt) are centred around the average probability for this item. There is a blog post by Andrew Heiss with some neat infographics for different types of marginal effects.

4.1 Interpretation of continuous person covariates

If the person covariate is a continuous variable, we can visualize its (counterfactual marginal) effect on a specific item with a pair of plots, such as these for the person covariate `A03` ("What is the surface of your woodlot? [Hectares]"):



The plot shows shows the willingness to adapt for adaptation action `D06` along the y-axis (this is a probability and must be between 0 and 1), for a range of specific values of the person covariate along the x-axis. Each point indicates a value for which predictions have been produced (I settled on 21 for each continuous variable, which already takes quite a lot of computational time). Shown are the median willingness to adapt (black point and line), as well as several equal-tailed uncertainty intervals (shades of purple). The "rug" just above the x-axis shows the values for `A03` that occur in our sample. We can

see that the willingness to increase the use of pesticides or chemical fertilizers increases with the total surface of woodlots – albeit with considerable uncertainty. To know how certain we can be about the increase in willingness, we can look at the plot on the right-hand side. Here the slope (i.e. instantaneous change; mathematically this is the first derivative of willingness) is shown on the y-axis. The slope is calculated at the same values of the covariate for which willingness has been evaluated. If the slope is positive willingness increases. The vertical black lines around the points indicate a 90% credibility interval. If this credibility interval is fully above 0, we can be at least 95% certain that willingness increases at this point. If the vertical line is fully below zero, we are less than 5% certain that willingness increases (which means, in turn, that we are more than 95% certain that willingness decreases). Colour give a more fine-grained idea of certainty that willingness increases. When fitting a linear effect of a continuous variable, the certainty should be exactly the same at every point (which is the case here): The certainty that willingness to adapt increases with woodlot size is close to 95% for all evaluated points.

For detailed statements, we can look up the information for variable `A03` and adaptation action `D06` in the corresponding tables (that is where `var.code` is equal to `A03`, and where `adapt.code` is equal to `D06`). For the left plot, the corresponding table is `predictions.csv` (only some columns are shown here):

var.level	prob.median	prob.q5	prob.q95
0.486	0.034	0.023	0.048
0.805	0.034	0.023	0.048
1.335	0.034	0.023	0.048
2.214	0.034	0.023	0.048
3.672	0.034	0.023	0.048
6.089	0.034	0.023	0.048
10.096	0.034	0.023	0.048
16.742	0.034	0.023	0.048
27.762	0.034	0.023	0.048
46.036	0.034	0.023	0.048
76.339	0.035	0.024	0.048
126.587	0.035	0.024	0.049
209.910	0.036	0.025	0.050
348.077	0.037	0.026	0.051
577.191	0.039	0.027	0.054

var.level	prob.median	prob.q5	prob.q95
957.113	0.043	0.029	0.062
1587.109	0.050	0.031	0.080
2631.785	0.064	0.033	0.124
4364.093	0.095	0.034	0.242
7236.651	0.168	0.033	0.535
12000.000	0.356	0.031	0.914

A full description of the columns is provided in `metadata.pdf`. We can make a statement with a 90% credible interval by taking information from the columns `prob.median`, `prob.q5`, `prob.q95`. I will assume here that we have already stated in the methods section that we will provide medians as point estimates, and 90% equal-tailed quantile intervals as credible intervals (90% uncertainty intervals are useful because they are equivalent to being 95% certain about a one-sided hypothesis). We can say, for example:

> For respondents that own woodlots with a total size of 10 hectares, the willingness to increase the use of pesticides or chemical fertilizers is 0.034 (CI: 0.023 – 0.048).

Of course, this information by itself may not be very interesting (other than expressing the notion that the willingness is low). How does willingness change with increasing woodlot size? This information was contained in the right-hand side plot and the corresponding values are provided in `comparisons.csv`. Here are some of the columns for `adapt.code` equal to `D06`, and `var.code` equal to `A03`:

var.level.cont	prob.slope.median	prob.slope.ci.l	prob.slope.ci.u	cert.pos
0.486	0.006	0	0.012	0.943
0.805	0.006	0	0.012	0.943
1.335	0.006	0	0.012	0.943
2.214	0.006	0	0.012	0.943
3.672	0.006	0	0.012	0.943
6.089	0.006	0	0.012	0.943
10.096	0.006	0	0.012	0.943
16.742	0.006	0	0.012	0.943
27.762	0.006	0	0.012	0.943

var.level.cont	prob.slope.median	prob.slope.ci.l	prob.slope.ci.u	cert.pos
46.036	0.006	0	0.012	0.943
76.339	0.006	0	0.013	0.943
126.587	0.006	0	0.013	0.943
209.910	0.006	0	0.013	0.943
348.077	0.006	0	0.014	0.943
577.191	0.006	0	0.016	0.943
957.113	0.007	0	0.019	0.943
1587.109	0.008	0	0.024	0.943
2631.785	0.010	0	0.035	0.943
4364.093	0.013	0	0.056	0.943
7236.651	0.020	0	0.071	0.943
12000.000	0.026	0	0.047	0.943

When talking about the effect, we can use the columns `var.level.cont`, `prob.slope.median`, `prob.slope.ci.l` and `prob.slope.ci.u` to make a statement about the slope (i.e. rate of change), with a corresponding 90% uncertainty interval:

> For respondents that own woodlots with a total size of 10 hectares, the willingness to increase the use of pesticides or chemical fertilizers grows by 0.006 per hectare (CI: 0 – 0.012).

If we only want to talk about the certainty that an increase exists, we can use the information in `cert.pos`:

> We are 94% certain that the willingness to increase the use of pesticides or chemical fertilizers grows with total woodlot size.

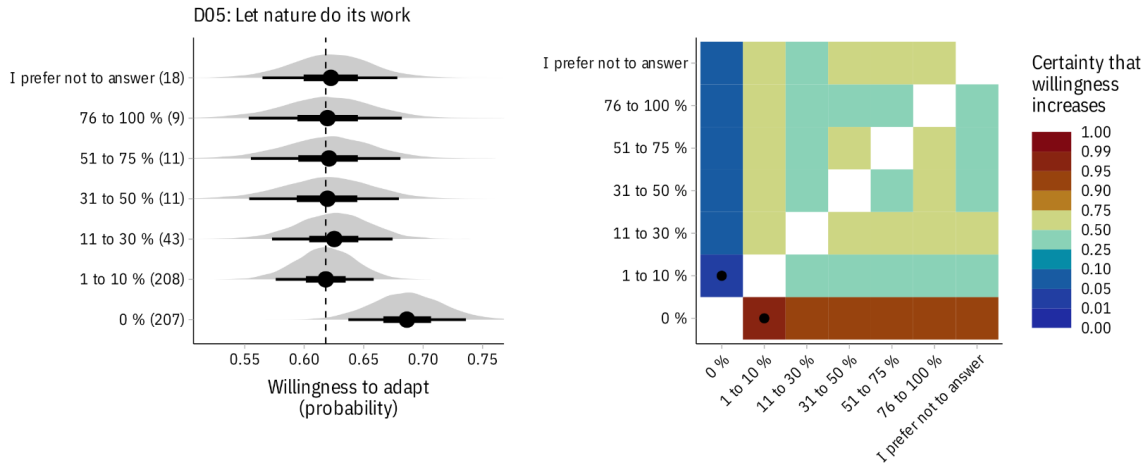
If we use the predefined certainty levels, we could also say:

> Across all respondents, it is highly certain (94%) that larger woodlots are associated with a greater willingness to increase the use of pesticides and chemical fertilizers.

In case you are wondering why the certainty that the effect is positive is exactly the same everywhere (even though the value of the slope changes): this is because the variable `A03` was included as a linear effect at the scale of the linear predictor. We could include it as a non-linear effect but doing so does not improve the model performance in this case.

4.2 Interpretation of categorical person covariates

If the person covariate is a categorical variable, we can use the following plots to visualize its counterfactual marginal effect on a specific item:



The left-hand side plot shows the marginal distribution for each level of the person covariate `F15` ("On average, how much of your annual family income comes from your woodlots?"), with covariate levels along the y-axis, and willingness to adapt on the x-axis. The point estimates are medians, the thick lines are 50% credible intervals, and the thin lines 90% credible intervals. The right-hand side plot shows the certainty that one level increases the willingness, compared to another level.

The actual values shown in the left plot are again provided in `predictions.csv`.

var.level	prob.median	prob.q5	prob.q95
0 %	0.686	0.637	0.736
1 to 10 %	0.618	0.576	0.658
11 to 30 %	0.625	0.573	0.674
31 to 50 %	0.620	0.554	0.680
51 to 75 %	0.621	0.555	0.681
76 to 100 %	0.620	0.553	0.682
I prefer not to answer	0.622	0.565	0.678

As before, we can use this information to make statements about the expected willingness to adapt, this time for each level of the categorical variable. For example:

> Respondents that do not derive any income from their woodlots show a willingness of 0.69 (CI: 0.64 – 0.74) to "let nature do its work".

Values for comparisons between levels are provided in `comparisons.csv`.

var.level.1	var.level.2	prob.diff.median	prob.diff.ci.l	prob.diff.ci.u	cert.pos
0 %	1 to 10 %	0.067	0.012	0.130	0.982
0 %	11 to 30 %	0.060	-0.001	0.131	0.947
0 %	31 to 50 %	0.065	-0.004	0.148	0.938
0 %	51 to 75 %	0.065	-0.004	0.147	0.938
0 %	76 to 100 %	0.065	-0.006	0.147	0.931
0 %	I prefer not to answer	0.063	-0.005	0.137	0.936

Only some of the pairwise comparisons are shown here, the actual table contains more. We can make statements about the effect by using information in the columns `var.level.1`, `var.level.2`, `prob.diff.median`, `prob.diff.ci.l` and `prob.diff.ci.u`. For example:

> Respondents that do not derive any income from their woodlots show a willingness to "let nature do its work" that is 0.067 (CI: 0.012 – 0.130) higher than respondents that obtain between 1% and 10% of their income from their woodlots.

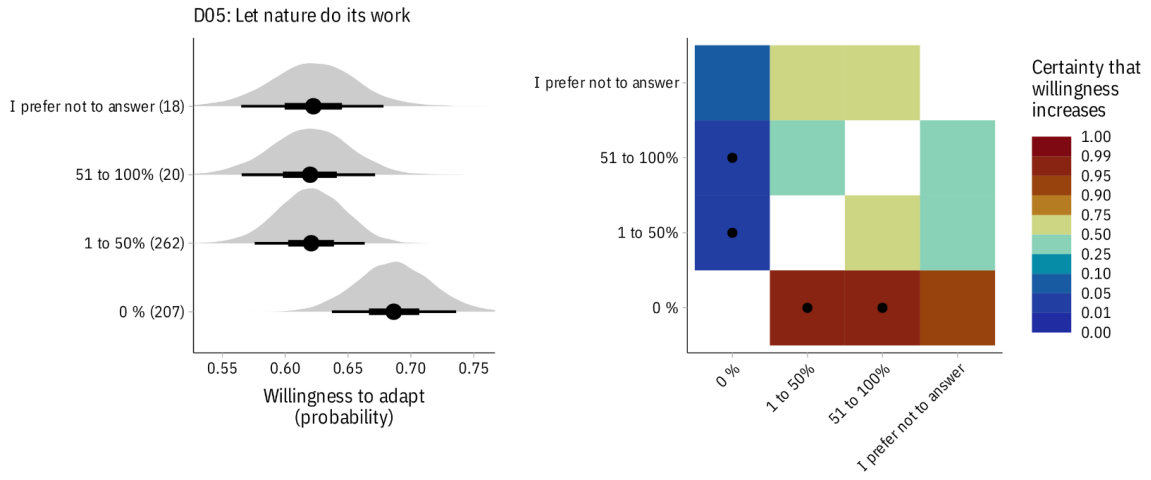
We can see that for people with no income from their woodlots, the certainty that the willingness is higher is actually always above 0.93 (lowest is 0.931, highest 0.982), no matter what other level it is compared against. We can thus also make the following statement:

> Within our sample, we can assert with high certainty (> 93%) that not obtaining any income from woodlots is associated with an increased willingness to "let nature do its work".

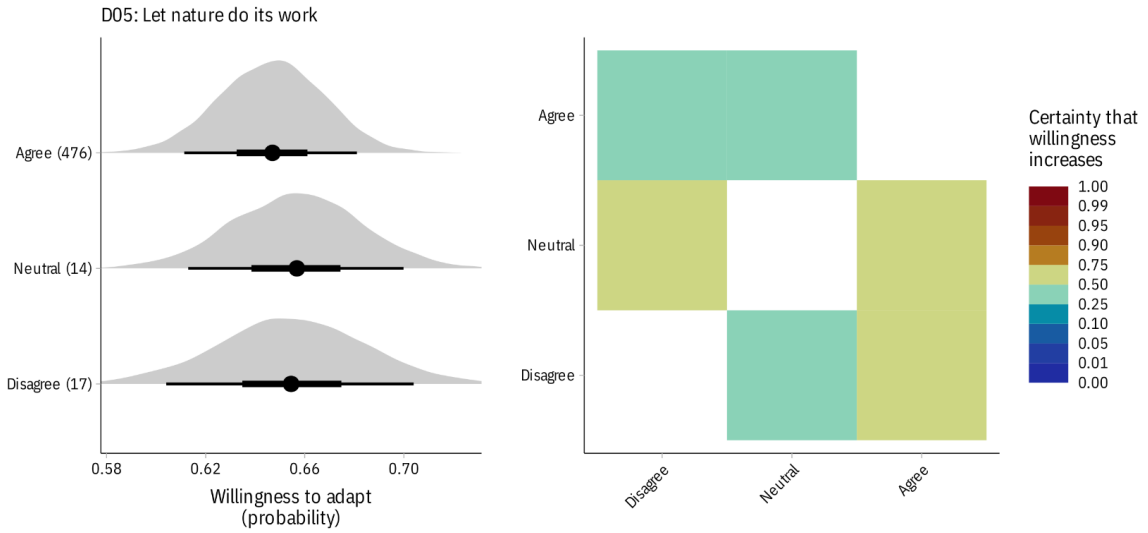
There are definitely many ways to phrase this, and the different types of information (and statements) can be mixed according to the point that one wants to make.

4.3 Simplification: aggregated estimates

Bayesian models are very flexible when it comes to calculating outputs. If specific estimates are required (e.g. comparing one level of a factor against the average of *all* other levels combined; or calculating an estimate for a woodlot size of exactly 1000 hectares) these can be easily obtained. Some of the items in the questionnaire are quite detailed, and we may be interested in aggregating some of the options (and we may not be all that interested in recoding the variables at the beginning and fit new models entirely; at least I have little interest in such an endeavour). For example, for variable `F15` it may be useful to compare the group of respondents that have no income from their woodlots to the group that obtains less than half of their income from their woodlots, and to the group that obtains more than half of their income from woodlots. Again, this can be visualized (as above, for adaptation item `D05`):



Another example would be recoding a 7-degree Likert scale to a 3-degree Likert scale. For person covariate `C01` ("The climate is currently changing all over the world") and adaptation item `D05` ("Let nature do its work"), we obtain the following simplified marginals and comparisons:



In this case the conclusions are the same as those from the full Likert scale: agreeing or disagreeing that the climate is currently changing has practically no effect on the willingness to let "nature do its work" (which is actually a good example that the absence of an effect can also be interesting).

I have calculated aggregated predictions and comparisons for `F15` and `C01` merely as examples. If you would like to do this for any of the variables, just tell me how you would like to aggregate them and I update the files (producing these results is only a matter of minutes since the scripts are already in place).

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*, 1–54. https://doi.org/10.18637/jss.v100.i05

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (pp. 73–80). PMLR. Retrieved from https://proceedings.mlr.press/v5/carvalho09a.html

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b, July 6). Sparsity information and regularization in the horseshoe and other shrinkage priors. https://doi.org/10.1214/17-EJS1337SI

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4