C01-1: Simple predictive models

The following example uses data from a field survey of areas in the Fogo natural park in 2007 by K. Mauer. For more information, please refer to this report.

Regarding libraries, the following packages are necessary.

[redacted]

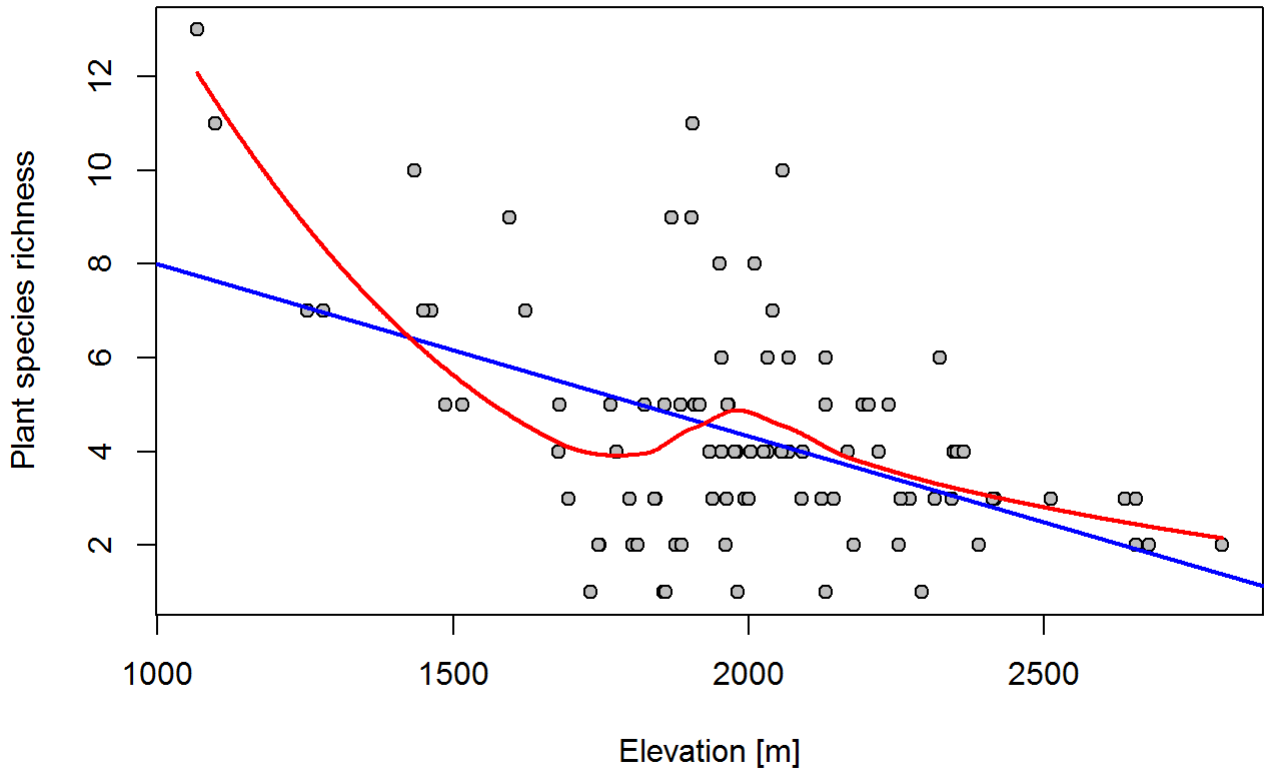
Simple visualization and models

In many cases, species richness changes along elevational gradients. Hence, a first analysis could check the relationship between richness and elevation.

The following example first creates a scatter plot and then computes a linear and a loess (i.e. local polynomial regression) model. The predicted linear and polynomial functions are added to the scatter plot in blue and red color.

Since the loess model is a polynomial function, individual y-axis values must be computed in a sufficient resolution in order to plot a “continuous” line. Therefore, the model is used to predict the y-axis values (i.e. species richness) for any full meter between the minimum and maximum elevation displayed in the scatter plot.

[redacted]



Advanced model selection for explaining species richness

A general problem with multiple variable models is overfitting. For example, the R-squared value will get larger and larger the more explanatory variables are included in the model equation (or at least R-squared will not decrease). If one chooses the model with the largest R-squared (i.e. with the most variables), the model might actually explain very much of the particular data sample used to train the model but the explanatory value might very likely be next to nothing if the model is applied to another sample.

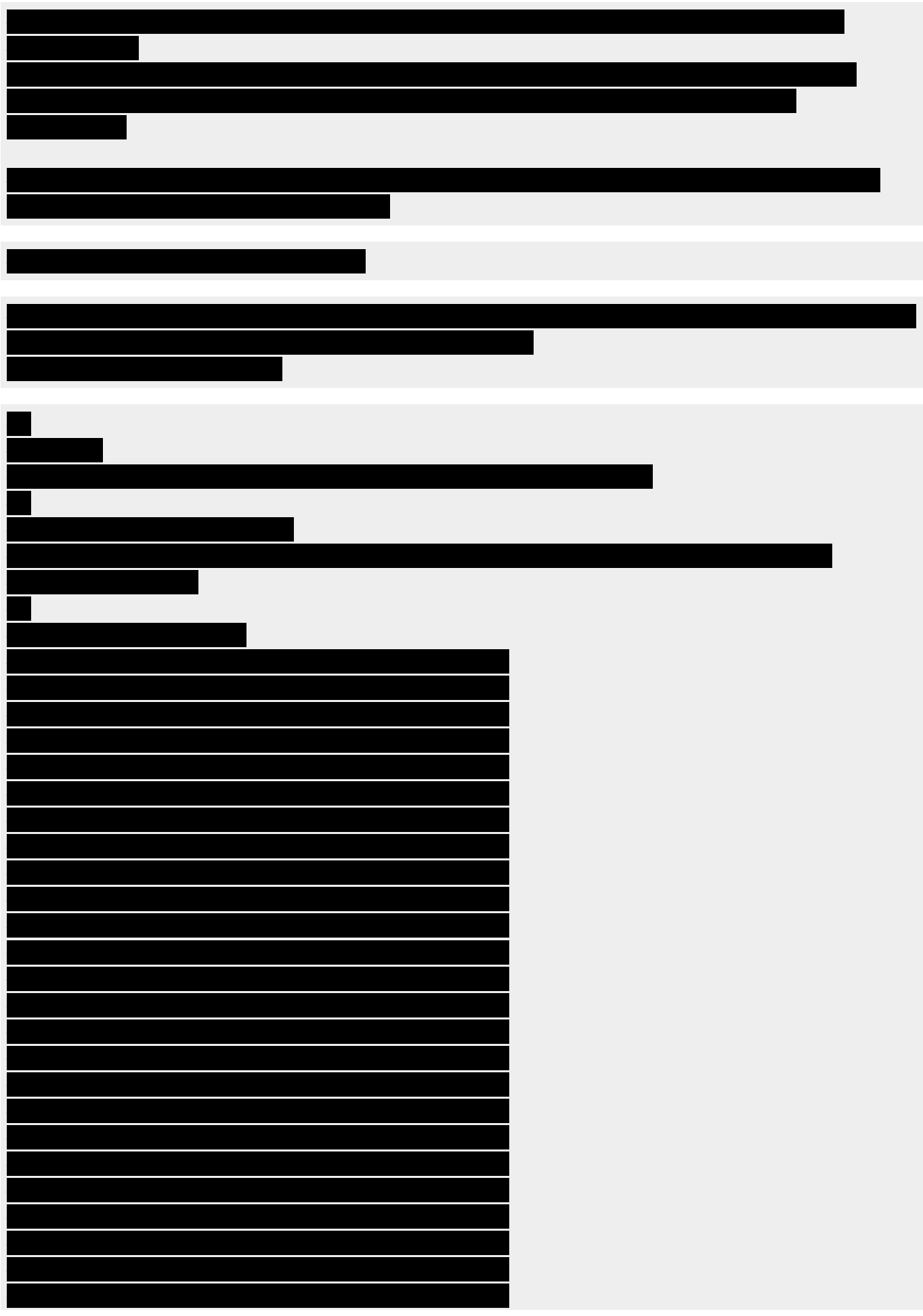
Hence, the best model is not the one which explains most but which explains quite a lot with only a few variables.

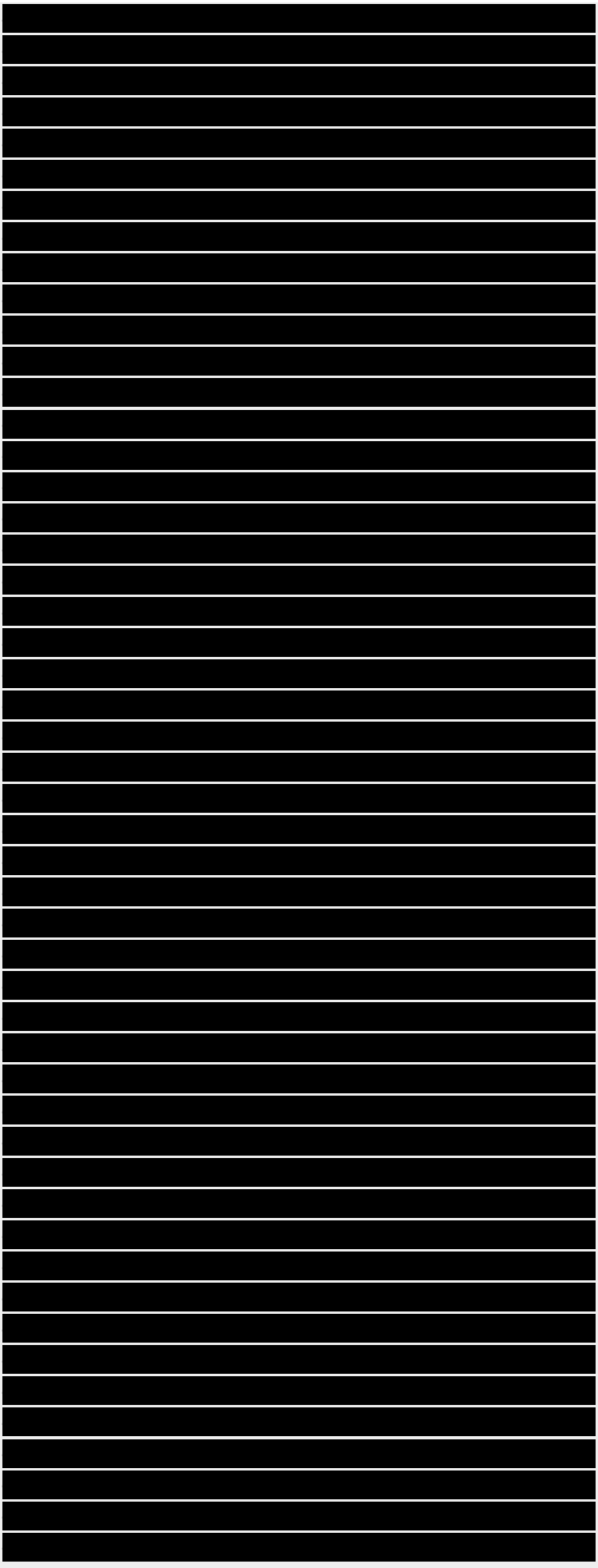
While there are quite many approaches which can be used for training the best model (e.g. cross-validation or boot-strapping), the example below illustrates a model averaging approach where 256 unique linear models are built which use a different (sub-)set of all explanatory variables available in the dataset.

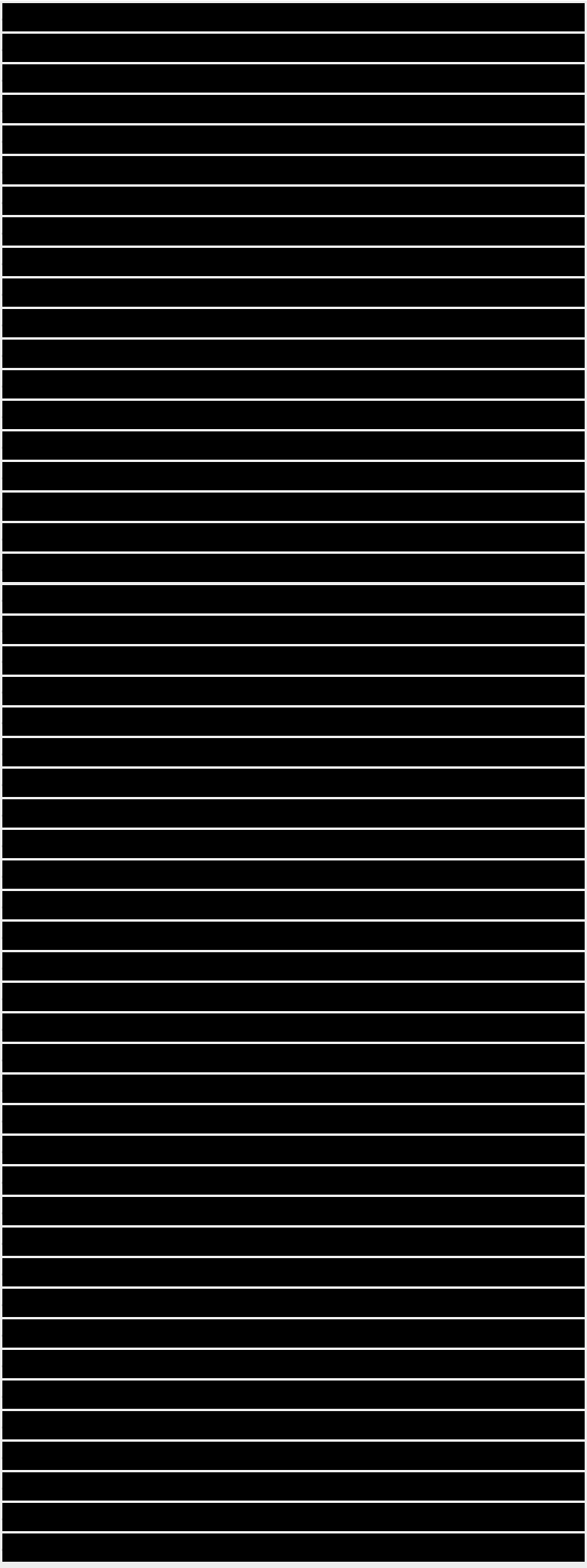
Applying such an automated model selection is quite easy. We will use the [redacted] function of the [redacted] package and just pass our linear model (which has been built using all available explanatory variables) to the function.

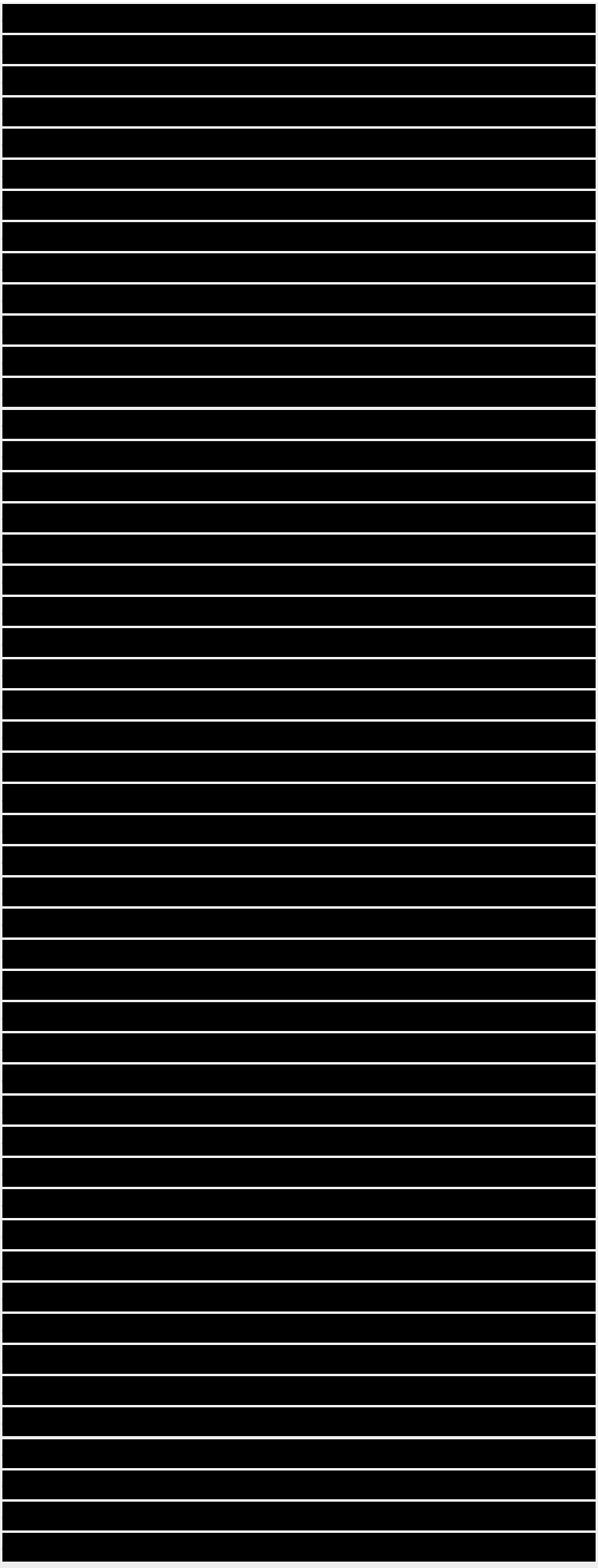
Since the results from [redacted] are quite confusing in the first place, we apply a model averaging (function [redacted]) afterwards and just look at a summarized output. As you can see, aside from returning information on the best model using e.g. the AICc value (the best model has the smallest), it

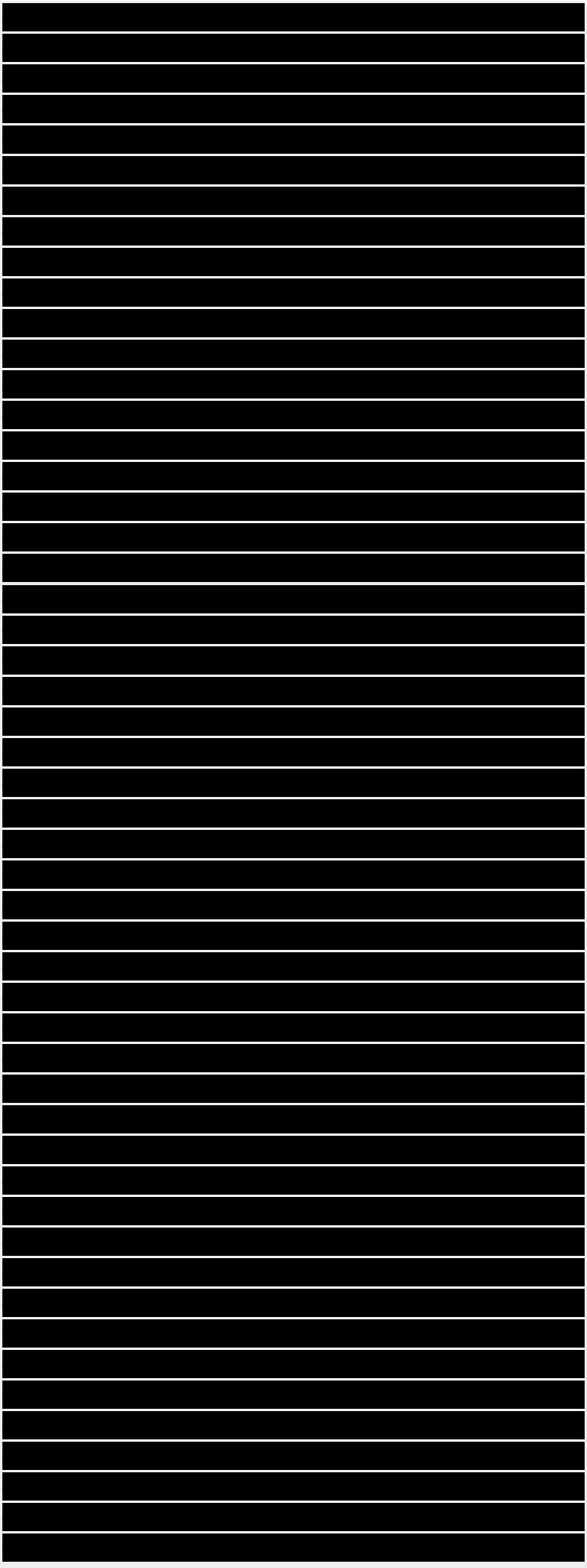
also returns significance information for each of the explanatory variables.

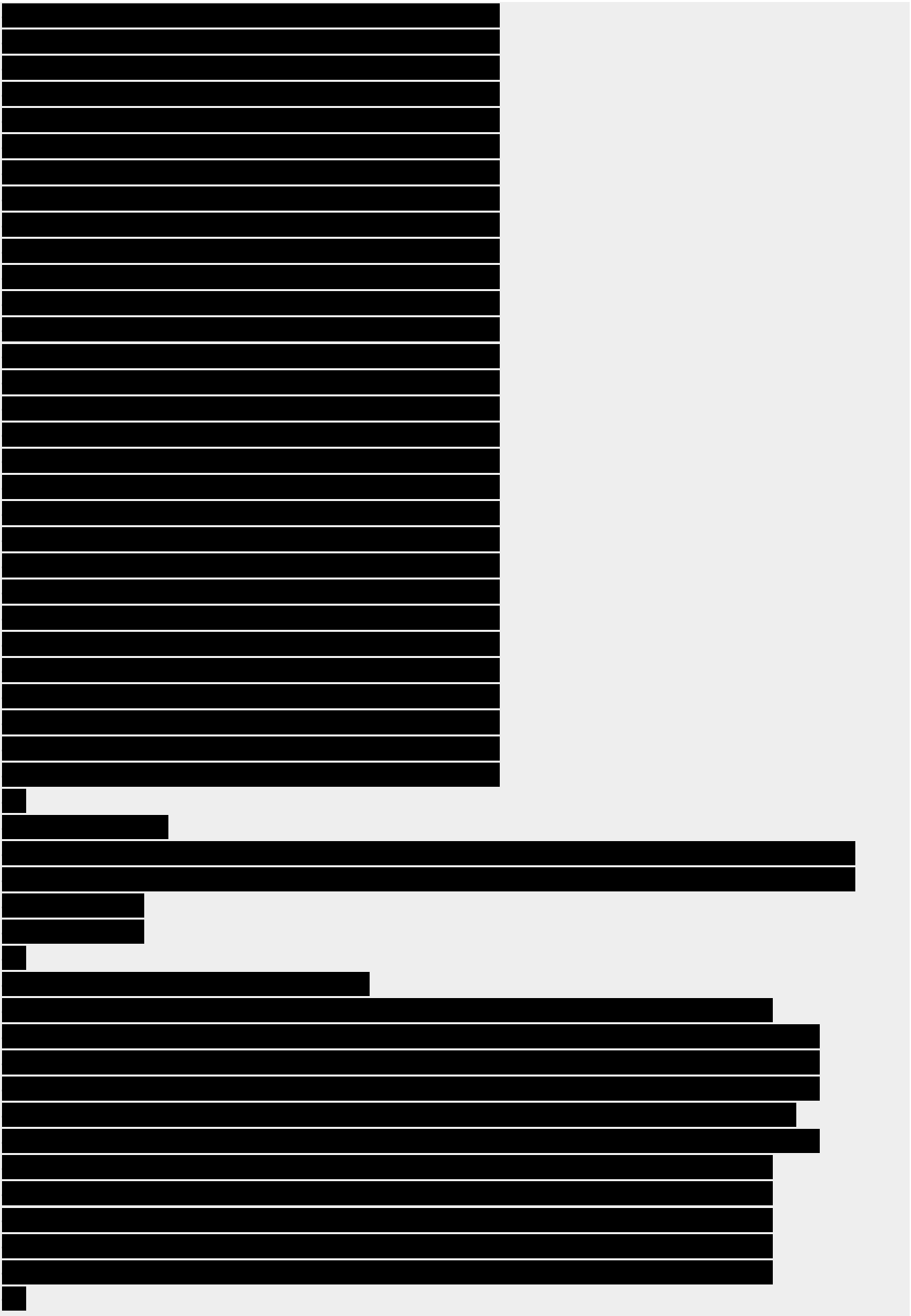


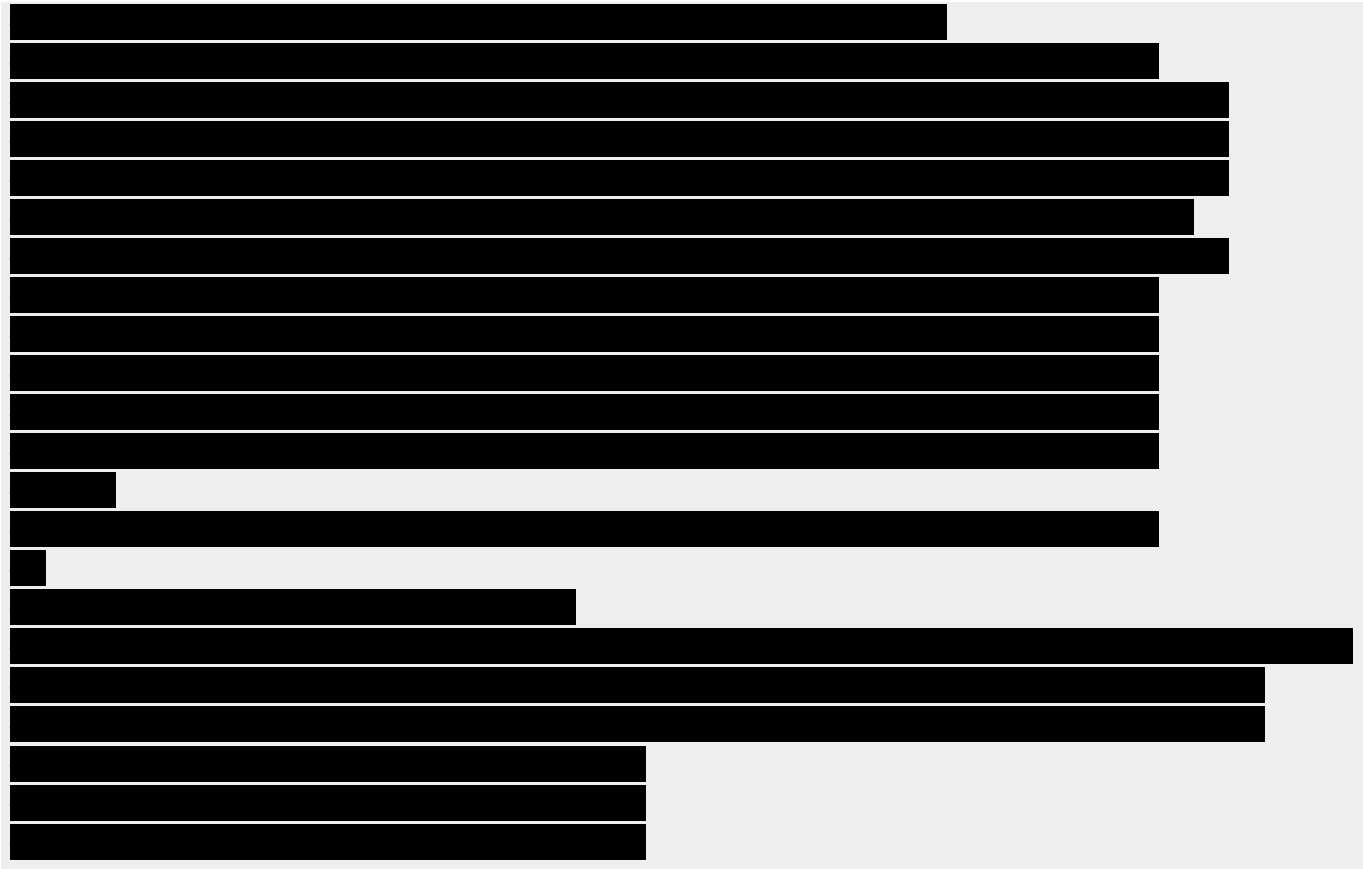












In the case of the linear model approach above, the best model is the one which just uses ALT_GPS_M, CAT_USO and DECL_GR as explanatory variables.