**Chapter 1 results summary**

**Macroeconomic predictors of forest cover change**

This first part of the analysis aimed to model the relationships between national-level macroeconomics and forest loss. I ended up with 17 variables in the final set, and these were further divided into three sets: 1) macroeconomics (e.g. GDP, foreign investment), 2) commodity prices (e.g. regional market price for rice, sugar), and 3) producer prices (e.g. farmgate prices for rice, cassava).

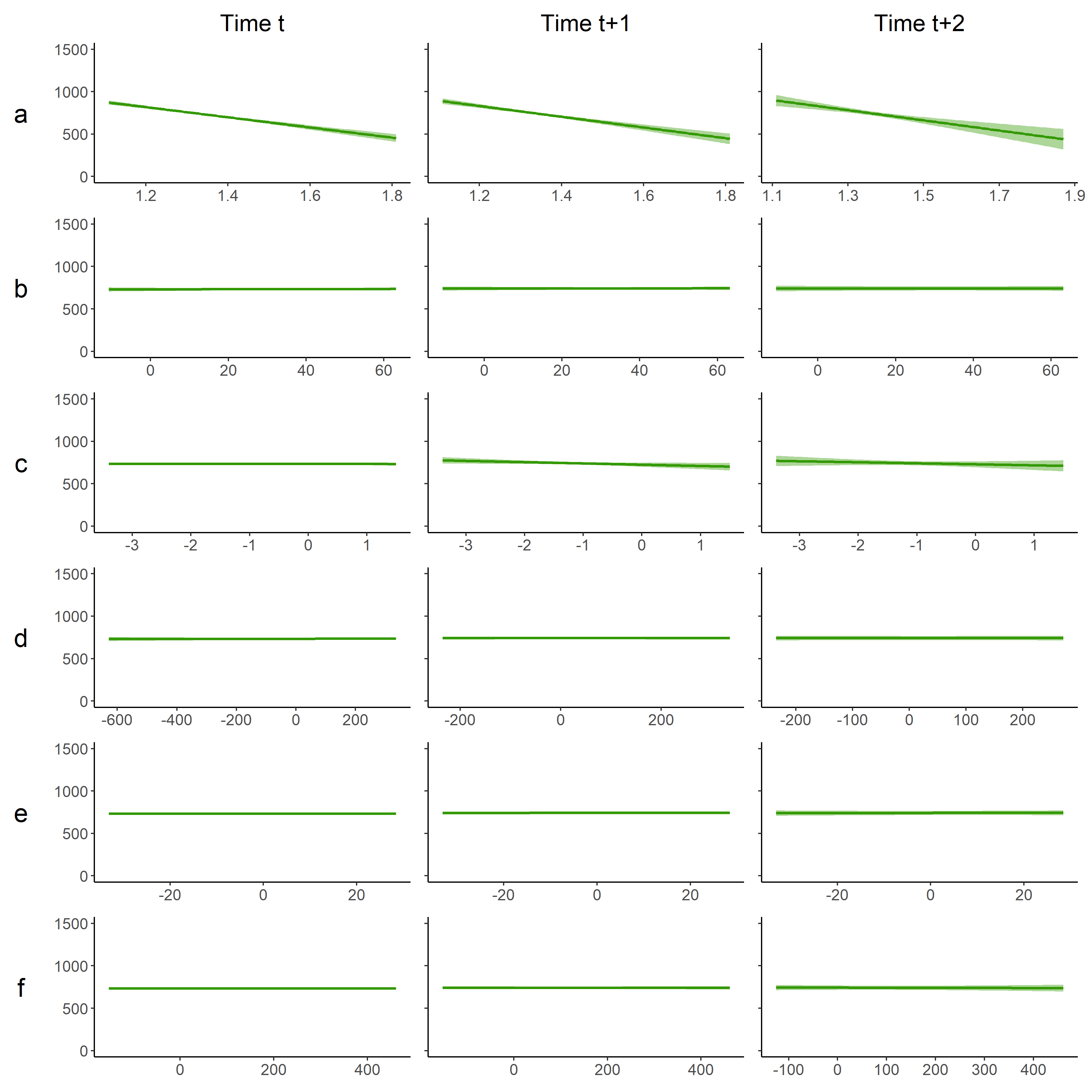
For each set, I ran models with all combinations of the variables, selected all models with a dAIC < 6, and then model averaged across the final model set. To account for time I modelled forest cover against time (year), and then used the residuals of that model as a covariate in all other models. I also included a forest remaining variable in all models, as this has been shown to be an important predictor in several studies. I further tested whether there were time lags by modelling forest loss in the subsequent year and subsequent two years (i.e. do changes in the predictors take one or two years to affect forest loss?).

The results were remarkably uninteresting. There was a small negative effect of population density on forest loss – in other words as population density increases, the amount of forest lost decreases, which is counterintuitive. This effect was there at times t, t+1, and t+2 (Fig. 1).

For the other two variable sets, there were no significant effects at all (Figs. 2 & 3).

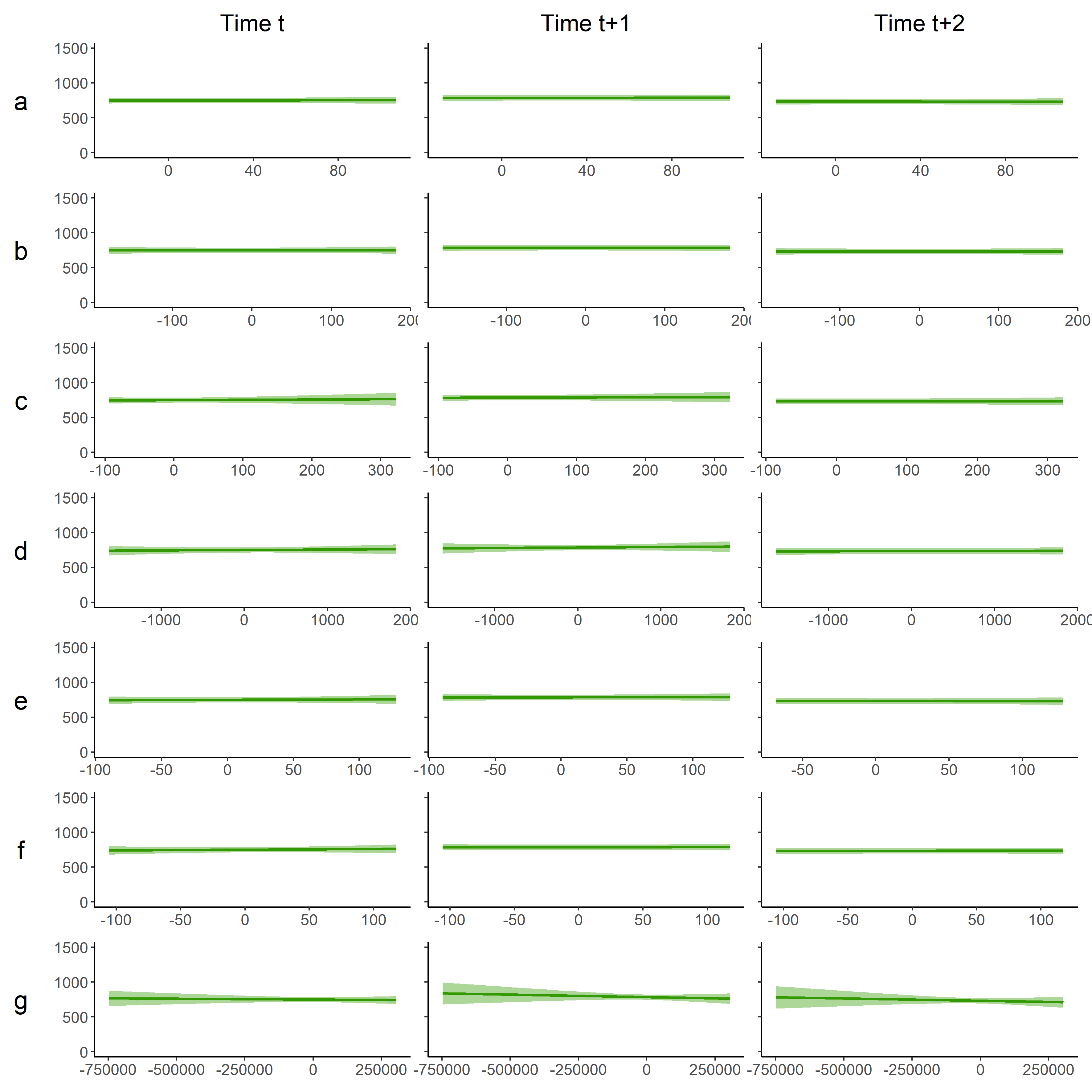
**Models with new ELC allocation as the response**

Following Kate’s advice, instead of using actual forest cover loss as the response, I looked at the count of new ELCs allocated in a given year as the response. Less direct than forest loss, but ELCs are known to be a major driver of forest loss, and so this is a good proxy. The issue with forest loss as the response is that the lag time is probably longer than 2 years, but I do not have enough data in the time series to look at longer time lags.

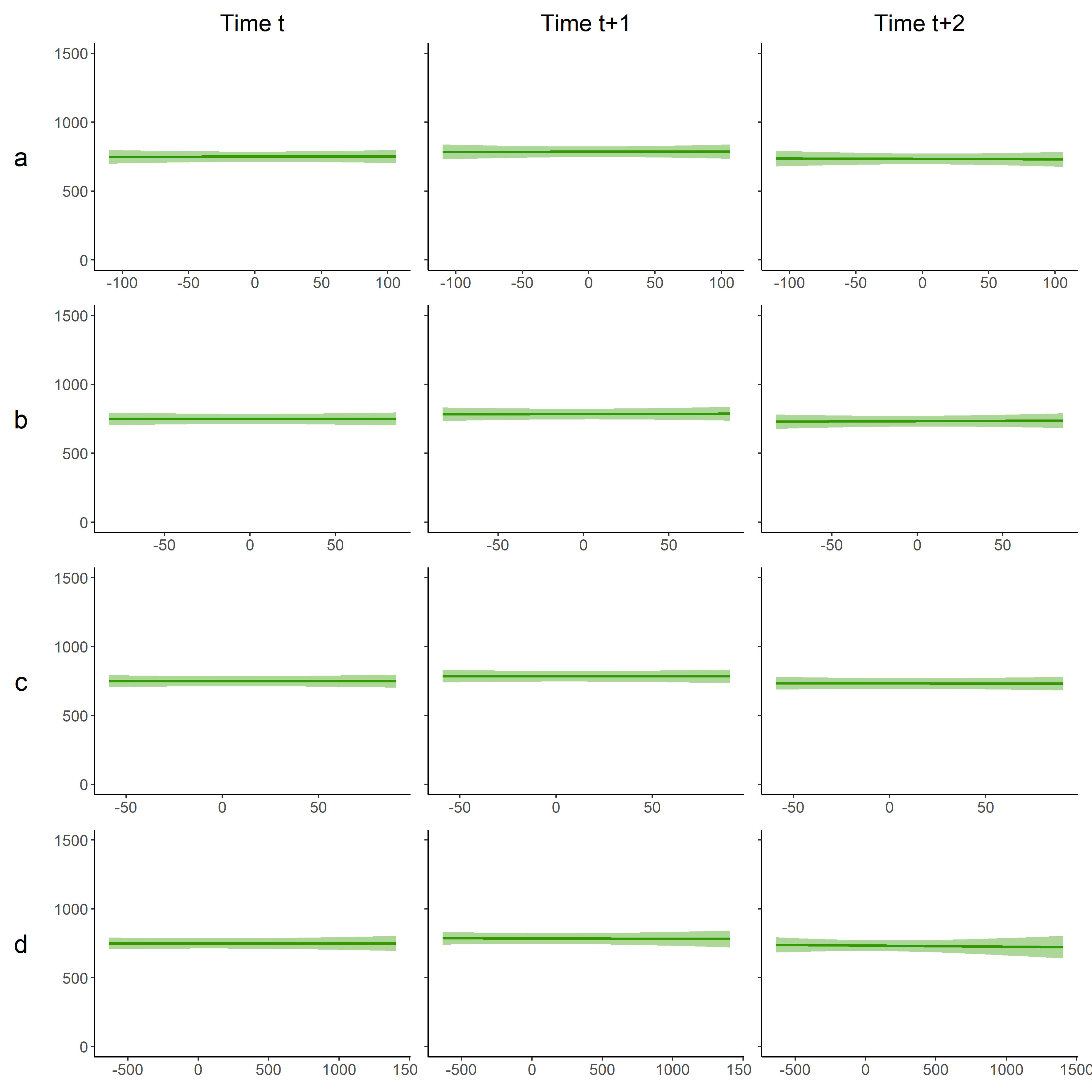


**Figure 1. Predicted relationship between forest loss and macroeconomic variables. All y-axes are the amount of forest lost in hectares. Row a: population density (individuals/km2), row b: Gross Domestic Product (GDP), row c: agricultural sectors contribution (%) to GDP, row d: development flows to the agricultural sector (USD millions), row e: development flows to the environment sector (USD millions), row f: Foreign Direct Investment (USD millions).The left column of plots are the effects on forest cover at time t (i.e. the variable values and forest loss values from the same year), the middle column of plots are the effects at time t+1 (i.e. the effects on forest loss in the subsequent year), and the right column of plots are the effects at time t+2 (i.e. the effects on forest loss two years after the variable values).**

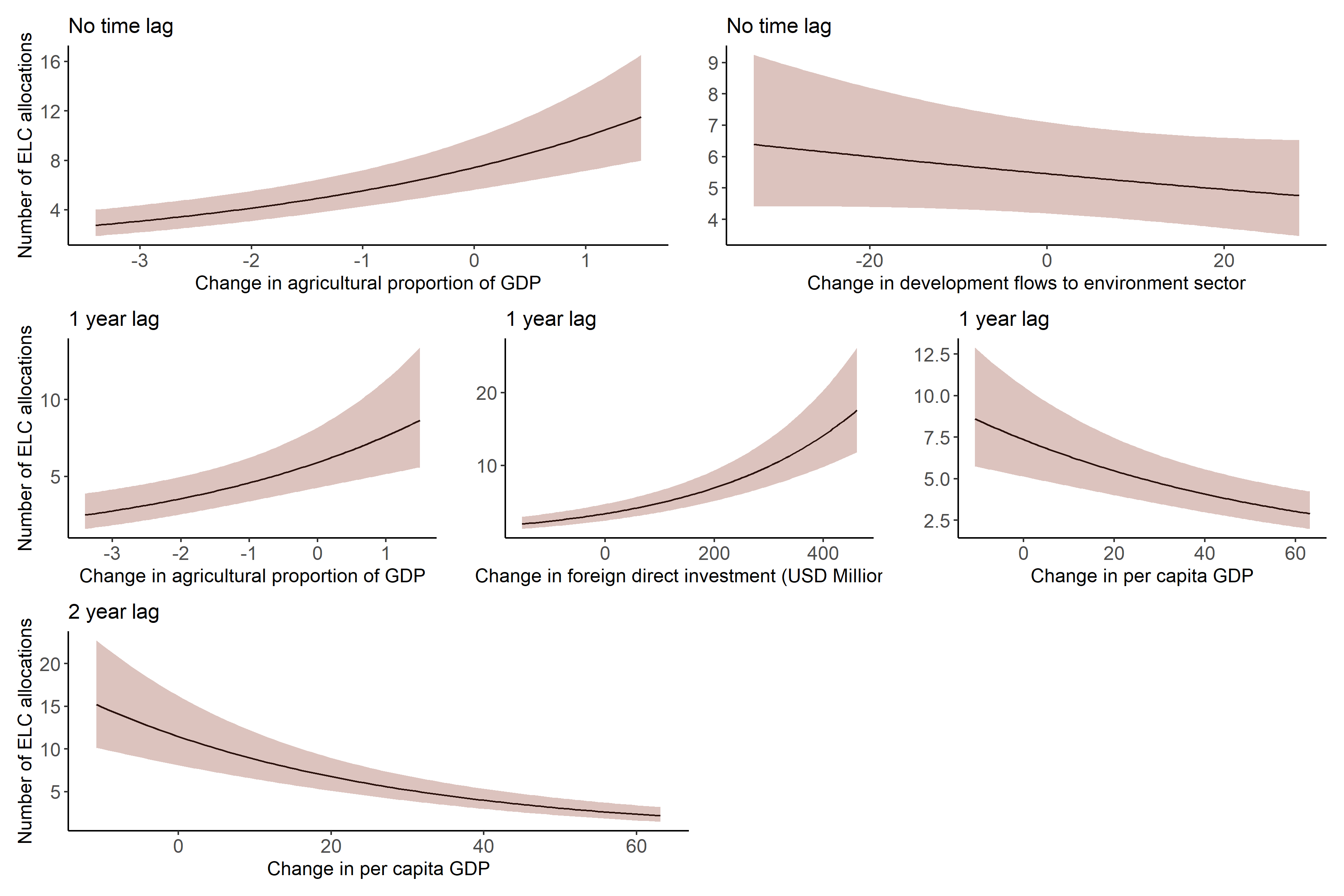
**REMOVE POP\_DEN – IT’S A CONTROL VARIABLE**



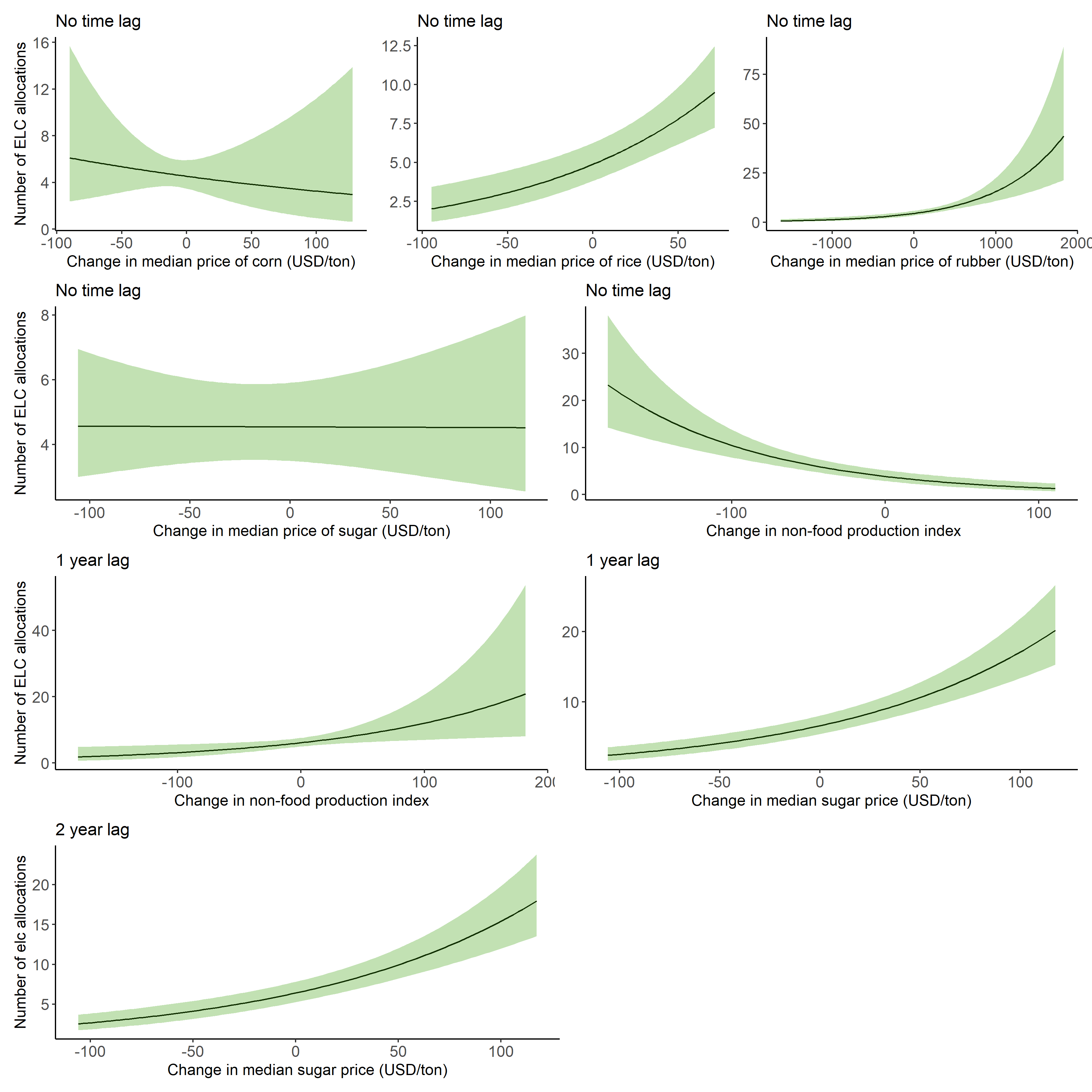
**Figure 2.** **Predicted relationship between forest loss and commodity variables. All y-axes are the amount of forest lost in hectares. Row a: Crop Production Index, row b: Non-food Production Index, row c: median annual market price for rice (USD/t), row d: median annual market price for rubber (USD/t), row e: median annual market price for corn (USD/t), row f: median annual market price for sugar (USD/t), row g: total production from forestry (m3). The left column of plots are the effects on forest cover at time t (i.e. the variable values and forest loss values from the same year), the middle column of plots are the effects at time t+1 (i.e. the effects on forest loss in the subsequent year), and the right column of plots are the effects at time t+2 (i.e. the effects on forest loss two years after the variable values).**



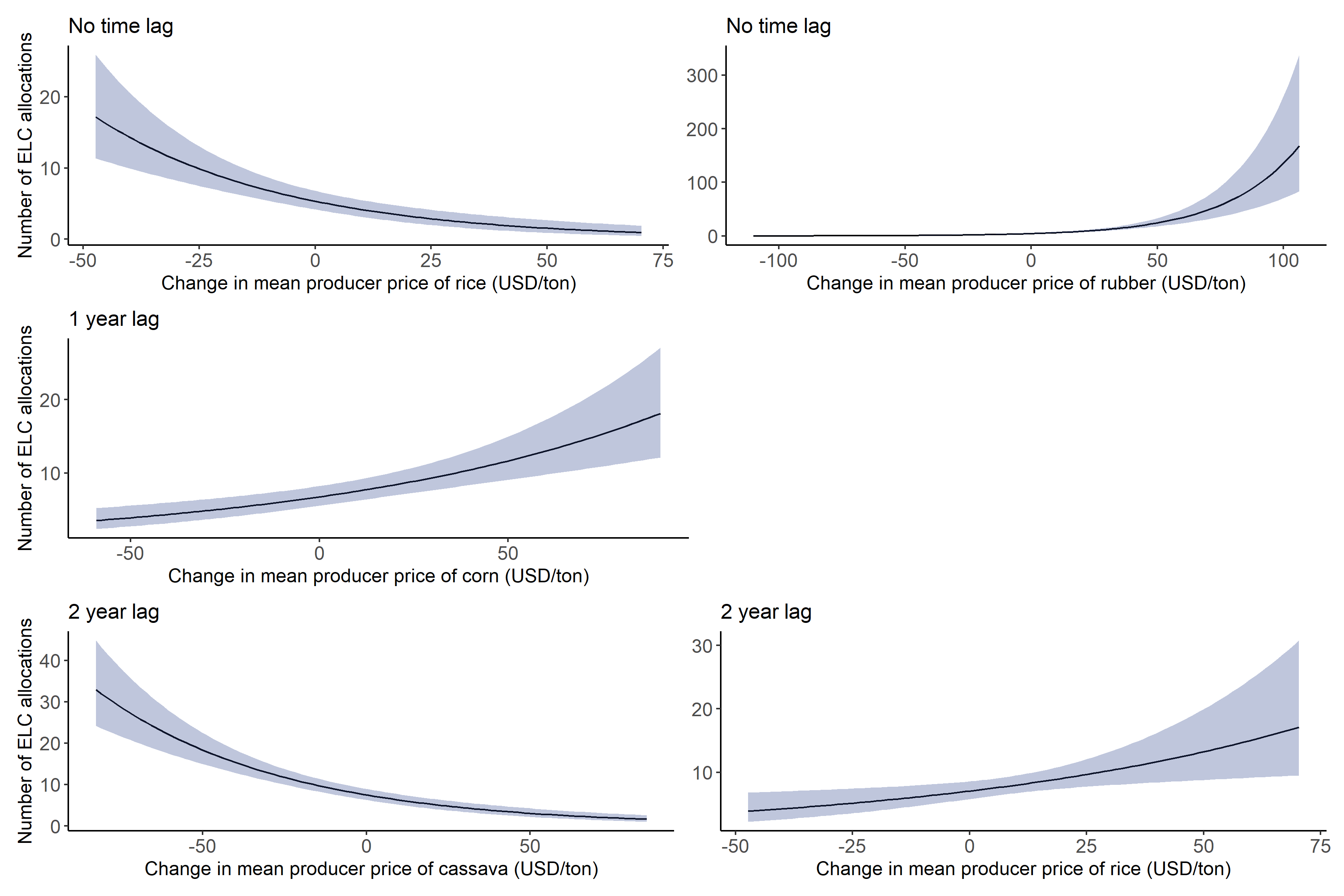
**Figure 3.** **Predicted relationship between forest loss and the producer prices (i.e. farmgate prices) variables. All y-axes are the amount of forest lost in hectares. Row a: producer price for rubber (USD/t) row b: producer price for cassava (USD/t), row c: producer price for corn (USD/t), row d: producer price for sugar (USD/t). Left column of plots are the effects on forest cover at time t (i.e. the variable values and forest loss values from the same year), the middle column of plots are the effects at time t+1 (i.e. the effects on forest loss in the subsequent year), and the right column of plots are the effects at time t+2 (i.e. the effects on forest loss two years after the variable values).**



**Figure 4. Predicted relationship between the number of new economic land concessions and macroeconomic variables in the same year (top row), and with a 1 year lag (middle row) and 2 year lag (bottom row).**



**Figure 5. Predicted relationship between the number of new economic land concessions and commodity/production variables in the same year (top two rows), and with a 1 year lag (3rd row) and 2 year lag (bottom row).**



**Figure 6. Predicted relationship between the number of new economic land concessions and producer price variables in the same year (top row), and with a 1 year lag (middle row) and 2 year lag (bottom row).**

**Socioeconomic predictors of forest cover**

This section of the analysis uses a national socioeconomic survey called the Commune Database. I have data from 2007 to 2012. The purpose of this analysis was to model the relationships between socioeconomic factors and forest cover to see if forest cover could be predicted by socioeconomics. The dataset was originally at the village level, but I aggregated up to the next administrative level – the commune.

I had 21 predictor variables in total, split into 9 sets. Because I was not explicitly interested in changes over time, I used generalised linear mixed models, with year as a random effect. I had commune as a random effect to account for repeated measurements, and commune was nested within Province, as I was not explicitly interested in the effect of province, but I was interested in how the other predictor effects might vary between provinces.

First, I fit GLMMs to each of the predictor sets to explore potential effects. These models highlighted that there were very few effects of note. Nevertheless, I used these results to select the predictors with the (relatively) strongest effects, and these were carried forward to a global model. I then built a full global model and removed variables stepwise until I had a final model.

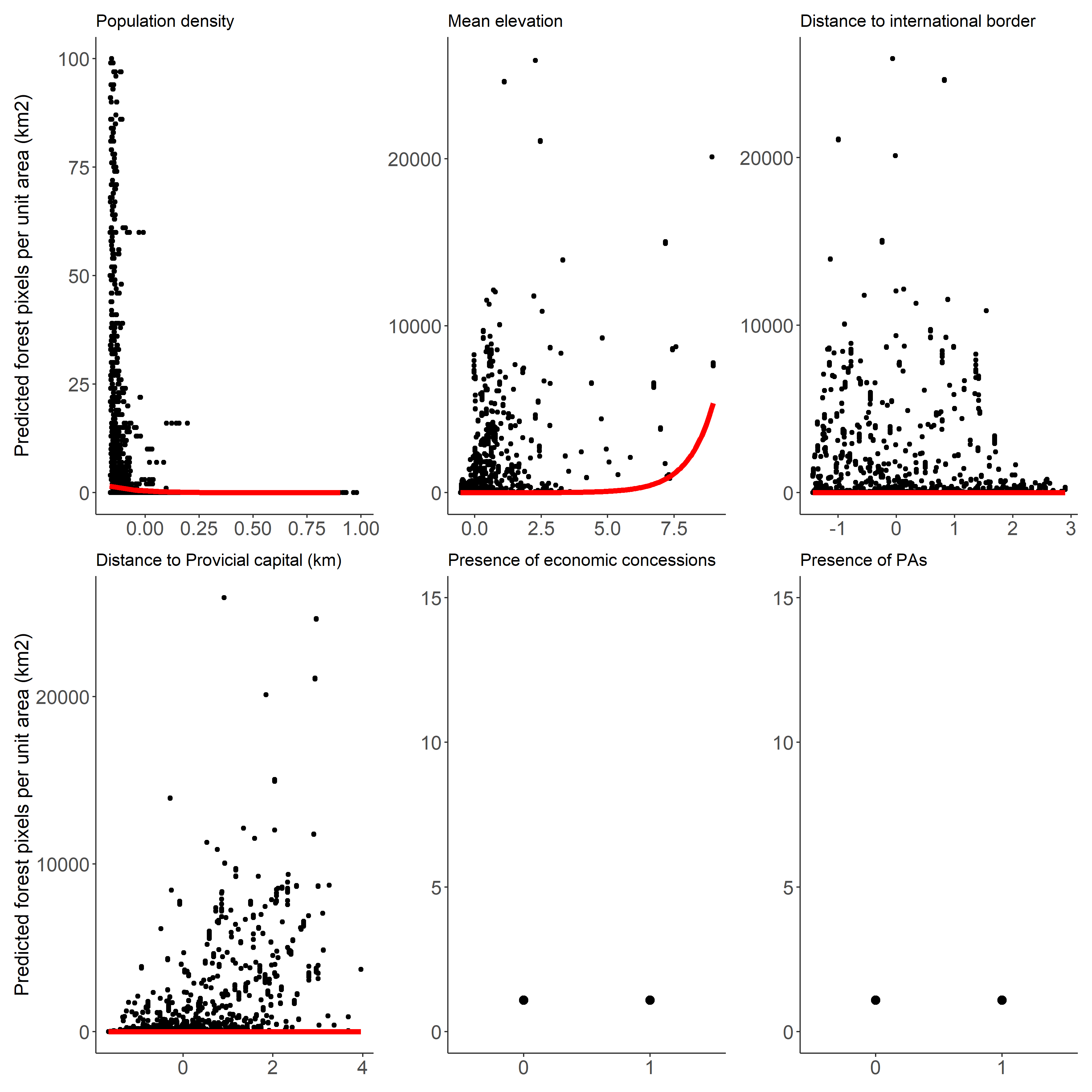
I ran global predictions from the final model (Fig 7), and I also ran predictions between provinces (Fig 8 & 9). For the between-province predictions, I predicted for each commune within a province, and then took the 50% quantile as the “mean” provincial prediction and used the 2.5 and 97.5% quantiles as confidence intervals, or “variance intervals” to display the within-province variation.

As there were serious issues with the above global model, and the resulting predictions (see Fig 7), I further aggregated the data up to the provincial level. This was intended to neutralise the between-commune variation. Here the random effects structure was year as a random slope and Province as a random intercept. First, I ran GLMMs with the variables as continuous as I did above, but the raw data suggested that there were two “peaks” in the data for some of the variables, and the resulting predictions were clearly very poor (Fig 10). The shape of the data suggested there were potentially two “types” of province for some of the variables, and so I converted the data into “high” and “low” categories by splitting the variables by their mean. These models again returned very little of interest (Fig. 11). The only significant effects were from economic land concessions and protected areas, where the model predicted significantly higher forest cover in provinces where there were land concessions and protected areas. This was not surprising, as the provinces with very little forest cover are unlikely to have PAs. It is quite interesting (although not at all surprising) that provinces with high forest cover are more likely to have economic concessions, as it is well known that economic concession (i.e. industrial agricultural concessions) were almost exclusively awarded on forested land. This allowed the concession companies to make large profits from timber extraction prior to planting their agricultural crop.

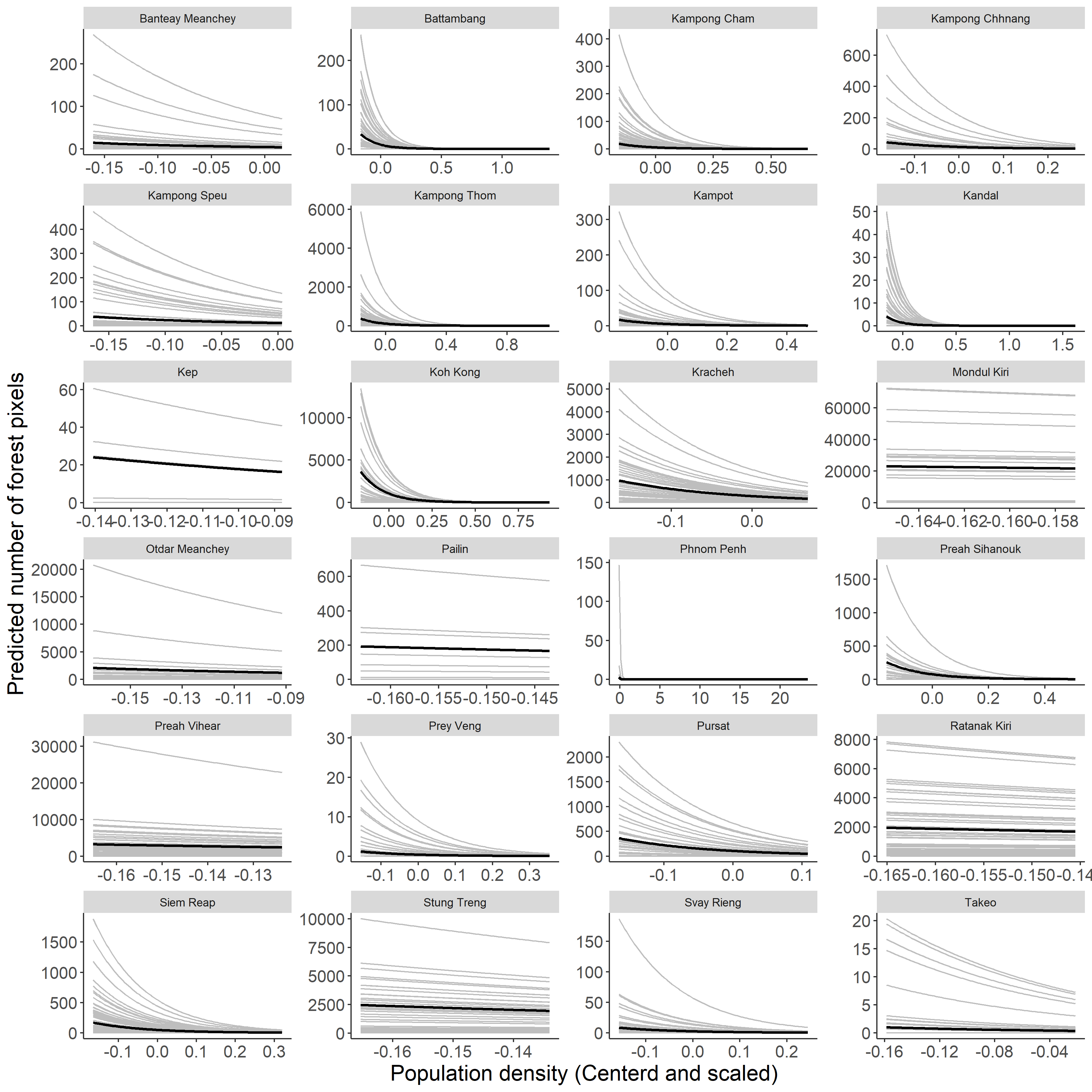
*Cluster analysis*

Finally, in the absence of any particularly interesting results I conducted a cluster analysis to see if the provinces could be categorised into broader typologies based on socioeconomics (and environmental factors, including forest cover). I used two clustering approaches, to cluster the provinces in different ways. First, I used K-means clustering to get a set of “unrelated” clusters (i.e. not agglomerative). I ran 100 iterations to find the optimal number of clusters and their contents, which resulted in 9 clusters (Figs 12 & 13). I then ran some statistical tests (lm -> aov -> Tukey) to establish if there were significant differences in the variables between cluster (see Figs. 14 & 15 for examples). In summary, there are significant differences in access to services, education, economic security, migration, population demographics, social justice, and elevation between clusters, and there are some differences in the control variables (distance to border, distance to provincial capital). There are no statistical differences in forest loss between clusters, and only slight differences in forest cover between clusters.

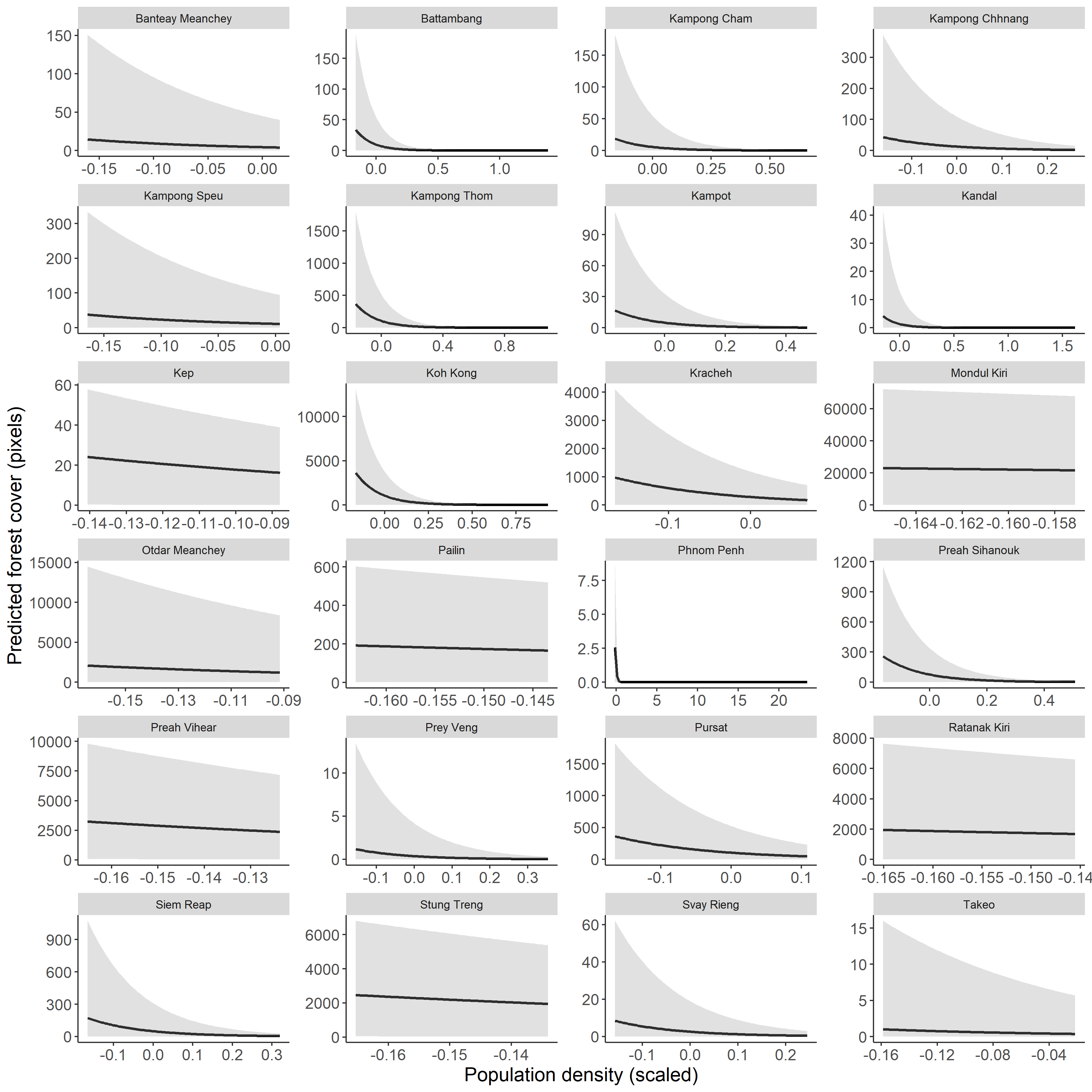
I then used an agglomerative approach to better describe the relationships between all of the provinces and to get a broader “typology” for the provinces – i.e. fewer, more interrelated clusters. I tested a range of agglomerative clustering techniques and tested them against each other using various metrics. The best method from these techniques was the unweighted pair group using arithmetic averages (UPGMA), and the optimal number of clusters, based on matrix correlation, was 5. Thankfully, the two methods (k-means and UPGMA) largely agreed on the clustering, with some obvious differences resulting from the different number of clusters. The UPGMA clustering allowed me to describe a broader socioeconomic typology of Cambodia (Figs. 16 & 17, Table 1).



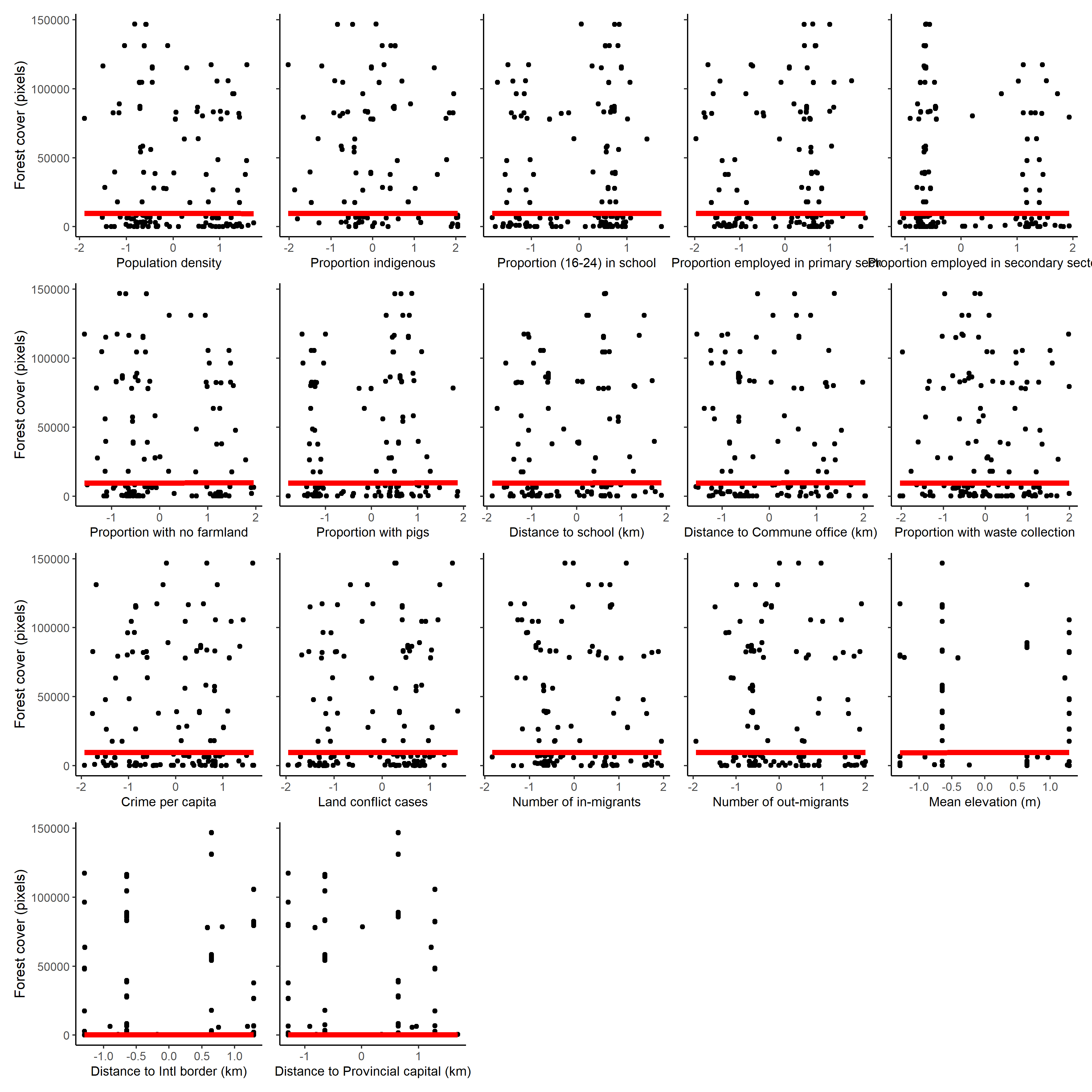
**Figure 7. Global predictions (i.e. predictions for an average commune within an average province) using the final model. X axes are the scaled predictors. Plots show the predicted values (red line) over the raw data.**



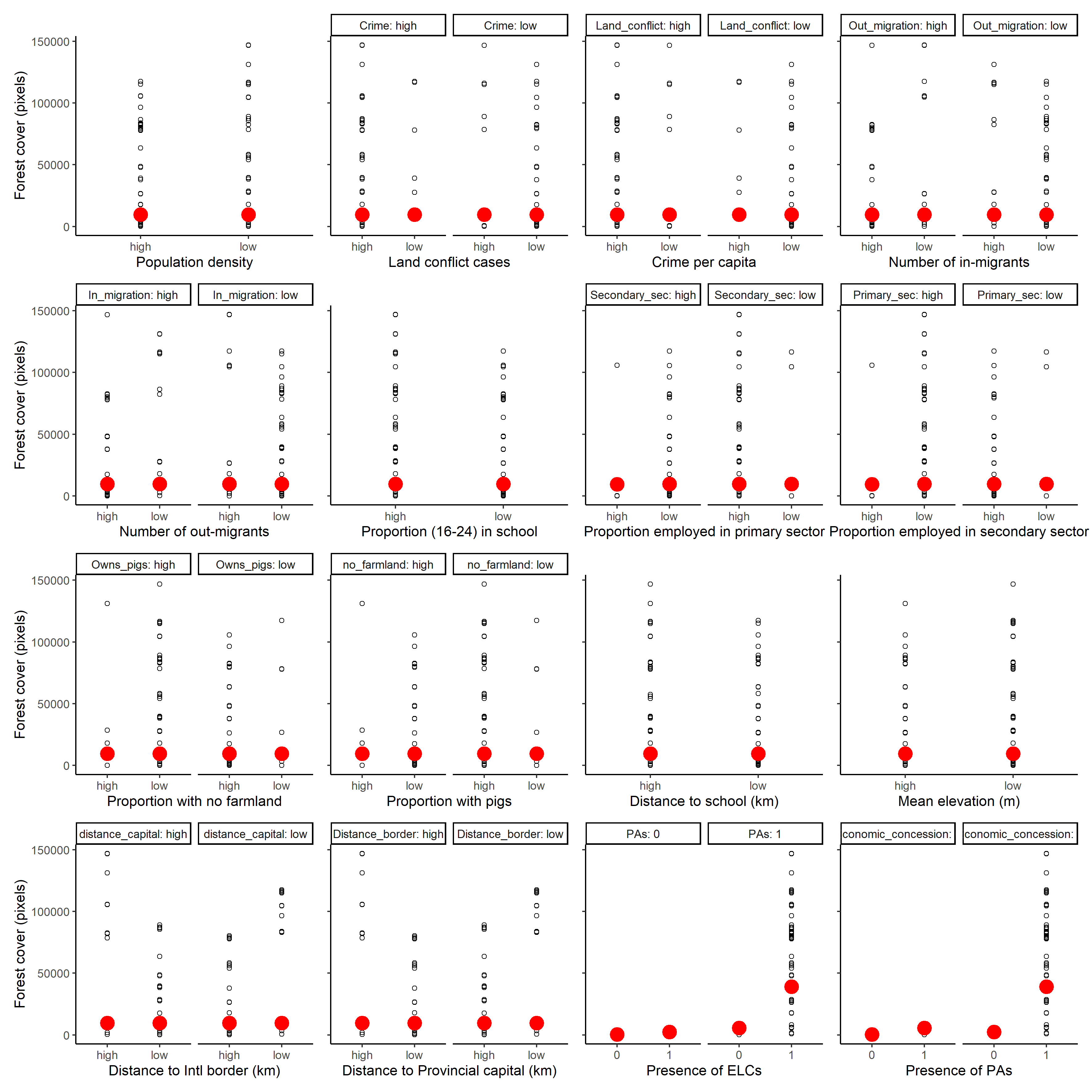
**Figure 8. Predictions within each province using the global model. Thick lines are the mean provincial predictions which are calculated as the 50% quantiles from all the commune-level predictions. Faded grey lines are the individual commune-level predictions.**



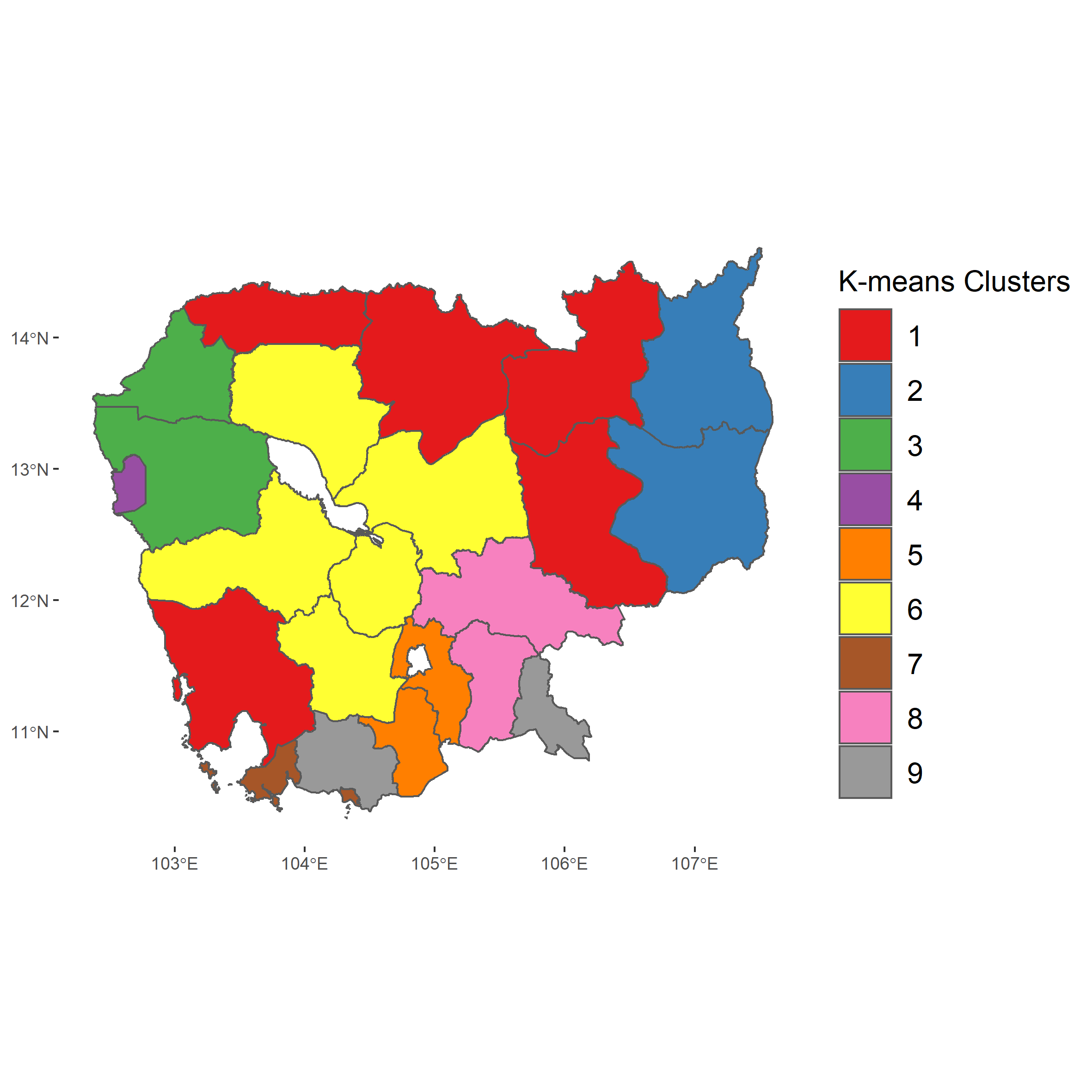
**Figure 9. As above - predictions within each province using the global model. Thick lines are the mean provincial predictions which are calculated as the 50% quantiles from all the commune-level predictions. Error ribbons are the 2.5% and 97.5% quantiles from all the commune-level predictions within each province.**



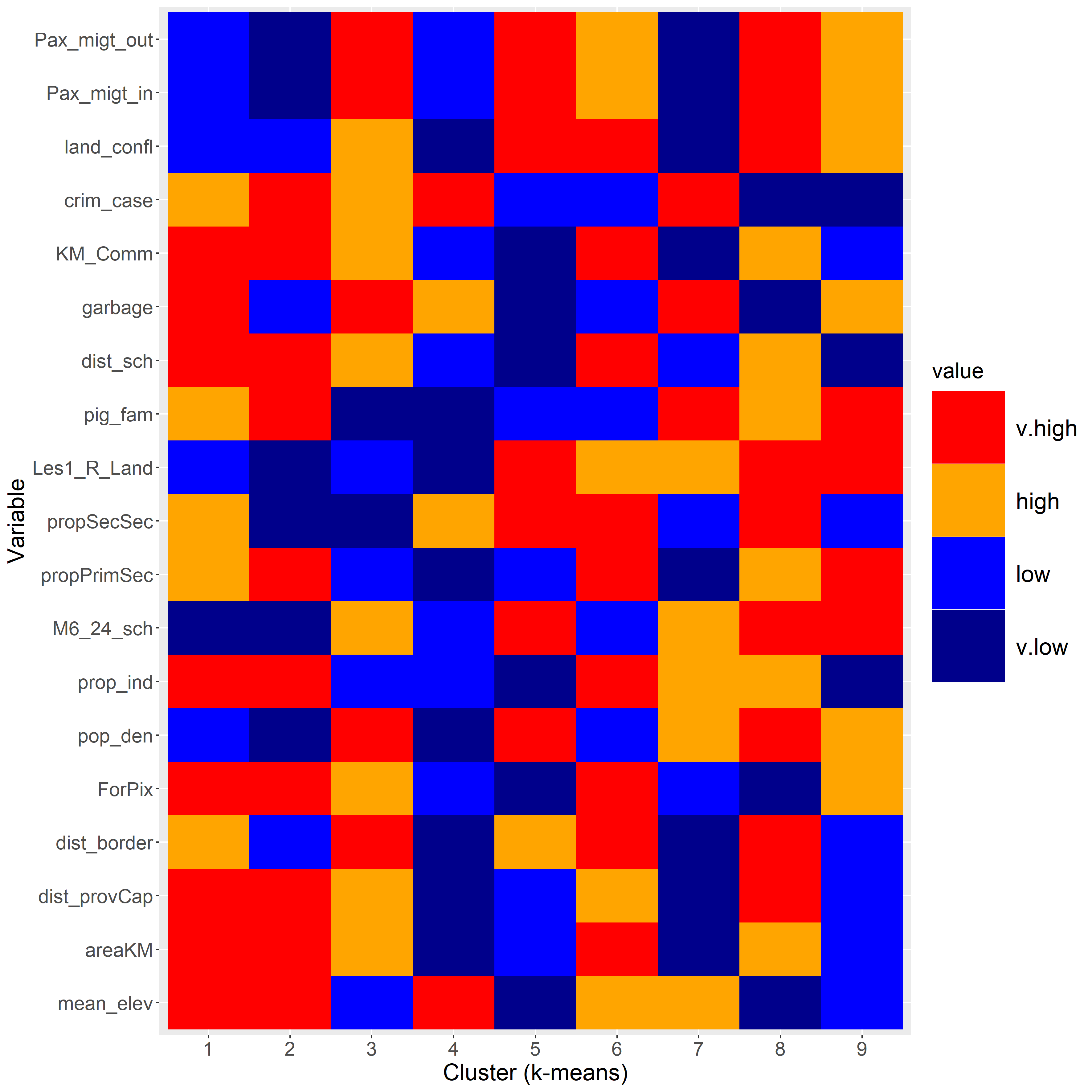
**Figure 10. Predictions from model sets using provincial-level data. Red lines are predictions, black points are raw data.**



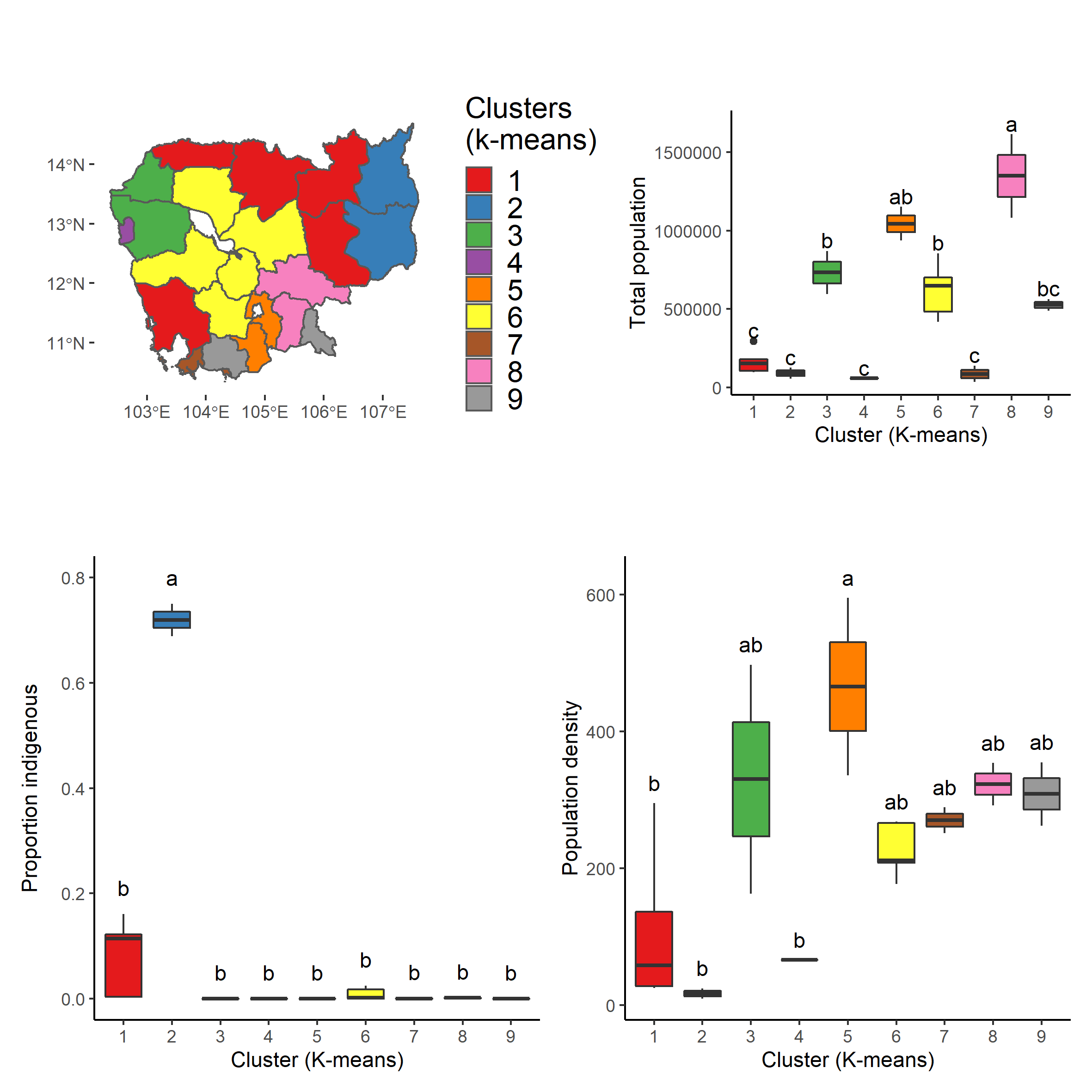
**Figure 11. Predictions from model sets using province-level data. Red points are predictions, black points are raw data. For model sets with more than one variable, the plots have been faceted to show predictions from all combinations of variable levels.**

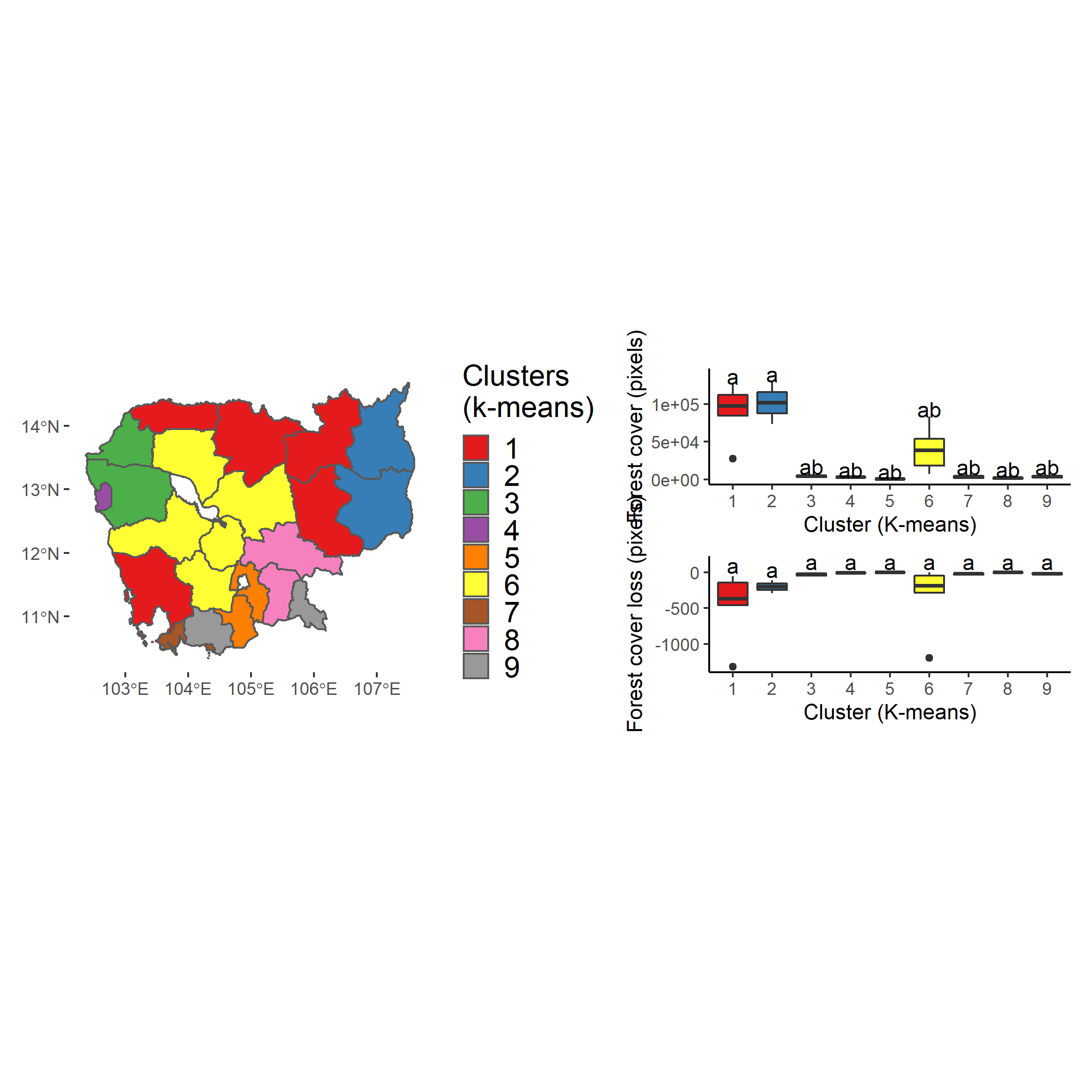


**Figure 12. Map of provinces coloured by the cluster using the k-means method.**



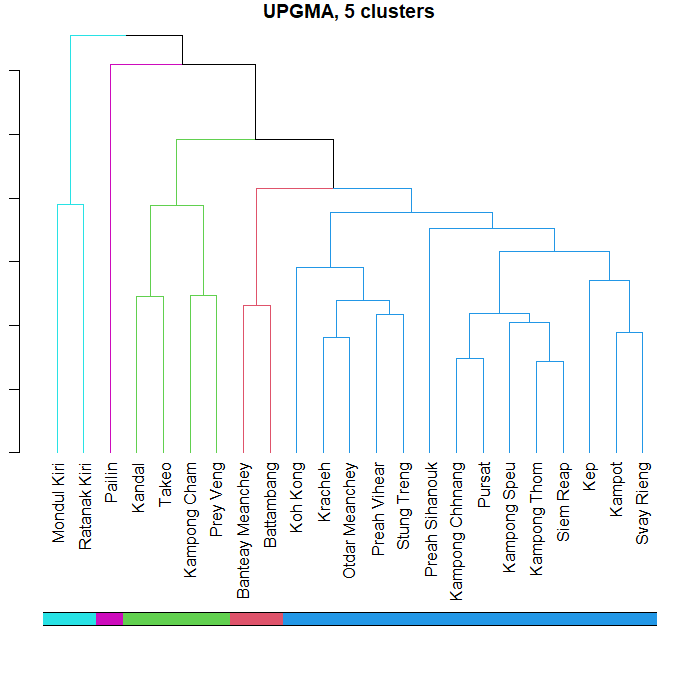
**Figure 13. Heatmap of all socioeconomic and environmental variables for each cluster (k-means). Variable values were meaned within each cluster, and then categorised based on quartiles across the full range of the variable (i.e. across all provinces). >75% = v.high, 50-74% = high, 26-50% = low, <25% = v.low.**



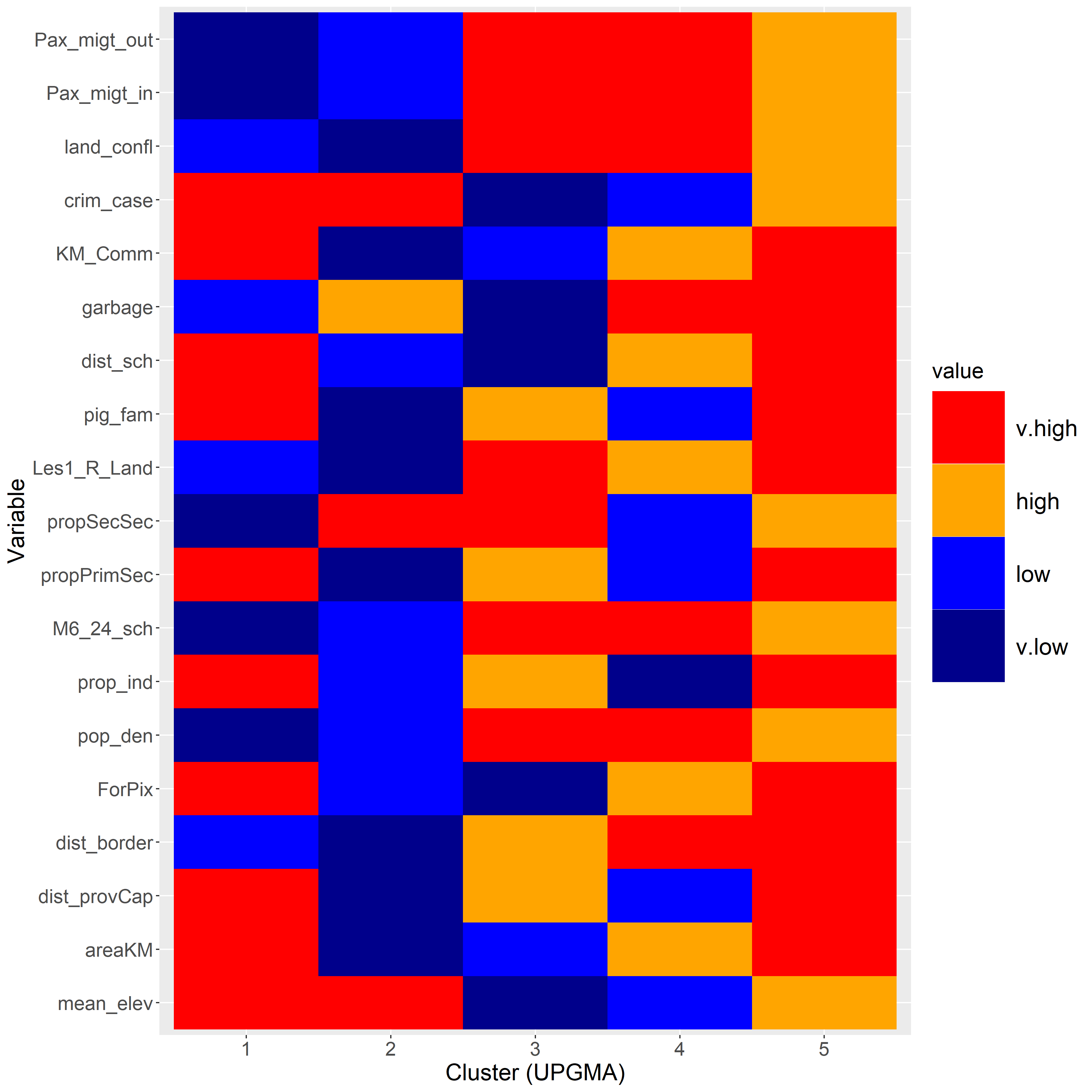


**Figure 14. Differences in population demographics between k-mean clusters. Letters denote groups according to post-hoc statistical tests (Tukey test).**

**Figure 15. Differences in forest cover (top) and forest loss (bottom) between k-means clusters. Letters denote groups according to post-hoc statistical tests (Tukey test).**



**Figure 16. Provinces clustered according to the unweighted pair-group using arithmetic means (UPGMA) method, with K=5.**



**Figure 17. Heatmap of all socioeconomic and environmental variables for each cluster (UPGMA). Variable values were meaned within each cluster, and then categorised based on quartiles across the full range of the variable (i.e. across all provinces). >75% = v.high, 50-74% = high, 26-50% = low, <25% = v.low.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **UPGMA cluster** | **K-Means cluster** | **Provinces** | **Description** | **Forest cover** |
| 1 | 2 | Mondulkiri, Ratanikiri | Very large provinces with very high elevations. Very low population density, and very high proportion of indigenous people. Very low education levels, very high proportion of primary sector workers and very low proportion of secondary sector workers. Economic security provided by rural livelihoods - few people have no farmland and livestock ownership is common. Very low access to services, high crime per capita, low land conflict and very low migration levels. | VERY HIGH |
| 2 | 4 | Pailin | Very small provinces with very high elevations. Low population density and low proportion of indigenous people. Low levels of education, low proportion of people in the primary sector but higher proportion of people in the secondary sector. Very few people with no farmland, but very little livestock ownership. High access to services and high crime per capita. Low land conflict and low migration. | LOW |
| 3 | 5, 8 | Kampong Cham, Kandal, Prey Veng, Takeo | Small provinces with very low elevations. Very high population density and high proportion of indigenous people. Very high levels of education, high proportion of people in the primary sector, but very high proportion of people in the secondary sector. High proportion of people with no farmland, but high levels of livestock ownership. High access to services and low crime per capita. But very high migration levels and very high rates of land conflict. | VERY LOW |
| 4 | 3 | Banteay Meanchey, Battambang | Large provinces with low elevations. Very high population density and very low proportion of indigenous people. Very high levels of education, and relatively low proportion of workers in the primary and secondary sectors (suggesting higher proportions in the other sectors e.g. tertiary). High proportion of people with no farmland, and low levels of livestock ownership (suggesting very urban). Low access to services, but this may be explained by the mean size of the provinces in this cluster (there is high access to garbage collection). Low crime per capita, but very high migration and very high rates of land conflict | HIGH |
| 5 | 1, 6, 7, 9 | Kampong Chhnang, Kampong Speu, Kampong Thom, Kampot, Kep, Koh Kong, Kracheh, Otdar Meanchey, Preah Sihanouk, Preah Vihear, Pursat, Siem Reap, Stung Treng, Svay Rieng | Very large provinces with high elevations. High population density and very high proportion of indigenous people. High levels of education, and a high proportion of workers in both primary and secondary sectors. Very high proportion of people with no farmland, but also very high proportion of people with livestock. Low access to services (although very high access to garbage collection) - this may be an artefact of the very large mean area of the provinces in this cluster. Very high crime rates, very high migration, and very high rates of land conflict. | VERY HIGH |

**Table 1. Descriptive typology of provinces in Cambodia based on socioeconomic and environmental variables**