**Introduction**

**Methods**

Study area?

*Data sources*

Socioeconomic variables were extracted from the Cambodian Commune Database for the years 2007 – 2012 (Table 1) which are available from Open Development Cambodia ([www.opendevelopmentcambodia.net](http://www.opendevelopmentcambodia.net)). Data on economic land concessions, protected areas, and elevation (digital elevation model), and shapefiles for the country, provinces, and communes were provided by the Royal Government of Cambodia (via the Wildlife Conservation Society). Forest cover layers were taken from the publicly available European Space Agency Climate Change Initiative (ESACCI) satellite data.

*Variable selection*

The response variable was forest cover area and was calculated using the ESACCI data product (see ‘data processing’ below). Socioeconomic and control variables were selected based on a combination of previous studies, data availability, and the authors’ knowledge of Cambodia. Socioeconomic variables were selected to create 8 variable sets reflecting different aspects of socioeconomic status and development, each of which was hypothesised to be either a driver or predictor of forest cover (Table S3, Luck et al. 2009; Ty et al. 2012; Kristensen et al. 2016; Bonilla-Bedoya et al. 2018). The variable sets were population demographics (n=8), education (n=4), employment (n=5), economic security (n=2), access to services (n=4), social justice (n=2), migration (n=2), and control variables (n=6). Control variables were included to account for the effects of environmental and other human factors including economic land concessions (Abdullah and Nakagoshi, 2007; Davis et al., 2015), protected areas (Bonilla-Bedoya et al., 2018), elevation (Ty et al., 2012), and distance to human infrastructure (Ty et al., 2012). A habitat control variable was excluded because the response variable (forest cover) was extracted from a land cover layer and represented a specific type of habitat, resulting in non-independence between the response and habitat.

*Data processing*

The forest cover response variable was extracted from the ESACCI product by totalling the number of pixels (1 Pixel = 0.09km2) in each year classified as bands 50, 60, 61, 62, 70, 71, 72, 80, 81, 82, 90, and 100 (Table S4). The forest cover layer was stratified into forest cover per commune and forest cover per province. Forest cover data processing was done in QGIS (QGIS Geographic Information System v3.16). Predictor variables were checked for collinearity, and if two variables in the same set had a correlation coefficient of >0.6 then generally one was removed (Supporting Information).

Data from the Commune Database were at the resolution of individual village, and so the selected variables (Table 1) were aggregated to the commune and province level after error checking and cleaning (Supporting Information). This resulted in between 1,317 and 1,512 communes, and 23 Provinces (excluding Phnom Penh). The number of communes changed between years due to administrative changes. Some variables were converted from raw values to proportional data to account for large differences in commune and province size and human population (Table 1). Data were checked for errors in R (Supporting Information, R Core Team, version 4.0).

*Modelling*

This analysis aimed to model the relationships between forest cover and socioeconomic variables within communes between 2007 – 2012. The results of initial commune-level modelling prompted further aggregation of the data to the province-level and models were built to investigate the relationships between forest cover and socioeconomic variables within provinces for the same time period.

Commune-level models

Generalised linear mixed models (GLMM) with Poisson errors were built with commune nested within province as random intercept terms to account for repeat measurements and the hierarchical data structure, and year as a random slope term to account for temporal autocorrelation (Zuur et al., 2009). The natural logarithm of commune area (km2) was used as an offset term in all models to account for large variation in commune size. Due to the large number of available predictor variables, maximal within-set models were run first for each of the 8 variable sets (Table S8), and variables with very weak, or no effect were dropped. Simplified models were compared with maximal models using likelihood ratio tests and analysis of variance tests. If a variable set had only one variable, this was automatically taken forward. Because assessment of term significance in GLMMs is complex, predictions and plots were made for all terms before being dropped to ensure noteworthy effects were not being missed. This process resulted in a final set of 13 variables which were used to create a candidate set of 10 models (Table S19). Following an information theoretic approach (Burnham and Anderson, 2007) models were compared via AIC to select the top model or models. The resulting final model fit was assessed via diagnostic plots (residuals versus fitted, quantile-quantile of random effects, Supporting Information, Harrison et al (2018)). Marginal (fixed effects only) and conditional (fixed and random effects) pseudo-R2 values were calculated based on Nakagawa & Schielzeth (2017) using the R package ‘MuMIn’ (Bartoń, 2020). To investigate the variation in effects between provinces, predictions were made for each variable within each commune and the 50% quantile (median) from all commune-level predictions within each province was extracted as the provincial mean prediction.

Province-level models

The same GLMM model formulation was used for the province-level models except that commune was removed from the random effects structure. Based on provincial-level histograms of predictor variables, 14 predictors were converted to categorical variables by splitting the data by the mean, resulting in “high” and “low” values (Table 1). Following an information theoretic approach, a candidate set of models was created (Table Sx) and model comparison was done using AIC.

*Cluster analysis*

Agglomerative clustering was conducted to create a typology for provinces based on the socioeconomic variables used in the analysis above. Several agglomerative clustering approaches were assessed. These were single linkage, complete linkage, unweighted pair-group using arithmetic averages (UPGMA), unweighted pair-group using centroids (UPGMC), Ward’s minimum variance, and flexible clustering. The methods were compared using cophenetic correlation and Gower distance metrics, and the appropriate number of clusters (*k*) was selected using the matrix correlation statistics (Borcard et al., 2018). The capital city of Phnom Penh, which is technically a province in itself, was removed prior to clustering because it has extreme values for many of the variables and is thus an outlier that affects the clustering.

**Table 1. Variables selected for the socioeconomic analysis. Variables range from 2007 – 2012.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Set** | **Variable** | **Transformation for analysis** | **Province-level class** | **Details** |
| Demographics | Total population |  | NA | Includes women, men, and children of all ages |
|  | Population density |  |  |  |
|  | Number indigenous | Proportion of total population | Categorical | Total number of people who are indigenous/ethnic minority |
| Education | Males aged 6 – 24 in school | Proportion of total number of males aged 6 - 24 |  | Number of males aged 6 - 24 in full time education |
| Employment | Number of adults employed in primary sector | Proportion of total adult population | Categorical | The primary sector includes agriculture (rice and other crop farming), fishing, livestock farming, forestry, and non-timber forest product collection (Kenessey 1987) |
|  | Number of adults employed in secondary sector | Proportion of total adult population | Categorical | The secondary sector includes wood-based production (e.g. furniture), metal- and glass-based production, foodstuff production, plastic- and rubber-based production, textiles production (Kenessey 1987) |
| Economic security | Number of families with <1ha rice land (including no rice land) | Proportion of total number of families | Categorical |  |
|  | Number of families who keep pigs | Proportion of total number of families | Categorical |  |
| Access to services | Distance to nearest school |  | Categorical | Median distance from any village in the commune to the nearest school (primary or secondary) |
|  | Number of families with access to waste collection | Proportion of total number of families |  |  |
|  | Distance to the Commune Office |  |  | Median distance from any village in the commune to the Commune Office (government administration office) |
| Social justice | Number of criminal cases | Criminal cases per capita | Categorical | Includes murder, theft, and other criminal cases |
|  |  |  |  |  |
|  | Number of land conflict cases |  | Categorical | In the previous 12 months |
| Migration | Number of in-migrants |  | Categorical | Migration into the commune |
|  | Number of out-migrants |  | Categorical | Migration out of the commune |
| Control | Mean elevation (masl) |  | Categorical | Mean elevation for the commune |
|  | Distance to international border (km) |  | Categorical | Distance from the centre of the commune to the nearest international border |
|  | Distance to Provincial Capital (km) |  | Categorical | Distance from the centre of the commune to the centre of the provincial capital (town or city) |
|  | Presence of economic land concessions |  |  | Binary. 1 = part or all of an economic land concession falls within the boundary of the commune, 0 = no economic land concession falls within the commune boundary |
|  | Presence of protected area |  |  | Binary. 1 = part or all of an protected area falls within the boundary of the commune, 0 = no protected area falls within the commune boundary. "Protected area" includes Wildlife Sanctuary, National Park, Protected Landscapes, Multiple-use areas, RAMSAR sites |
|  | Protected area category |  |  | None = no protected area falls within commune, MULTI = more than one category of protected area falls within commune, WS = wildlife sanctuary, NP = national park, PL = protected landscape, MUA = multiple-use area, RMS = RAMSAR |

**Results**

*Commune-level model*

Initial within-set model selection resulted in a final candidate set with 10 models and 13 unique variables (Table S19). There was a single top model according to AIC (m1), with all other models having delta AIC values of more than 18 (Table S19). The top model only had one non-control variable - population density (Table 4). The random effects term with the highest variance was Commune (10.45 [SD = 3.23], 60% of the total random effect variance), followed by Province (6.77 [SD = 2.60], 39% of the total random effect variance, Table 4). The variance explained by year at both the commune and province level was low (0.005 [SD = 0.068] and 0.0005 [SD = 0.022] respectively), contributing approximately 1% of the total random effect variance (Table 4). The marginal R2 (fixed effects only) was 0.78 (78%), and the conditional R2 (fixed and random effects) was 1, suggesting that most of the model variance was explained by the fixed effects. The largest positive effect was from mean elevation (rate ratio = 2.861, Table 4) which relates to 0.6 forest pixels (0.06 km2) predicted within an “average” commune (i.e., all other fixed and random effects set to their mean) when mean elevation is at the minimum within the country. When the mean elevation is at the maximum found within the country (and all other terms are set to their mean), the number of forest pixels predicted is 13,380 (1,204 km2). This highlights that higher elevation areas of Cambodia are much more likely to be forested than lower elevation areas. The strongest negative effect was from population density (rate ratio = 0.001, Table 4) which relates to approximately 1.5 predicted forest pixels (0.14 km2) at the minimum value of population density found within the country, contrasting with a prediction of effectively zero (2.22 × 10-16) forest pixels at the highest value of population density within the country. All other model terms, excluding the presence of ELCs, had positive effects on forest cover (Table 4). These effects suggest that remote communes (large distances to provincial capitals) that are centrally located within the country (far away from international borders) are predicted to have high forest cover. Interestingly, although the effects are weak, communes that contain ELCs are predicted to have lower forest cover than those without, and communes with protected areas are predicted to have higher forest cover than those without (Table 4).

The results from the final commune-level model must, however, be viewed with extreme caution because model validation revealed some serious underlying issues. As is suggested by the variance associated with the commune-level random effect term, there was extreme variation between communes for all variables (predictors and response, Figure 5). This between-group variance results in the model being unsuitable for generalised (i.e., ‘global’) predictions (Figure 5). Intercept and slope estimates between communes, even within the same province, varied hugely (Figure 6), and this issue was highlighted in diagnostic plots where we see that the assumption of normality of deviations of the conditional means of the random effects (for commune) from the global intercept is violated (Figure S6). Furthermore, the model residuals displayed heteroskedasticity, with the model predicting particularly poorly for lower values of the response (Figure S7). Therefore, drawing general inferences about the relationships between forest cover and socioeconomics at the country level using this model is inappropriate.

*Province-level model*

The province-level models were run to eliminate the commune-level variation and to identify any broader relationships between forest cover and socioeconomics. A candidate set of 19 models was built and an evaluation of AIC selected a single model (m8) as the top model (Table S20). Model m5 had some support (delta AIC = 5, Table S20) but was a simpler version of m8 and therefore inferences were drawn from m8 alone. The random effects term with the highest variance was Province (1.18 [SD = 1.08], which constituted 99% of the total random effects variance), followed by year (0.006 [SD = 0.077], which was 1% of the total random effects variance). The marginal R2 (fixed effects only) was 0.71 (71%) and the conditional R2 (fixed and random effects) was 0.99 (99%), suggesting that the majority of model variance was explained by the fixed effects. Presence of ELCs and presence of PAs had the largest two positive effects relative to their refences levels (no ELCs, no PAs), suggesting that provinces that have those two features are predicted to also have higher forest cover (rate ratios = 1.51 and 1.64 respectively). In provinces where the proportion of males in school and distance to school are both low, higher levels of forest cover are predicted compared with provinces where these variables are high. Furthermore, in provinces where elevation, distance to an international border, and distance to the provincial capital are low, forest cover is predicted to be higher than in provinces where these variables are high. However, all the above effects are weak (Figures 7 & 8). For example, the difference in the predicted number of forest pixels between a province with a low proportion of males in school and a province with a high proportion (with all other variables set to low), is 200 (18 km2). The difference in the number of predicted forest pixels between a province with low median distances to schools and a province with high median distances (with all other variables set to low), is 689 (62 km2). As standalone figures these appear large, but in the context of the range of the response variable (minimum value of 54 forest pixels to a maximum of 146,876 forest pixels), the effects are relatively weak. Presence of PAs had the largest effect on predicted forest pixels. The number of forest pixels predicted for a province with PA presence is 36,890 (3,320 km2) higher than for a province with no PA presence. This emphasises the relationship between forested land and protected areas in Cambodia. The size of the effects for the two socioeconomic predictors (proportion of males in school, and distance to school) in the top model suggest that these variables have little power to predict forest cover at the provincial level in Cambodia, but that the presence of protected areas and economic land concessions do.

***Cluster analysis***

The UPGMA clustering had the highest cophenetic correlation (0.79) and the lowest Gower distance (254.14) and was therefore selected. The matrix correlation statistic suggested that 4 clusters were optimal, but that between 3 and 7 clusters had very similar support. When divided by 4 clusters, there was a large group (16) of provinces that fell into a single cluster, and so 5 clusters were chosen to add further nuance (Figure 9). The provinces within clusters were geographically contiguous (Figure 10), although clusters that had smaller cophenetic distances (i.e., were closer on the dendrogram, Figure 9) were not necessarily geographically contiguous. The largest cluster (cluster 5) dominated a central strip of the country, separating the smaller, and more similar clusters (Figure 10). Only clusters 2 and 4 were contiguous with each other. These results suggest that provinces often have similar socioeconomic conditions to that of their neighbours, but that there are also distinct regions within the country that can be characterised by their socioeconomics rather than their geography. A heatmap of the socioeconomic variable values for each cluster revealed some distinguishing patterns (Figure 11). The largest cluster (cluster 5) was distinguished by high or very high values of all variables, which translates to generally large provinces with high population density, high education levels, high proportions of primary and secondary sector workers, and high migration (Table 5). This contrasts with cluster 2, which has predominantly low values for the socioeconomic variables which translates to very small provinces with low population density, low levels of education, low levels of primary sector employment (higher secondary sector employment), and low levels of migration (Table 5). Clusters 3 and 4 had the highest levels of migration (and interestingly the highest levels of land conflict), education, and population density, reflecting the presence of two of the three largest cities and significant urban development. Cluster 1 had the lowest population density, education, proportion of secondary sector workers, and migration, reflecting the clusters remote geography and rural character. Provinces within cluster 1 were also the most forested but had also lost the most forest during the study period (Figure 12). Provinces within cluster 5 were generally the next most forested after cluster 1 and had also lost large areas of forest during the study period (Figure 12). Cluster 3 had the least amount of forest, which was expected due to high levels of urbanisation and agriculture. Clusters 1 and 2 had the highest elevation, and clusters 1 and 5 had the highest mean distance to a provincial capital (Figure 12).

**Table 4. Model outputs and rate ratios from the top models from the socioeconomic analysis. Outputs are for the commune-level analysis and the province-level analysis. Reported coefficients are on the link (log) scale.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Variance** | **Std.Dev** | **Coefficient** | | **SE** | **Rate ratio*a*** |
| ***Commune-level final model*** |  |  | |  |  |  |
| *Random effects* |  |  |  | |  |  |
| Commune (intercept) | 10.4500 | 3.2334 | - | | - |  |
| Year/Commune (slope) | 0.0046 | 0.0680 | - | | - |  |
| Province (intercept) | 6.7730 | 2.6025 | - | | - |  |
| Year/Province (slope) | 0.0005 | 0.0220 | - | | - |  |
| *Fixed effects* |  |  |  | |  |  |
| Intercept | - | - | -4.6240 | | 0.5620 |  |
| Population density | - | - | -7.5140 | | 1.1270 | 0.001 |
| Mean elevation | - | - | 1.0510 | | 0.1220 | 2.861 |
| Distance to In'tl border | - | - | 0.5805 | | 0.2036 | 1.787 |
| Distance to Provincial capital | - | - | 0.6929 | | 0.1114 | 2.000 |
| ELC presence | - | - | 0.0000 | | 0.0025 | 0.999 |
| PA presence | - | - | 0.0093 | | 0.0143 | 1.009 |
| ***Province-level final model*** |  |  |  | |  |  |
| *Random effects* |  |  |  | |  |  |
| Province (intercept) | 1.1762 | 1.0845 | - | | - |  |
| Year/Province (slope) | 0.0058 | 0.0765 | - | | - |  |
| *Fixed effects* |  |  | - | | - |  |
| Intercept | - | - | -2.9900 | | 0.4497 |  |
| Males in school (low) | - | - | 0.0051 | | 0.0019 | 1.002 |
| Distance to school (low) | - | - | -0.0174 | | 0.0022 | 1.002 |
| Mean elevation (low) | - | - | -0.0223 | | 0.0024 | 1.002 |
| Distance to border (low) | - | - | 0.0061 | | 0.0019 | 1.002 |
| Distance to Prov capital (low) | - | - | -0.0072 | | 0.0019 | 1.002 |
| Presence of economic concessions (1) | - | - | 1.9974 | | 0.4090 | 1.505 |
| Presence of PAs (1) | - | - | 2.8063 | | 0.4965 | 1.643 |

*a* Rate ratio = exp(coefficient)

Chart, scatter chart

Description automatically generated

**Figure 5. Predicted relationships (red lines) between socioeconomic variables and forest cover in Cambodia between 2007 – 2012 from the top commune-level model. Predictions are ‘global’ i.e., all random effects were set to their mean values, and thus predictions are not for any specific commune. Black dots are the raw data points of each predictor versus forest cover.**

Diagram, shape, arrow

Description automatically generated

**Figure 6. Predicted relationships between population density and forest cover within Cambodian provinces between 2007 – 2012 using the top commune-level model. Faded grey lines are the predictions for each individual commune within each province. Black lines are the mean provincial predictions, which were computed using the 50% quantile from all commune predictions. Plot panels have non-standard y axis ranges.**

A picture containing diagram

Description automatically generated

**Figure 7. Predicted forest cover within each Cambodian province given high and low levels of school attendance (males aged 6 – 24 in school) from the top province-level model. All other variables in the model were set to their reference level (distance to school = low, elevation = low, distance to international border = low, distance to provincial capital = low, economic land concession = yes, protected area = yes).**

A picture containing diagram

Description automatically generated

**Figure 8. Predicted forest cover within each Cambodian province given high and low distances to the nearest school from the top province-level model. All other variables in the model were set to their reference level (school attendance = low, elevation = low, distance to international border = low, distance to provincial capital = low, economic land concession = yes, protected area = yes).**

Chart

Description automatically generated

**Figure 9. Cambodian provinces clustered based on socioeconomics. Data were averaged across the study period 2007 – 2012. Variables included were total population, population density, number of land conflict cases, number of criminal cases per capita, number of in- and out-migrants, the proportion of the population classified as indigenous, proportion of males aged 6 – 24 in school, proportion of the population employed in the primary and secondary sectors, proportion of families with no access to agricultural land, proportion of families who kept pigs, distance to the nearest school, proportion of families with access to waste collection, and distance to the commune (administrative) centre. The clustering method was unweighted pair-group using arithmetic averages (UPGMA).**

Map

Description automatically generated

**Figure 10. Map of Cambodia showing the clusters resulting from the unweighted pair-group using arithmetic averages (UPGMA) method. Provinces are labelled. The upper white polygon is the Tonle Sap lake, and the lower white polygon is the city of Phnom Penh, both of which were excluded from the analysis.**

Chart, bar chart

Description automatically generated

**Figure 11. Heatmap showing the variable values for each cluster. Variables were categorised as “v.low” if the mean (across provinces within that cluster) was below the 25% quantile for that variable across the whole country, “low” if the mean was above 25 and below 50%, “high” if the mean was above 50% but below 75%, and “v.high” if the mean was above the 75% quantile. Pax\_migt\_out = numbers of out-migrants, Pax\_migt\_in = numbers of in-migrants, land\_confl = number of land conflicts, crim\_case = criminal cases per capita, KM\_Comm = distance to commune office, garbage = proportion of families with access to waste collection, dist\_school = distance to nearest school, pig\_fam = proportion of families who keep pigs, Les1\_R\_Land = proportion of families with no rice land, propSecSec = proportion of adults employed in the secondary sector, propPrimSec = proportion of adults employed in the primary sector, M6\_24\_sch = proportion of males aged 6-24 in education, prop\_ind = proportion of the population that is indigenous, pop\_den = population density.**

Diagram

Description automatically generated

**Figure 12. Boxplots showing the distribution of environmental variables for each cluster: *a* = mean forest area, *b* = mean area (km2), *c* = change in forest cover (between 2007-2012), *d* = mean elevation (masl), *e* = mean distance to international border, *f* = mean distance to a provincial capital. Boxplots show the median (centre line within boxes), 25 and 75% percentiles (box edges), and minimum and maximum values (upper and lower whiskers, not exceeding 1.5 × interquartile range). 5 UPGMA clusters.**

**Table 5. Descriptive typology of the provinces and clusters within Cambodia, clustered using socioeconomic variables and the unweighted pair group using arithmetic mean (UPGMA)**

|  |  |  |
| --- | --- | --- |
| **UPGMA cluster** | **Provinces** | **Description** |
| 1 | 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. |
| 2 | Pailin | Very small province 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. |
| 3 | 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. |
| 4 | 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 |
| 5 | Kampong Chhnang, Kampong Speu, Kampong Thom, Kampot, Kep, Koh Kong, Kracheh, Otdar Meanchey, Preah Sihanouk, Preah Vihear, Pursat, Siem Reap, Stung Treng, Svay Rieng | 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. |