**Introduction**

Tropical deforestation is a significant threat to biodiversity, ecosystem processes, and local people (Frewer & Chan 2014; Estoque et al. 2019), and is particularly insidious in its complexity (Mena et al. 2006; Rowcroft 2008; Kong et al. 2019). The drivers of forest loss in the tropics are not only numerous and multifaceted, but they operate at multiple scales and are comprised of complex feedback loops between ecological, biophysical, social, cultural, political, and economic factors (Geist & Lambin 2002, 2003; Shrestha et al. 2018; Xu et al. 2019; Mannan et al. 2019). This complexity means that underlying drivers operating at national, regional, or even global scales manifest themselves in a variety of proximate causes, which themselves are governed and shaped by local conditions (Geist & Lambin 2002; Fox & Vogler 2005; Van Den Hoek et al. 2014). The dynamics between drivers at different scales makes disentangling the causes of deforestation highly contextual, increasing the value of local studies. Local socioeconomic conditions are important factors in understanding the link between broader drivers of land use change (LUC) and deforestation, and can be effective predictors of forest loss (Redo et al. 2012; Liu et al. 2016; Bonilla-Bedoya et al. 2018). Proximate causes of deforestation such as agricultural expansion and infrastructure development are often closely linked via feedback loops and dependencies to socioeconomic conditions including poverty, migration, local economies, and land and wealth inequality (Geist & Lambin 2002; Khuc et al. 2018). Therefore, understanding the link between socioeconomics and forest cover at different scales is a crucial step in the development of effective economic and environmental policies that have positive effects on both people and forests.

Socioeconomics can encompass a huge variety of conditions that describe the social, demographic, and economic state of local people, and can affect which drivers of LUC are most influential in causing forest loss (Mena et al. 2006). The complexity of social-ecological systems means that it is challenging for researchers to identify and model all the correct and appropriate socioeconomic predictors of forest loss, but there is a wealth of research that has helped to untangle certain relationships in specific locations and at specific scales. At the local scale, socioeconomic drivers and economics that affect decision-making processes within households, coupled with institutional factors, are often the most relevant for influencing LUC (Van Den Hoek et al. 2014; Gatto et al. 2015). Poverty was once believed to be the single most important socioeconomic driver of deforestation (e.g., Lomborg 2001), although more recent research has added significant nuance to this argument (Geist & Lambin 2003), while other studies have demonstrated that poverty played very little role at all in deforestation (Onojeghuo & Blackburn 2011). Poverty itself is a complex metric encompassing a multitude of factors such as income, wealth, land, agriculture, migration, education, and healthcare, all of which interact to influence deforestation (Khuc et al. 2018). Inequalities in land, income, and wealth, and insecure land tenure and forest rights for local people are all common factors in driving deforestation (Ceddia 2019). Such inequalities, in combination with debt and overpopulation, drive the expansion of agriculture and other natural resource-based activities, to the detriment of forests, as local people strive for subsistence and economic development (Culas 2012; Ceddia et al. 2015).

Studies from Asia have highlighted the importance of socioeconomics in influencing the effects of economic and other underlying drivers on deforestation. The differences in population density between urban and rural locations and the choice of agricultural crop had an interaction effect on deforestation in Indonesia (Gatto et al. 2015), and changes in urban structure and local economic development boosted in-migration in Shenzhen, China, which drove urban forest fragmentation (Gong et al. 2013). Population density has been shown to drive forest loss in India (Krishnadas et al. 2018), and in Pakistan, Mannan et al (2019) found that a combination of geographic, socioeconomic, and environmental factors were effective predictors of LUC, whereas Zeb et al (2019) found that household demographics and poverty were underlying factors of forest clearance for livestock and agricultural expansion. In Thailand and Vietnam livelihoods and local economics were highly influential in farmer’s decision-making related to LUC (Nguyen et al. 2017), and in the mountainous regions of Southeast Asia it has been national policies in combination with local economics that governed LUC (Fox & Vogler 2005). The extensive literature on the socioeconomic predictors of deforestation emphasises the breadth and complexity of relationships between local socioeconomic conditions, broader economic factors, the environmental context, and government policy. The scale at which studies are undertaken is revealed to be important, as is local context. The two examples from Pakistan (Mannan et al. 2019; Zeb et al. 2019) are from very similar areas within Pakistan, and conduct their analyses at a similar spatial resolution, yet identify different socioeconomic predictors of forest loss. Therefore, environmental and economic policies that improve socioeconomic conditions for local people without forest loss and environmental degradation will require an understanding of the relationships between socioeconomics and forest cover at different scales. Successfully isolating the signals of these relationships is however, challenging, due to the complexity of social-ecological systems, the non-linear feedback loops, and the heterogeneity in system dynamics at different scales.

Since the early 1990s, after more than a decade of civil war and unrest, Cambodia has experienced extraordinary economic development and social change (Solcomb 2010; Milne & Mahanty 2015). There have been significant improvements in rates of poverty, access to services, and education, yet inequality between the rich and the poor has increased (World Bank 2014). There has been dramatic socioeconomic development within Cambodia’s major urban centres, yet rural, marginalised groups and ethnic minorities, particularly in the remote provinces, largely remain poor with minimal access to services, and where insecure land tenure leaves them exposed to land grabbing and conflict (Hammer 2008; Ironside 2008; Neef & Touch 2012; Phillips & Davy 2021). Such rapid social and economic changes make identifying the drivers of deforestation particularly challenging, as broad-scale drivers and their effect on local landscape actors, and the subsequent proximate causes of deforestation, are likely to fluctuate rapidly over time and space. There are several studies from Cambodia that have focussed on socioeconomic predictors of deforestation, which provide some important context. At the national scale, human population pressure has been identified as an important driver of deforestation (Dasgupta et al. 2005), and in Northwest Cambodia there have been many direct and indirect drivers of deforestation since 1975, including repatriation of Khmer Rouge soldiers and in-migration following the end of the civil conflict, refugee repatriation, the subsequent clearance for subsistence agriculture, and the expansion of cash crops such as cassava (Hought et al. 2012; Kong et al. 2019). In the Angkor Basin, home to the Angkor temples, a complex mix of global, regional, and local drivers including tourism, climate change, government policies, economic development, and environmental management between 1989 and 2005 caused over 23% of the existing forest cover to be lost to agricultural expansion and charcoal production (Gaughan et al. 2009).

Integrated conservation and development projects (ICDPs) that aim to tackle both forest loss and socioeconomic development at the same time have had mixed results (e.g., Geist & Lambin 2003; Chambers et al. 2020; Bernhard et al. 2021). These projects can have unintended consequences for a number of reasons, including poor management of incentives or weak enforcement of protective laws, misinterpretation of stakeholder motivations, or failure to account for underlying economic drivers operating at a broader scale (Chambers et al. 2020). National economic and environmental policies, and interventions such as ICDPs, will be vulnerable to failure if the relationships between forest cover and loss and 1) broad economic drivers and 2) local socioeconomic conditions are not understood and accounted for. In chapter 1 I addressed the relationships between macroeconomic drivers and forest loss at the national scale. Previous studies have evaluated the relationships between forest cover and socioeconomics in small, discrete locations within Cambodia (Dasgupta et al. 2005; Gaughan et al. 2009; Hought et al. 2012; Kong et al. 2019), but to my knowledge, no study has attempted this at a national scale. Therefore, in this chapter I aim to fill this research gap by exploring whether high resolution spatially explicit socioeconomic variables from the across Cambodia can predict forest cover, and whether a non-statistical approach can reveal further patterns between socioeconomics and forest cover. My objectives are therefore to 1) model the relationship between socioeconomic variables and forest cover for the whole country at two spatial resolutions (Province, Commune) using generalised liner mixed models, and 2) use a cluster analysis to create a provincial-level socioeconomic typology which further describes the relationships between socioeconomic development and forest cover.

**Methods**

*Study area*

Please see the section “*Study area*” in Chapter 1 for a detailed description of Cambodia, and for a map of Cambodia within Southeast Asia. There are 24 provinces in Cambodia, each of which is made up of several further administrative layers (Figure 1). Districts are below province, and each district is comprised of multiple communes, with each commune containing multiple villages. The number of communes is not static, with changes in the number of communes between years reflecting shifting administrative boundaries. Between 2007 and 2012 the number of communes ranged from 1,317 (2007) to 1,512 (2012).

*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 x, Dasgupta et al. 2005; Mena et al. 2006; Rowcroft 2008; 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 presence of economic land concessions (Abdullah & Nakagoshi 2007; Davis et al. 2015; Xu et al. 2019), presence of 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 Map

Description automatically generatedthe 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.

**Figure 1. A map of Cambodia with the 24 provinces coloured and numbered, with the smaller communes shown with black lines within each province. The provinces are 1 – Ratanak Kiri, 2 – Stung Treng, 3 – Otdar Meanchey, 4 – Preah Vihear, 5 – Banteay Meanchey, 6 – Siem Reap, 7 – Kampong Thom, 8 – Mondul Kiri, 9 – Kracheh, 10 – Kampong Chhnang, 11 – Pursat, 12 - Kampong Speu, 13 – Prey Veng, 14 – Svay Rieng, 15 – Takeo, 16 – Kampot, 17 – Koh Kong, 18 – Kep, 19 – Preah Sihanouk, 20 – Battambang, 21 – Pailin, 22 – Kampong Cham, 23 – Kandal, 24 – Phnom Penh (capital city).**

*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 (averaged using either mean or median, or summed) to the commune and province level after error checking and cleaning (see Supporting Information for details on aggregation and error checking). 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 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 & 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 from all commune-level predictions within each province was extracted as the provincial median prediction. Confidence intervals for all prediction plots were calculated as 2 × SE (Zuur et al. 2009).

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. Confidence intervals for all prediction plots were calculated as 2 × SE (Zuur et al. 2009).

*Cluster analysis*

Agglomerative clustering was conducted to create a typology for provinces based on the socioeconomic variables in Table 2 (excluding control variables). 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 models and the transformations done for the modelling. Variables with a \* indicate they were included in the cluster 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 (non-Khmer, as defined by the RGC) |
| 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**

*Socioeconomic predictors of forest cover at the Commune level*

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 >0.99. 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.

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.

*Socioeconomic predictors of forest cover at the Province level*

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%). 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 (*n* = 16) of provinces that fell into a single cluster, and so I chose 5 clusters 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). 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 categorised by high or very high values of all variables, which translates to 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 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 2. 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, diagram

Description automatically generated

**Figure 2. Predicted relationships (red lines, blue bars) 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 points are the raw data points. a = population density, centred and scaled, b = mean elevation (masl), centred and scaled, c = distance to international border (KM), centred and scaled, d = distance to provincial capital (KM), centred and scaled, e = presence of economic land concessions, f = presence of protected areas.**

Calendar

Description automatically generated

**Figure 3. 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 4. 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 5 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).**

Map

Description automatically generated

**Figure 6. 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 7. 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 8. 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 3. 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. |

**Discussion**

In this study, I have modelled the relationships between socioeconomic variables and forest cover at multiple scales and have investigated these relationships using two different approaches. This study has revealed some important relationships from which we can make inferences regarding the socioeconomic, geographical, and biophysical predictors of forest cover across Cambodia. Furthermore, I have revealed key methodological issues, particularly around scale and model variance, that are likely to be common in these types of analyses, but which often remain unexplored or unreported in the literature. Understanding the drivers and proximate causes of deforestation is critical for the development of sustainable environmental policies and forest conservation initiatives. Studies need to target multiple scales to build a cohesive picture of the social-ecological systems within which deforestation occurs, so that policy development is appropriate and effective. Importantly, researchers need to select the appropriate method to answer specific questions at specific scales, and the complexity of the system must be understood.

*Socioeconomic predictors of forest cover*

The commune-level model revealed that population density was the only non-control variable with any effect on forest cover, and the effect was weak. I was limited in the socioeconomic variables that were available, and it is possible that the variables selected were simply poor predictors of forest cover. However, the modelling process revealed very large between-commune variation in both predictor and response variables, in addition to many random effect levels (between 1,317 and 1,512). Model predictions from the final model, and from preliminary models, showed that the parameter estimates (intercepts and slopes) for a given socioeconomic variable (see Figure 6 for an example from population density) varied widely from commune to commune, even within the same province. Therefore, it is possible that the difficulty in estimating a single parameter from the surrounding “noise” resulted in the detection of weak, or no effects, rather than a genuine lack of effects. The province-level model was built to counter the issue of excessive between-commune variance by approaching the analysis from a different scale. Two socioeconomic variables remained in the final province-level model but again, the effects were relatively weak. It is still possible that the weak effects represented a genuine lack of correlation between socioeconomics and forest cover, however, modelling the effects at a larger scale will simply mask the large variation that exists at the finer scale, rather than eliminating it. An advantage of GLMMs is the ability to quantify between-group variance, which not only offers crucial insight about the differences between groups (e.g., countries) from which inference can be drawn (Zuur et al. 2009), but can also highlight potential problems with ‘global’ predictions (i.e., predictions that are made with all random effect terms set at their mean). Yet very few studies that use these models for LUC report any values for variance associated with the random (group-level) effects. For example, Bhattari and Hammig (2004) use data from 63 countries to produce a single effect for GDP per capita on deforestation, yet do not report any value for country-level variance. The effect size is relatively small, and therefore if there was large between-country variance then the country-level effects could be vastly different, rendering the single global effect misleading.

*Scale*

This analysis highlights the importance of scale when modelling complex social-ecological systems; researchers must not only select the scale of the analysis carefully but must also be aware of underlying variation which may be affecting estimates, requiring cautious interpretation of results. The results of the socioeconomic analysis have further highlighted the effect of scale on drivers with larger effects. The direction of the effect of distance to an international border changes depending on whether you are looking at the commune-level or the province-level; it was positive within communes and negative within provinces. This reversal of effect direction between scales also occurs for the distance to the provincial capital (positive within communes and negative within provinces). Taken together, the two models can add important nuance to the interpretation of results. Provinces that are close to international borders have higher forest cover, but within those provinces, the communes that are furthest away from the border and the provincial capital are predicted to have the highest forest cover. Provinces furthest away from the major the urban centres of Phnom Penh, Siem Reap, and Battambang tend to be the large, rural provinces that have an international border (e.g., Mondul Kiri, Ratanak Kiri, Stung Treng, Koh Kong) and have high forest cover. My results suggest that within those provinces, the communes furthest from the border, and furthest from the provincial capital, are likely to have higher forest cover. Increases in human population density over time, including from in-migration, often result in agricultural expansion, exploitation of forest resources (e.g., timber), and increased urbanisation, all of which could be reducing forest cover around the provincial capitals. International borders promote the movement of people, commodities, economic activity, and all the associated infrastructure that is required to maintain such activity (ref). When combined with illegal cross-border activities such as logging, land clearance, and the wildlife trade (see Evans et al. 2013), it is plausible that communes closer to the international borders are more likely to have reduced forest cover (Grogan et al. 2015).

The inherent complexity within social-ecological systems results in significant challenges when researchers attempt to model them (Basse et al. 2014). Taking this study as an example, a researcher has a choice between modelling at a large scale (e.g., national, regional) where effects may be weak or unrepresentative of much of the country or region, or modelling at a fine scale where effects may be swamped by variation resulting in the loss of the true signal. By reframing analytical goals and aiming for description of the data, for example using cluster analyses, over statistical hypothesis testing and attempts at explanation, researchers can reduce the need for increasingly complex data and models.

*Cluster analysis*

The purpose of the cluster analysis was to explore an approach that was different to the traditional statistical modelling I had done previously, and to remove the above issues of variance. I was interested to see what patterns would emerge when the underlying goal of statistical hypothesis testing (i.e., the effect of the predictor *x* on response *y* is significantly different from 0) was removed. The cluster analysis revealed interesting patterns of distinct regions, suggesting that in many cases provinces that are adjacent to each other tend to have similar socioeconomic characteristics, resulting in clusters that are comprised of spatially contiguous provinces. The two cluster that generally display the largest differences are clusters 1 and 3. Cluster 1 contains the provinces of Mondul Kiri and Rattank Kiri which are large, remote, and some of the least developed provinces in the country. They are home to the Eastern Plains Landscape which is one of the most important areas in SEA for biodiversity (Chapter 3, Gray et al., 2012; Griffin and Nuttall, 2020; Nuttall et al., 2017). The cluster has the highest forest cover, low population density, low access to services, and low migration. Economic development in the first two decades after the civil war was focused almost entirely on the major cities: Phnom Penh (cluster 3), Sihanoukville (cluster 5), and Battambang (cluster 4), with rural provinces remaining underdeveloped, inaccessible, and poor (Hughes & Un 2011). The lack of infrastructure and access, coupled with low population density and few employment opportunities that limited in-migration, have all likely contributed to forest cover remaining high (Evans et al. 2013). Conversely, cluster 3 has the lowest levels of forest cover and contains the capital city of Phnom Penh and the surrounding provinces which are the hubs for industry and economic activity (such as the garment sector). Cluster 5 is interesting because it contains the largest number of provinces. The expectation was that the provinces that most closely resembled cluster 1 (i.e., large, rural provinces with high forest cover) such as Stung Treng, Preah Vihear, and Koh Kong, would have been clustered either with cluster 1, or within a separate cluster. However, they were clustered with the central belt of provinces (e.g., Kampong Speu, Kampong Chhnang, Kampong Thom) which are almost exclusively low elevation agricultural provinces that are geared towards rice production. The inclusion of Stung Treng, Preah Vihear, and Koh Kong within this cluster and the resulting cluster typologies, suggest that there has been some success in increasing the socioeconomic status of rural, highly forested provinces without excessive loss of forest cover.

The advantage of clustering techniques such as UPGMA is that although there are metrics that can suggest optimal numbers of clusters, the researcher can select the number of clusters that is most useful for their particular investigation (Borcard et al. 2018). Unlike statistical models, cluster analysis does not produce estimates of effect sizes, nor can predictions be made. Nevertheless, by altering the number of clusters, investigating different clustering approaches, followed by considered exploratory analysis and plotting, a comprehensive picture of the study system can be produced. This may be a sensible first step in a larger analysis which can increase understanding of the system before modelling approaches are decided upon. Furthermore, methods such as cluster analysis are conceptually simpler than advanced statistical and mechanistic modelling, making interpretation and explanation, simpler.

*Implications / broader perspectives of the results*

The results of this study have highlighted that the regions of Cambodia that have the highest forest cover also tend to be the rural, remote, poor provinces with high proportions of indigenous people. It is people living within these areas that will be reliant on natural resources and forest products for their subsistence and livelihoods. In these circumstances, the efforts of an individual actor to increase their socioeconomic status is likely to include agricultural expansion, resulting in forest loss. Therefore, to avoid forests being the price of socioeconomic development, national and sub-national government need to develop economic policy frameworks that deliver economic benefits whilst encouraging forest protection, such as payments for ecosystem services schemes, and support for agricultural improvement technologies and diversification (Eliste & Zorya 2015). The Cambodian government, however, has shown enthusiasm in the past for economic development via private land leases for industrial-scale commercial agriculture (Chapter 1), many of which have been awarded in rural, remote, forested land. These economic land concessions (ELCs) have frequently been placed on traditional lands of indigenous people, on the lands of the rural poor who had yet to be awarded legal land titles, and in areas of high forest cover, including protected areas (Vrieze & Kuch 2012; Global Witness 2013; Davis et al. 2015). Remote provinces with a low density of relatively poor inhabitants, low levels of land tenure security, and plentiful forests, are particularly vulnerable to the allocation of new ELCs, particularly if the economic policies of the last decade are pursued. Despite government claims, ELCs often contribute very little to local economies, and are sources of land conflict, illegal settlement, and extensive, unregulated, and often illegal deforestation (Vrieze & Kuch 2012; Global Witness 2013; Watson et al. 2014; Davis et al. 2015; Milne & Mahanty 2015). There has been a reduction in new ELC allocations in recent years (www.opendevelopmentcambodia.net), which may suggest that new avenues for economic development and growth in the agriculture sector are being developed.

Since the end of civil conflict in the early 1990s, there has been significant migration and resettlement into rural provinces as people move back into traditional homelands or move in search of new land to settle (Milne & Mahanty 2015). This post-war migration has come at a cost to forest cover, as families look to establish and expand their agricultural land (Hought et al. 2012; Kong et al. 2019). The rural provinces with high forest cover are still vulnerable to in-migration and land speculation, as access has improved significantly over the last decade and poor, landless families seek to establish themselves in frontier areas (Evans et al. 2013). This, and other studies, have demonstrated that increases in human population density can predict forest loss (Dasgupta et al. 2005; Krishnadas et al. 2018). In the context of poor environmental governance and weak institutions, as in Cambodia (Riggs et al. 2018), rural in-migration could continue to drive forest loss. Government settlement and land titling policies need to pre-empt increased migration into rural areas, particularly those with protected areas, and ensure forest loss is minimised. The two most prominent settlement initiatives – social land concessions and Directive 01 – have been widely criticised for poor management and implementation, and both have resulted in the loss of forests inside protected areas (Milne 2013; Oldenburg & Neef 2014; Grimsditch & Schoenberger 2015, also see Annex 1).

The cluster analysis placed the provinces Preah Vihear, Stung Treng, and Koh Kong into cluster 5, suggesting that these provinces have socioeconomic conditions similar to the wealthier, more developed provinces such as Siem Reap, and to provinces with extensive agriculture such as Kampong Chhnang. This placement was despite many similarities with the provinces in cluster 1, including being large, rural, and with high forest cover. This clustering suggests that Preah Vihear, Stung Treng, and Koh Kong have made progress in increasing the socioeconomic conditions of the population, without extensive forest loss (median forest loss within cluster 5 is less than for cluster 1). These improvements may have been driven by the expansion of economic sectors that do not rely on natural resource extraction or agricultural expansion, or indeed by the growth of nature-friendly sectors, such as ecotourism. These results warrant further investigation to identify whether there are specific policies, initiatives, economic conditions, or social movements that have improved socioeconomic conditions with minimal deforestation, and which could be replicated in other poor, forested provinces.