Supplementary Material A: Data processing

Towards a global behavioural model of anthropogenic fire: The spatiotemporal distribution of land-fire systems

This appendix covers processing of secondary data undertaken to support modelling and findings presented in Perkins et al., (2022). It covers rescaling, extrapolation, sampling and smoothing of secondary data sets.

1. Rescaling of data

The ultimate goal of the land-fire system distribution presented in Perkins et al., (2021) is to develop a model of human fire impacts that may be coupled with the JULES-INFERNO DGVM. JULES-INFERNO runs at a resolution of 1.25^{o} x 1.875^{o}. Therefore, all secondary data sets employed in our model were re-scaled to this (coarse) resolution. This was done in R using the 'raster' package version 3.3.13 (Hijmans 2020) using bilinear interpolation.

2. Extrapolating data sets

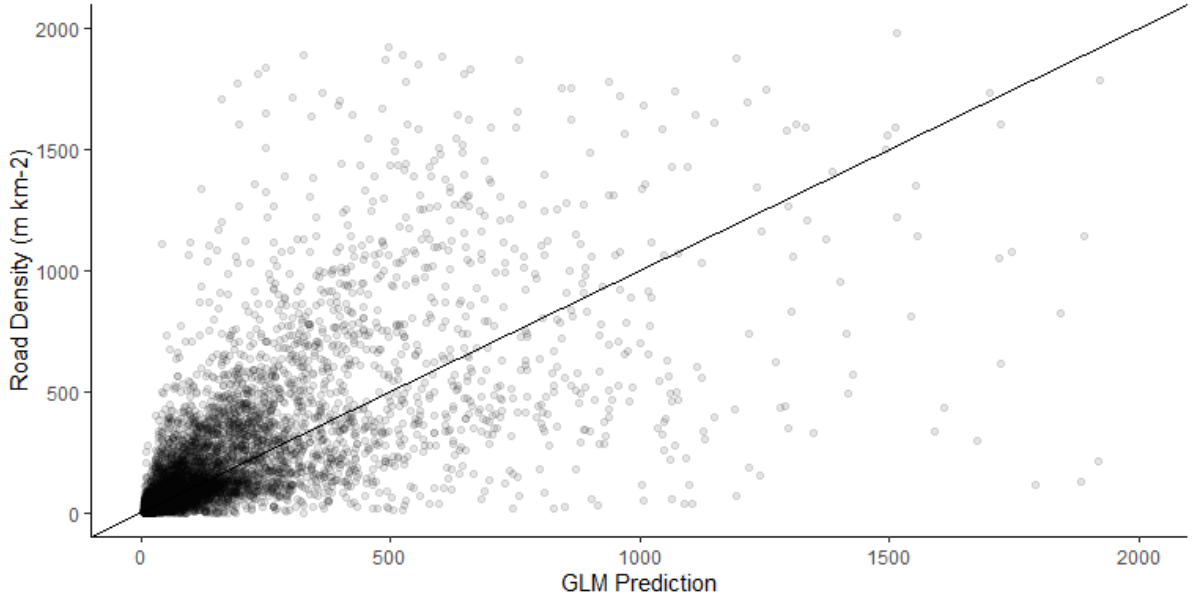
Data sets used came with a varying level of temporal coverage over the period of model runs (1990-2014; see Perkins et al., 2022, Table 1 for a complete list). Missing years in data sets were handled in two ways. Data which contained direct measurements (for example population density) which did not have full coverage across the study period were extrapolated using a simple last observation carried forward or first observation carried backwards approach. By contrast, market access data were not directly measured but were themselves compiled from calculations based on secondary data. Therefore, as these were only available for one study year, they were extrapolated across years study years using a generalised linear model.

2.1 Extrapolation of market access data

Market access data (Verburg et al., 2011), (which describes the travel time to the nearest city or port on a 0-1 scale), were found to be valuable as a driver of the distribution of land-fire systems (LFS). However, they were only available for the year 2000. Furthermore, as the original data were themselves derived from secondary data - the location of ports and cities and road density – it should be possible to extrapolate the measure across all study years. This was done using a generalised linear model (glm) using the following steps.

Given the importance of travel times in calculating the accessibility of the nearest city or port from a given location, the first step was to find predictor variables to capture this aspect of market access. A number of methods were tried, including use of the 'accessibility' and friction layers for 2015 developed by the Global Malaria Project (Weiss et al., 2018). This was projected across 1990-2014 by modelling the friction layer as a generalised linear model, and using this extrapolated temporal variable to calculate the least cost path from each grid cell to the nearest city or port. However, perhaps due to the complexity of this calculation, the extrapolated accessibility and friction layers were found not to be predictive of the 2000 market access data. The eventual adopted approach, therefore, was to use the global road density data set of Meijer et al., (2018) as a proxy for travel times across a given grid cell. These road density data were extrapolated across the study period using a generalised linear model, with GDP, HDI and the natural logarithm of population density as predictor variables (Figure A). The underlying model achieved a pseudo r^2 of 0.71.

A)



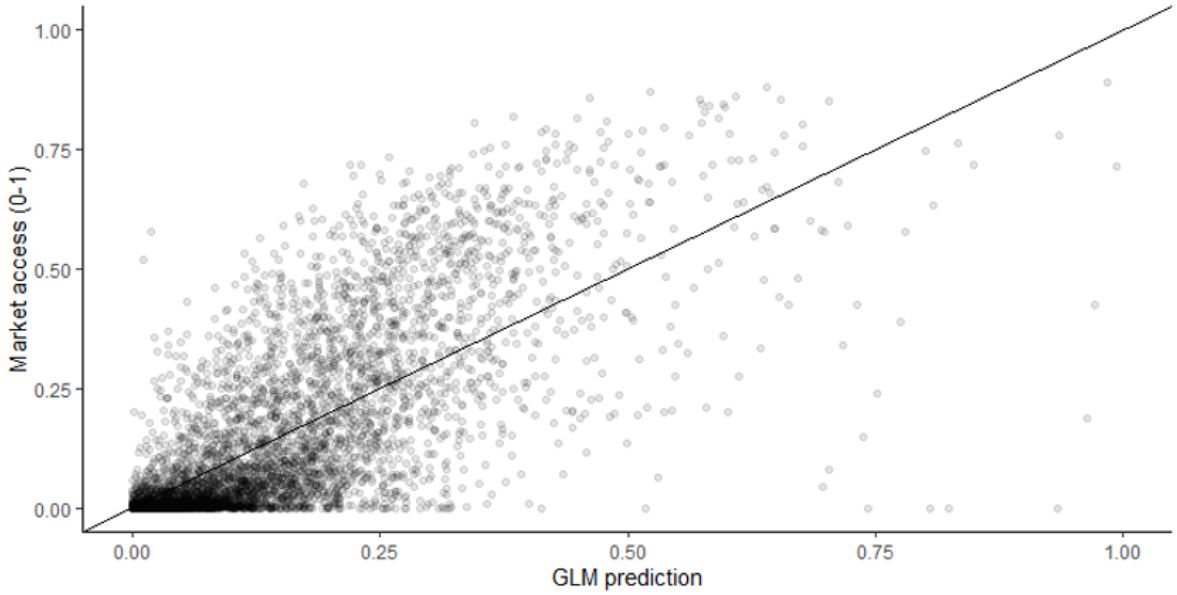
B)

Variable	Coefficient
Intercept	0.7709
Log(Population)	0.1013
GDP	-1.689 e-06
HDI	0.6874

Figure A: Overview of generalised linear model used to predict global road density: A) Predictions against original data set & B) model coefficients. The model achieves good predictive accuracy (pseudo $r^2 = 0.71$), but tends to under-predict the variance in the response variable (standard deviation: 604.8, data vs 306.3, modelled). The glm used a gaussian response variable with a logarithmic link. The original data were for 2015 (Weijer et al., 2018).

Along with this extrapolated road density layer, the glm of market access also used the logarithm of population density and the un-extrapolated accessibility layer of Weiss et al., (2018). The resulting model achieved a pseudo r^2 of 0.73 (Figure C). Whilst the use of the static accessibility layer of Weiss added bias to the model, the glm still achieved good predictive accuracy. Ultimately, the use of extrapolated market access data is justified in its empirical performance: the AUC of classification trees that used the market access or related market influence variables increased by an average of 0.01 when the extrapolated data was used. Furthermore, for the planned global model of human impacts on fire to be run into the future data will need to be able to be projected forwards. Our work suggests this is feasible for market access, which underpins its utility as a predictor variable of human fire use.

A)



B)

Variable	Coefficient
Intercept	-3.396
Population	7.998e-05
Log(Road density) +1	0.3383
Accessibility	-4.809e-03

Figure B: Generalised linear model used to predict global market access: A) Predictions against original data set & B) model coefficients. The model achieves good predictive accuracy (pseudo $r^2 = 0.73$), but underestimates market access at low to moderate levels (particularly 0.25-0.5). The glm used a gaussian response variable with a logarithmic link. The original data were for 2000 (Verburg et al., 2011).

3. Calculation of variable convolutions and other derivatives data sets

Three derivative predictor variables were calculated based on the original secondary data sets employed in the study. These were:

- HDI*log(GDP)

During the construction of classification tree models, GDP and HDI were found consistently to be chosen as the first split in a tree structure in approximately 50% of a bootstrapped ensemble. It was found that the product of HDI and the natural logarithm of GDP was chosen preferentially in place of GDP and HDI in the majority of such instances. This may be because HDI captures information most effectively at low GDP, where economic data may be distorted by a few vary large salaries, whilst GDP is more effective at capturing information in more developed contexts.

- Terrain Roughness Index

The Terrain Roughness Index (TRI; Riley et al., 1999) is a measure of the variance in topography. It was calculated using using the spatialEco package in R version 1.3.7 (Evans 2020).

- *Wealthy flat index*

Similar to the case of GDP and HDI, some LFS classification trees were split approximately evenly between (low or flat) topography and (high) GDP as the first split measure – primarily for intensive land uses. Therefore, a combined variable was created to capture both these concerns, calculated as: GDP x 1/TRI. A high TRI represents very rugged terrain, so this index is highest in areas of high GDP and flat terrain. We term this the 'Wealthy flat index'.

4. Processing of HANPP data

Data for the human appropriation of net primary productivity (Haberl et al., 2007; Kastner et al., in review) were available at 5 arcminute resolution. Therefore, these were resampled to the resolution of JULES-INFERNO, as described above.

5. Use of data for modelling

5.1 Data sampling

In order to train the classification tree models that drive our LFS distribution, we sampled the secondary data sets at the locations of case studies in DAFI. This was done using the central point of a DAFI case study area as the sampling location. The year sampled was the mean of the study period, rounding upwards – so a study beginning in 2002 and ending in 2005 would be allocated the values from secondary data sets for 2004.

5.2 Data smoothing

During model runs, it was found that interannual variability in biophysical variables caused some cells on the boundary between LFS to oscillate between two states. For example, between intensive farming and small-holder cropping based on fluctuations of reference evapotranspiration. Therefore, a 10-year average was calculated from the data, comprised of the model year (t) and the previous 9 model years. This removed the oscillation issue. The impact of imposing a moving window on socio-economic variables was also explored, but not found to change model outputs significantly.

References

Evans, J. (2020). sptialEco. version 1.3-4. https://github.com/jeffreyevans/spatialEco

Hijmans, R. (2020). raster: Geographic Data Analysis and Modeling. R package version 3.3-13. https://CRAN.R-project.org/package=raster

Perkins, O., Matej, S., Erb, K-H., & Millington, J. (2022). Towards a global behavioural model of anthropogenic fire: The spatio-temporal distribution of land-fire systems. Socio-Environmental Systems Modelling, vol. 4, 18130. doi:10.18174/sesmo.18130

Riley, S., DeGloria, S., & Elliot, R. (1999). A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity. Intermountain Journal of Sciences, 5 (1-4), 23-27.

Verburg, P., Eliis, E., & Letourneau, A. (2011). A global assessment of market accessibility and market influence for global environmental change studies. Environmental Research Letters, 6 (3), 0304019. https://doi.org/10.1088/1748-9326/6/3/034019

Weijer, J., Huijbregts, M., Schotte, K., & Schipper, A. (2018). Global patterns of current and future road infrastructure. Environmental Research Letters, 13 (6), 064006. https://doi.org/10.1088/1748-9326/aabd42

Weiss, D., Nelson, A., Gibson, H… & Gething, P. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature, 553: 333–336. https://doi.org/ 10.1038/nature25181