Appendix 1. Methodological Details and Rationales.

This appendix is dedicated to expand on the details, rationales and performance evaluation of the methodology followed through this study. The sections are organized following the methodological steps listed in Fig. A1.1, however the actual work implied numerous feedback loops and reiterations of previous steps which are omitted for simplicity.

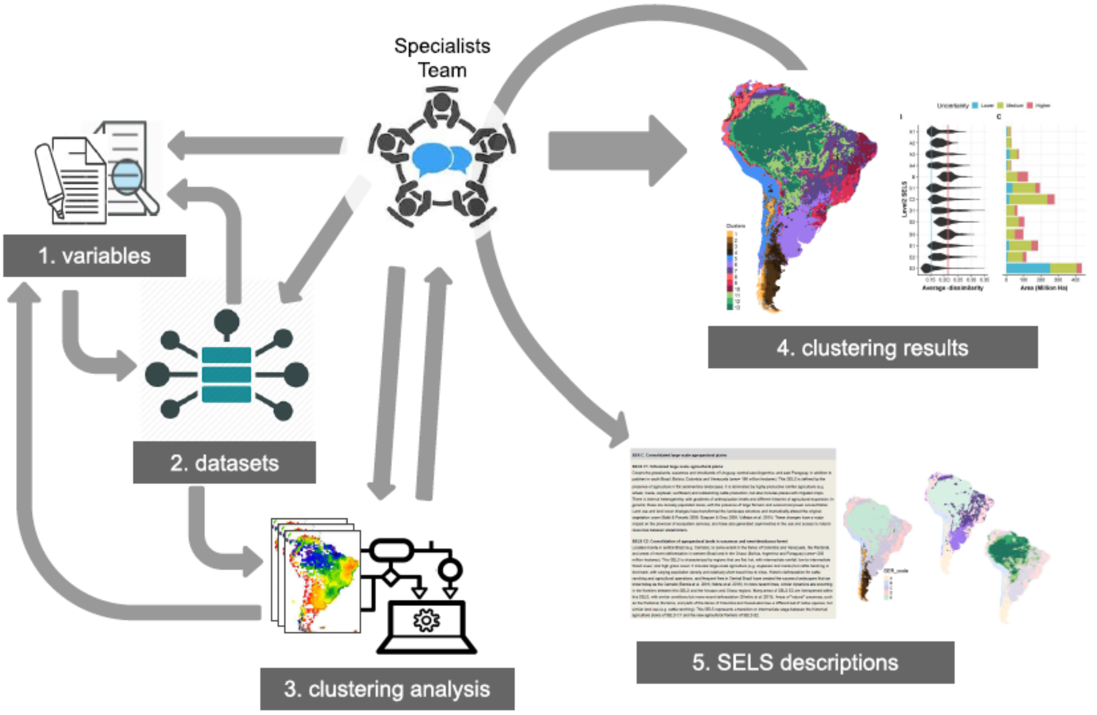


Figure A1.1. Diagram of the methodological steps followed by this study. (1) Variables: analyze and systematize the conceptual SELS descriptions in Table 1 of Boillat et al. (2017) defining a list of variables to use as inputs for the clustering. (2) Datasets: search and retrieve the spatial data to best represent the selected variables. (3) Clustering analysis: generate automated classifications through hierarchical cluster analysis. (4) Clustering results: analyze the clustering outputs and agree on the SELS representation according to the specialists group's territorial knowledge. (5) SELS descriptions: arrange in subgroups of regional specialists to discuss and describe each particular SELS. Arrows pointing backwards in relation to the numerical steps represent the feedback loops and local iterations of our process.

1. Variables

We used as a reference the biome-level SELS typologies described in Table 1 of Boillat et al. (2017; hereafter conceptual SELS) to guide the variable selection process. Such descriptions were in a narrative form with no shared standard structure. We first exhaustively analyzed the conceptual SELS descriptions and listed all attributes mentioned for each of them. We then synthesized the list of attributes into general variables that represented the data we needed to acquire in order to capture those properties. The product of this process was our ideal input data list including 25 general variables (Table A1.1), from which we discarded and added variables through a heavily iterative process connected with step (2) Datasets.

On one hand we had to discard all general variables lacking a dataset that was adequate (representative proxy) and spatially continuous (covering the whole continental extent) with a coherent methodology. On the other hand, we discarded all variables that referred to trends, since combining measures of state and trajectories raised concerns among the authors about methodological philosophical inconsistencies.

Finally, to visualize whether our data was balanced across different aspects of the social-ecological systems we arranged the general variables within broader dimensions following the framework of Winkler et al. (2018). Compared to other popular frameworks, such as Ostrom's framework for analyzing sustainability of social-ecological systems (Ostrom 2009) ideal for addressing specific issues, the Winkler's framework has a more general scope, which fits better the continental-scale broad multifaceted typologies of our study. We considered all Level III Winkler categories except *Health*, due lack of data. We recognized underrepresented dimensions in our original list of 25 general variables (Table A1.1), such as the *Physical* dimension, mentioned in the conceptual SELS names yet not in their descriptions; the *Political* dimension, which was indirectly suggested but not explicitly addressed; or infrastructural aspects of the *Economic* dimension. To complement and balance the representation of all different dimensions we incorporated the following variables: *Flat relief, Temperature, Precipitation, Irrigation, Cities traveltime, Ports traveltime,* and *Governance indicators*.

Table A1.1. Systematization process of the conceptual SELS's descriptions: synthesis of all mentioned attributes into general variables.

General Variable	Description in Boillat et al., 2017	Included in this study
Natural land cover	SAL: forested areas; DML: semi-arid shrublands	Yes
Rate of land use change	SAL: relatively rapid rate of land use change; high rate of deforestation	No (trends)
Change in cropland cover	SAL: expansion of agricultural frontiers; STFD: decreasing agriculture	No (trends)
Change in livestock	SAL: expansion of cattle ranching; STFD: decreasing livestock; DML:extensive livestock grazing	No (trends)
Main livestock Species	DML: particularly goats	No (concerns on representation of informal livestock on datasets)
Crop exports	SAL: commodity markets driving LUC; CAL: some areas have shifted to export-oriented agriculture	No (national level statistics)
Ecosystem Degradation	SAL: forest degradation due logging; DML: extensive degradation due capital-intensive land use and extensive cattle ranching; CAL: highly degraded and threatened natural ecosystems ('lomas costeras', dry tropical forests, wetlands) and Important biomes such as Brazil's Atlantic forest have become highly fragmented	No (unclear definition/lack of data)
Biodiversity loss	SAL: biodiversity loss	No (IUCN data not spatial)
Carbon emissions	SAL: high carbon emissions	No (unclear impact)
Protected areas	SAL: expansion of protected areas; STFD: extensive formal conservation	Yes

Cultural diversity	SAL: expansion of indigenous areas; SAHA: high cultural diversity	Yes
Endemisms	SAHA: high endemisms; DEM: high species endemisms	No (lack of spatial data)
Size of production units	SAL: shifting to larger management/production units in some areas; SAPL: expansion of sizes of agricultural and livestock farms, large scale land acquisitions in recent years; SAHA: small subsistence-oriented management units; CAL: large scale land acquisitions for tourism and other developments	No (lack of data covering the full continent)
Land use diversity	SAL: other areas with diversity of land systems; SAHA: high diversity of landscapes, limited mechanized agriculture and relatively high levels of biodiversity within anthropogenic landscapes; CAL: with mixed land and forest usages; SAPL: agribusiness surrounding indigenous and conservation areas in the Cerrado; DML: dominated by irrigated agriculture within matrix of semi-arid shrublands	Yes
Crop diversity	SAHA: livelihood diversification; high agro-ecological diverstity	Yes
Environmetal conflicts	SAL: new land uses in conflict with local and indigenous communities	No (lack of data)
Type of urbanization	SAL: chaotic urbanization and peri urban expansion; DML: various degrees of urbanization; CAL: home to high population densities	Yes
Historical land use	SAPL: long history of cropland and ranching settlements; SAHA: most landscapes with long history of human settlement; CAL: long history of human occupation	Yes
Migration rates	SAHA: eleveated rates of rural out-migration	No (lack of spatial data)

Political/economical relevance	SAL: enhanced contributions to national economic growth and food security; SAHA: may become peripheral as political power and people move to the lowlands; CAL: concentration of political and economic power	Yes
Change in agriculture yields	SAL: dramatic increases in ag productivity	No (trends)
Main crop types	SAPL: high tech agribusiness (soybeans, maize and other grains and fiber); DML: high capital crops (vineyards, olives, fruit orchards); CAL: traditional tropical crops (sugar cane and coffee) & expanding crops (oil palm and eucalyptus)	No (difficulties in creating the metric)
Plantations	STFD: growing forestry plantations (exotic conifers), low agriculture value; DML: high capital crops (vineyards, olives, fruit orchards); CAL: traditional tropical crops (sugar cane and coffee) & expanding crops (oil palm and eucalyptus);	Yes
Mining	SAHA: opened up to new wave of mining	Yes
Tourism	STFD: growing tourism; SAHA: opened up to new wave of tourism activities	No (lack of data)

The first column indicates the general variable we associated with the descriptions of column 2. The second column contains direct transcripts of all the descriptions on Table 1 of Boillat et al. (2017) sorted by the general variable we associated it with. The third column indicates whether the general variable was included in this study and the reason in case of not. Acronyms refer to the SELS typologies by Boillat et al. (2017): SAHA - South American Highlands and Altiplano; CAL - Coastal Agricultural Lands with long colonization history; DML - Dry and Mediterranean Lands; SAL - South American Lowlands: new agropastoral areas; SAPL - South American Plateau Lowlands: agropastoral historical areas; STFD - Southern Temperate Forests and Drylands.

2. Datasets

To be used in this study, all spatial datasets were required to cover the full extent of the South American continent (dismissing islands) with a consistent methodology, in addition we preferred those closer to the year 2010 and a spatial resolution not greater than our grid size (exceptions are the *governance indicators* which are at the national scale, and *plant diversity* at 110km pixels). Country level data, as well as biomes and ecoregions, were allegedly discarded since they imply an artificial homogenization of the territory within arbitrary boundaries which may impact on the spatial representation of the SELS by misleading them to resemble those boundaries. It was a decision taken by the group of authors to avoid using country resolution data for all our variables except those representing political aspects.

We tested for correlations, considering correlation coefficient of |0.75| (absolute value) as the maximum accepted correlation for two variables in the model (Fig. A1.2). We selected Spearman's rank correlation coefficient due it is non-parametric, assesses monotonic relationships, and poses less strict data requirements than Pearson's method (e.g. normal distribution or linear relationships).

The final list of input variables for our analyses consisted in 3 physical, 2 biological, 6 landscape, 7 economic (includes infrastructure), 2 demographic, 4 political, and 2 cultural variables; 11 of which corresponds to the biophysical domain and 15 to the socio-economic domain (Table 1). Most of the variables are non-normally distributed (Fig A1.3), the implications of this on the results are addressed in the next section. Below we expand on the details of calculation of hexagon values for all input variables.

Flat relief: Proportion of the hexagon covered by non-mountain classes in Karagulle et al. (2017) landforms classification. In this classification the mountain classes are four: high mountains, scattered high mountains, low mountains, and scattered low mountains. We chose this variable due it performs better than others in recognizing mountainous terrains embedded in other terrain types (Sayre et al. 2018).

Temperature: Hexagon median of mean annual temperature based on the climate maps generated by ClimateSA. ClimateSA data averages the climatic conditions between 1981 and 2010.

Precipitation: Hexagon median of mean annual rainfall based on the climate maps generated by ClimateSA. ClimateSA data averages the climatic conditions between 1981 and 2010.

Plant diversity: Vascular plant species richness based on the Kreft and Jetz (2007) global patterns of vascular plant species richness calculated with the ordinary co-kriging method. We consider Plant biodiversity as a proxy of overall biodiversity since diversity of different taxa such as mammals, birds, plants, reptiles and amphibia were found to be correlated regardless of environmental conditions (Qian and Ricklefs 2008) and vegetation heterogeneity has shown to be a strong predictor of species richness (Qian and Ricklefs 2008, Stein et al. 2014).

Protected areas: Percent of the hexagon covered by protected areas, considering all categories of protection in the World Database on Protected Areas by UNEP-WCMC and IUCN. The data was downloaded in May 2019 and there is no information to sort protected areas created after our year of reference 2010. Although not the ideal situation, we consider the potential error is acceptable for the purpose of this study.

Land cover: Percent of the hexagon covered by each of the considered classes (i.e. forest, shrublands, grasslands, crops and plantations) based on Graesser et al. (2015) annual land cover classification for South America. To represent our reference year we used the average land cover between 2009 and 2011.

Cover diversity: The land cover diversity of each hexagon was calculated as the shannon diversity index of the area covered by each of the nine land cover classes included in Graesser et al. (2015). To represent our reference year we used the average land cover between 2009 and 2011.

Centrality: This variable is a proxy of the hexagon share of the country's economy, indicating the economic relevance of a particular region to the country. It was calculated by distributing the national gross domestic product (GDP) over the country's territory following the relative distribution of nighttime lights (NTL). The value for each hexagon was calculated as the national GDP * hexagon sum NTL/national sum NTL. For hexagons that overlays with more than one country we consider it part of the one with major area. National 2012 GDP data was obtained from the World Bank database, and 2012 nighttime lights map from the NASA Earth Observatory.

Cattle density: Total cattle production by hexagon according to the gridded Livestock of the World 2.0 by Livestock Geowiki (Robinson et al. 2014) available to download from https://livestock.geo-wiki.org/home-2/.

Mine sites density: Number of mine sites by hexagon considering all categories of mines in the Mineral Resources Data System (MRDS) for the year 2011 available to download from https://mrdata.usgs.gov/mrds/.

Crop diversity: Shannon diversity of the area covered by all different crops in the hexagon based on the 175 crop types by Monfreda et al. 2008.

Irrigation: Percent of the hexagon equipped for irrigation based on the layer "gmia_v5_aei_pct_cellarea" of the Global Map of Irrigation Areas (GMIA) by FAO AQUASTAT (Siebert et al. 2005).

Cities travel time: Mean of travel time in hours to the nearest city of 50,000 or more people (Nelson et al. 2008).

Ports travel time: Mean of travel time in hours to the nearest port. The map was produced for this study following the methodology of Weiss et al. (2018). The road network data was downloaded from the Global Accessibility Map project repository (https://forobs.jrc.ec.europa.eu/products/gam/). We considered all sea ports and inland ports on rivers included in the Río de la plata and Amazonas basins. Ports locations were obtained from Natural Earth (https://www.naturalearthdata.com/) and Ports.com accessed in February of 2018. The distance to ports map together with a detailed explanation of its development (including input data and the reproducible script), are available to download through this link. https://github.com/luciazarba/SELS-SA.

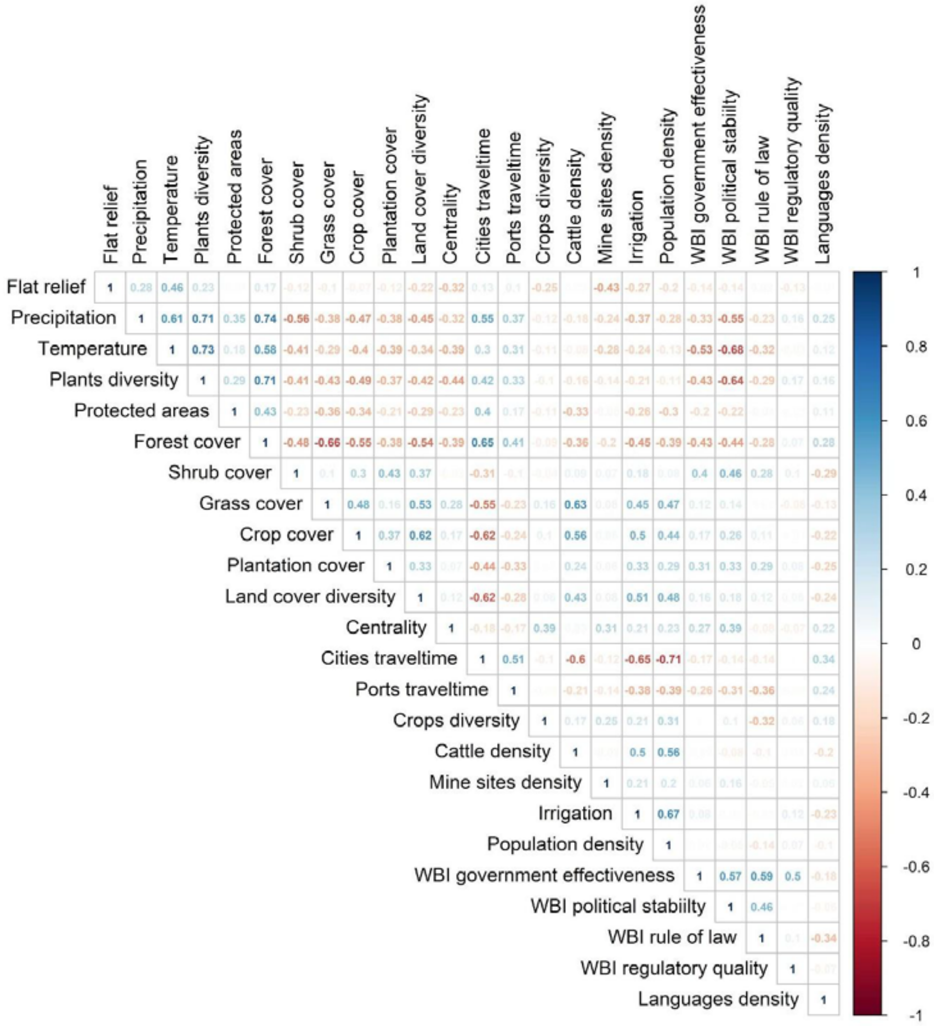
Population density: Mean environmental population by hexagon, based on the Landscan environmental population for the year 2012 (Bright et al. 2012).

Urbanization type: Category of biggest city in a 100 km buffer zone. Cities categories were: rural (no cities within the buffer zone), small city (less than 100,000 inhabitants), medium city (less than 1,000,000 inhabitants), big city (less than 10,000,000 inhabitants), and metropolis (more than $1x10^7$ inhabitants). Cities' data was downloaded from the Global Accessibility Map project repository (https://forobs.jrc.ec.europa.eu/products/gam/).

WBI governance indicators: Country values of the Worldwide Governance Indicators by the World Bank: Voice and Accountability, Political Stability and Absence of Violence, Government Effectiveness, Regulatory Quality, Rule of Law, Control of Corruption. Data was downloaded for the year 2010 from the World Bank website (https://databank.worldbank.org/source/worldwide-governance-indicators). Political Stability and Absence of Violence and Control of Corruption were eventually discarded due high correlation with other variables. In the model, the four political variables included as inputs were weighted down to 0.25 to minimize enforcing national boundaries. In this way the four political variables all together weigh as much as one of the variables in the other domains.

Languages density: Number of different languages spoken within a 100 km buffer zone around each hexagon. The map of language distributions for South America was kindly provided by Mutur Zikin (Zikin 2007), and it was georeferenced and vectorized by the authors of this publication.

Anthropization century: The earliest century in which a 30% of the hexagon was covered by anthropic land cover classes based on Ellis et al. (2010) classification. It consists of anthrome classification maps for each century from 1700 to 2000. We considered as anthropic all classes except for water, remote croplands, remote rangelands, remote woodlands, wild woodlands, and wild treeless and barren lands.



	Flat relief	Precipitation	Temperature	Plants diversity	Protected areas	Forest cover	Shrub cover	Grass cover	Crop cover	Plantation cover	Land cover diversity	Centrality	Cities traveltime	Ports traveltime	Crops diversity	Cattle density	Mine sites density	Irrigation	Population density	WBI government effectiveness	WBI political stabiilty	WBI rule of law	WBI regulatory quality	Languages density
Flat relief	1	0.28	0.46	0.23	[illegible]	0.17	-0.12	-0.1	-0.07	-0.12	-0.22	-0.32	0.13	0.1	-0.25	[illegible]	-0.43	-0.27	-0.2	-0.14	-0.14	[illegible]	-0.13	[illegible]
Precipitation		1	0.61	0.71	0.35	0.74	-0.56	-0.38	-0.47	-0.38	-0.45	-0.32	0.55	0.37	-0.12	-0.18	-0.24	-0.37	-0.28	-0.33	-0.55	-0.23	0.16	0.25
Temperature			1	0.73	0.18	0.58	-0.41	-0.29	-0.4	-0.39	-0.34	-0.39	0.3	0.31	-0.11	-0.08	-0.28	-0.24	-0.13	-0.53	-0.68	-0.32	[illegible]	0.12
Plants diversity				1	0.29	0.71	-0.41	-0.43	-0.49	-0.37	-0.42	-0.44	0.42	0.33	-0.1	-0.16	-0.14	-0.21	-0.11	-0.43	-0.64	-0.29	0.17	0.16
Protected areas					1	0.43	-0.23	-0.36	-0.34	-0.21	-0.29	-0.23	0.4	0.17	-0.11	-0.33	[illegible]	-0.26	-0.3	-0.2	-0.22	[illegible]	[illegible]	0.11
Forest cover						1	-0.48	-0.66	-0.55	-0.38	-0.54	-0.39	0.65	0.41	-0.09	-0.36	-0.2	-0.45	-0.39	-0.43	-0.44	-0.28	0.07	0.28
Shrub cover							1	0.1	0.3	0.43	0.37	[illegible]	-0.31	-0.1	-0.04	0.09	0.07	0.18	0.08	0.4	0.46	0.28	0.1	-0.29
Grass cover								1	0.48	0.16	0.53	0.28	-0.55	-0.23	0.16	0.63	0.08	0.45	0.47	0.12	0.14	[illegible]	-0.08	-0.13
Crop cover									1	0.37	0.62	0.17	-0.62	-0.24	0.1	0.56	[illegible]	0.5	0.44	0.17	0.26	0.11	[illegible]	-0.22
Plantation cover										1	0.33	0.07	-0.44	-0.33	[illegible]	0.24	0.06	0.33	0.29	0.31	0.33	0.29	0.08	-0.25
Land cover diversity											1	0.12	-0.62	-0.28	0.06	0.43	0.08	0.51	0.48	0.16	0.18	0.12	0.06	-0.24
Centrality												1	-0.18	-0.17	0.39	[illegible]	0.31	0.21	0.23	0.27	0.39	-0.08	-0.07	0.22
Cities traveltime													1	0.51	-0.1	-0.6	-0.12	-0.65	-0.71	-0.17	-0.14	-0.14		0.34
Ports traveltime														1	[illegible]	-0.21	-0.14	-0.38	-0.39	-0.26	-0.31	-0.36	[illegible]	0.24
Crops diversity															1	0.17	0.25	0.21	0.31		0.1	-0.32	0.06	0.18
Cattle density																1	[illegible]	0.5	0.56	[illegible]	-0.08	-0.1	[illegible]	-0.2
Mine sites density																	1	0.21	0.2	0.06	0.16	-0.05	[illegible]	0.05
Irrigation																		1	0.67	0.08	[illegible]	[illegible]	0.12	-0.23
Population density																			1	[illegible]	-0.06	-0.14	0.07	-0.1
WBI government effectiveness																				1	0.57	0.59	0.5	-0.18
WBI political stabiilty																					1	0.46	[illegible]	-0.08
WBI rule of law																						1	0.1	-0.34
WBI regulatory quality																							1	-0.07
Languages density																								1

Figure A1.2. Variables correlation matrix. Spearman correlation between all 24 numeric variables considered for the analysis. Positive correlations are in blue, negative correlations are in red, and the strength of the color reflects the strength of the correlation (white color corresponds to correlation coefficients close to zero, therefore not relevant for this purpose).

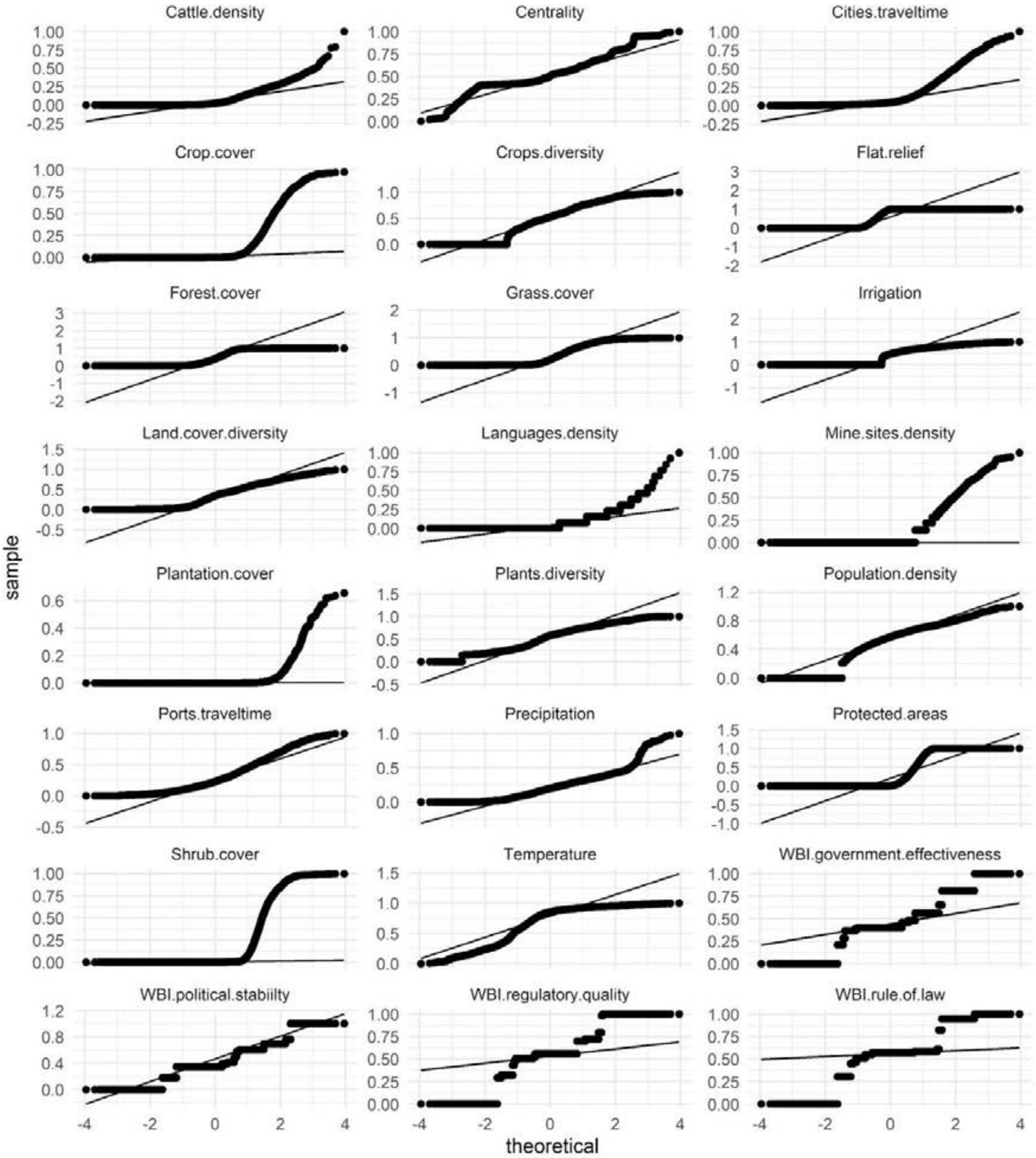


Figure A1.3. Variables Quantile-Quantile plots. These plots allow us to visualize the deviation of each continuous variable from a theoretical normal distribution. If the values (thick dots) lie along the thinner line the distribution has the same shape as the theoretical normal distribution.

3. Clustering analysis.

We analyzed 26 variables across a grid of 13287 hexagonal cells (40 km side to side, area ~1,400 km^2) covering the entire continent of South America in order to identify general typologies of social-ecological land systems (SELS). The process required to calculate the statistical distances between all pairs of hexagons along the multidimensional space and arrange them into groups based on such distances. All calculations were performed in R statistical software (R Core Team 2019) and the scripts are available through this link. https://github.com/luciazarba/SELS-SA.

Statistical distance

Two of our input variables were ordinal: *urbanization type* and *anthropization century*, which represented a major constraint due most distance calculation algorithms only accept continuous data. We followed the Gower distance method (Gower 1971) since it is the recommended algorithm for mixed data (Kassambara 2017, Boehmke and Greenwell 2019). As calculated in R with the *daisy* function (*cluster* package, Maecheler et al. 2019) the dissimilarity between two rows is computed as the weighted mean of the contributions of each variable. Contributions for numeric variables are defined as the absolute difference of both values, divided by the total range of that variable. For ordinal variables' the contribution calculation function applies "standard scoring" (replacement of the variable's levels by their integer codes); similar to using their ranks but avoiding ties.

Several of our input variables did not follow a normal distribution (Fig. A1.3). Despite many data analysis algorithms require specific data distributions, the reference literature for gower (Gower 1971) and DIANA (Kaufman and Rousseeuw 1990) algorithms do not mention particular requirements or considerations regarding data distributions. We found in more recent literature that the Gower distance algorithm is the appropriate metric when clustering non-normally distributed data (Kassambara 2017, Boehmke and Greenwell 2019) since it is less sensitive to outliers and non-normal distributions than other popular methods like Euclidean distances (Boehmke and Greenwell 2019). Furthermore, searching through the gray literature we found a very interesting statement in a scholarly blog discussing the applicability of normality tests for machine learning techniques. One user pointed out that he/she was not aware of any clustering method that assumes normality, and that the cluster-structured data implies a multimodal (and thus non-normal) distribution (Cross Validated blog entry "How to Cluster with Non-normal data" https://stats.stackexchange.com/questions/373404/how-to-cluster-with-non-normal-data).

To account for potential issues with non-normally distributed data we deliberately used the Gower distance metric. Nevertheless, to mitigate the effect of data artifacts on the distance calculations we applied logarithmic transformation to those variables that presented highly exponential distributions (Table 1), and min-max standardization to all variables (forcing them to range between 0 and 1) to avoid unequal impact of variables on the distance measures due their different scales of values.

Clustering Method

We decided *a priori,* based on conceptual adequation, that the most appropriate clustering algorithm for the purpose of this study was Divisive Hierarchical Clustering (DIANA).

As defined in the software vignette (sensu stricto Maechler et al. 2019 page 33): "The DIANA algorithm constructs a hierarchy of clusterings, starting with one large cluster containing all n observations. Clusters are divided until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected. The diameter of a cluster is the largest dissimilarity between any two of its observations. To divide the selected cluster, the algorithm first looks for its most disparate observation (i.e., which has the largest average dissimilarity to the other observations within the same cluster). This observation initiates the "splinter group". In subsequent steps, the algorithm reassigns observations that are closer to the "splinter group" than to the "old party". The result is a division of the selected cluster into two new clusters."

Most methods build their clusters starting from their terminal nodes (leaves), considering local patterns or proximate neighbors to make decisions. Instead, DIANA starts from the root of the tree, taking into consideration the overall distribution of the data points for the initial splits, gaining in accuracy and favoring larger groups coherence rather than smaller groups purity (Kassambara 2017, Dey 2019, Boehmke and Greenwell 2020). The first step of the algorithm involved consideration of all possible divisions of the data into two subsets (and so forth in every iteration), which is computationally demanding for large datasets, but allows to capture the main structure of the data (Kaufman and Rousseeuw1990).

Since this study is not about sorting elements into distinct natural units that exist in the field but classifying the landscape into general typologies of similarity along a multidimensional continuum, we consider DIANA to be the most appropriate approach. Anyways, for the sake of exploration and following the recommendations of an anonymous reviewer, we tested alternative clustering methodologies (Table A1.2) and compared them through a series of clustering stability and internal validation metrics

(Table A1. 3). The endeavor was not straightforward since many clustering algorithms were not compatible with mixed data nor gower distances, therefore we had to make adaptations: the two ordinal variables in our data set were converted to numeric (equidistant fractions of 1) and similarities were calculated with the Manhattan method, one of the most popular methods that is capable of dealing with outliers and no-normal distributions (similar to Gower). The results do not show any of the methods to be definitely better than the others (Fig. A1.4), therefore we found no reason not to use DIANA. Disclaimer, due the mentioned modifications the results of this experiment are incommensurable with the results of other analysis of our study.

Table A1.2. Clustering algorithms

Method	Definition
Hierarchical agglomerative[1]	each observation is initially considered as a cluster of its own (leaf). Then, the most similar clusters are successively merged until there is just one single big cluster (root).
K-means[1]	partition the points into *k* groups such that the sum of squares from points to the assigned cluster centres is minimized. At the minimum, all cluster centres are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre).
PAM[1]	it is based on the search for k representative objects or medoids among the observations of the data set, instead of using the mean, for partitioning a data set into k groups or clusters.
SOM[2]	type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction.
DIANA[1]	the inverse of agglomerative clustering. It begins with the root, in which all objects are included in one cluster. Then the most heterogeneous clusters are successively divided until all observations are in their own cluster.

[1] Kassambara A. 2017 Practical Guide To Cluster Analysis in R - Unsupervised Machine Learning, *STHDA Edition 1*.

[2] Wehrens R., and J. Kruisselbrink. 2018. Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, 87(7), 1–18. doi: 10.18637/jss.v087.i07

Table A1.3. Cluster validation metrics

Metric	Definition
APN[1]	measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed
AD[1]	computes the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed
ADM[1]	computes the average distance between cluster centers for observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed
FOM[1]	the figure of merit measures the average intra-cluster variance of the observations in the deleted column, where the clustering is based on the remaining (undeleted) samples. This estimates the mean error using predictions based on the cluster averages.
Connectivity[2]	reflects the extent to which items that are placed in the same cluster are also considered their nearest neighbors in the data space - or, in other words, the degree of connectedness of the clusters. And yes, you guessed it, it should be minimised.
Dunn index[2]	represents the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. As you can imagine, the nominator should be maximised and the denominator minimised, therefore the index should be maximized.
Silhouette width[2]	defines compactness based on the pairwise distances between all elements in the cluster, and separation based on pairwise distances between all points in the cluster and all points in the closest other cluster. Values as close to (+) 1 as possible are more desirable.
avg. within[3]	average distance within clusters.

[1] Brock, G., V. Pihur, and S. Datta. 2008. clValid: An R Package for Cluster Validation. Journal of Statistical Software, 25(4), 1-22. URL https://www.jstatsoft.org/v25/i04/

[2] Kulma, K. 2017. Cluster Validation In Unsupervised Machine Learning. https://kkulma.github.io/2017-05-10-cluster-validation-in-unsupervised-machine-learning/

[3] Hennig C. 2020. fpc: Flexible Procedures for Clustering. R package version 2.2-7. https://CRAN.R-project.org/package=fpc

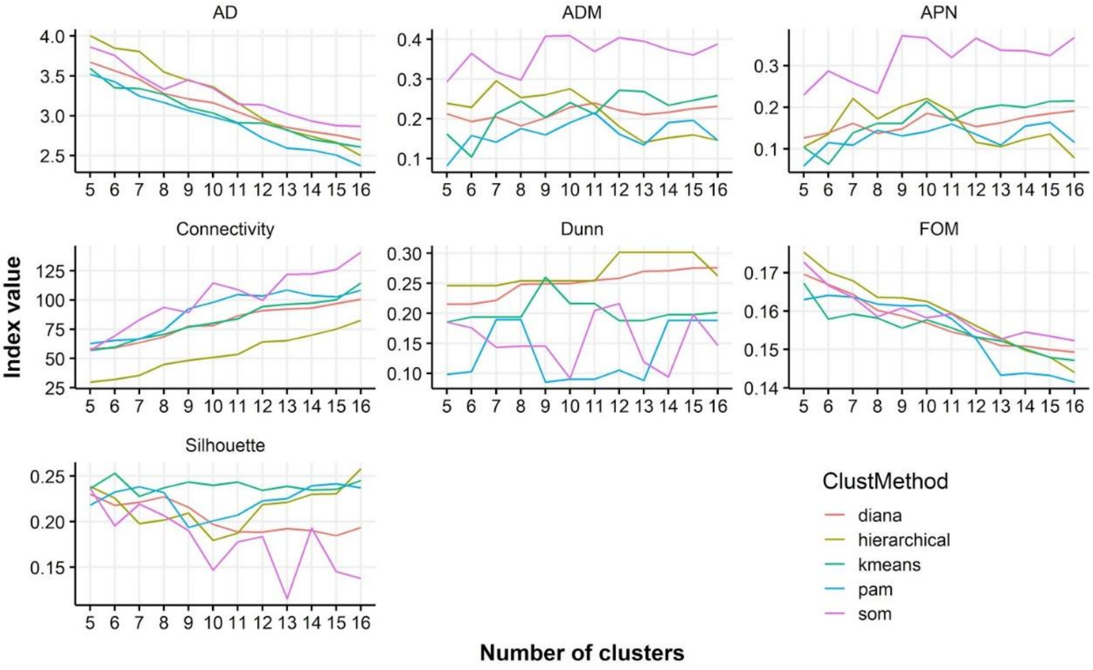


Figure A1.4. Comparison of clustering methods for reference. Performance of alternative clustering methods (line colors, Table A1.2) are compared in terms of stability and internal validation metrics (boxes, Table A1.3) along a gradient of number of clusters (K). Note the distance calculation algorithm for these analyses was Manhattan distance. Disclaimer: due the mentioned modifications the results of this experiment are incommensurable with the results of other analysis of our study.

4. Clustering results

In this section we describe how we analyzed the results of the DIANA analysis and agreed on a clustering output as the best SELS representation according to the specialists group's territorial knowledge. This included the decision on the number of clusters and its map layout, examination of the spatial representativity of the SELS across their territory, and evaluation of the relative contribution of each input variable to the classification.

Number of clusters

The output of DIANA is a dendrogram of hierarchical clusters. To decide at which height to cut the dendrogram we considered quantitative validation metrics (Figure A1.5) and analyzed the resulting spatial layout and clusters' statistics at the successive dendrogram cuts in relation to our territorial knowledge to agree on the optimal number of clusters. We disregarded clustering outputs with less than 5 or more than 16 clusters since we considered them not informative or too complex for the purpose of this study, respectively. As shown in Figure A1.5, alternative validation metrics did not converge into one unique “optimal number of clusters”, therefore the decision was made based mostly on expert’s knowledge. After analyzing the output maps and variable’s statistics the authors agreed the map depicting thirteen clusters was the most adequate representation of smaller-size SELS for the purpose of this study, and we found no evidence in the quantitative validation metrics to contradict that decision.

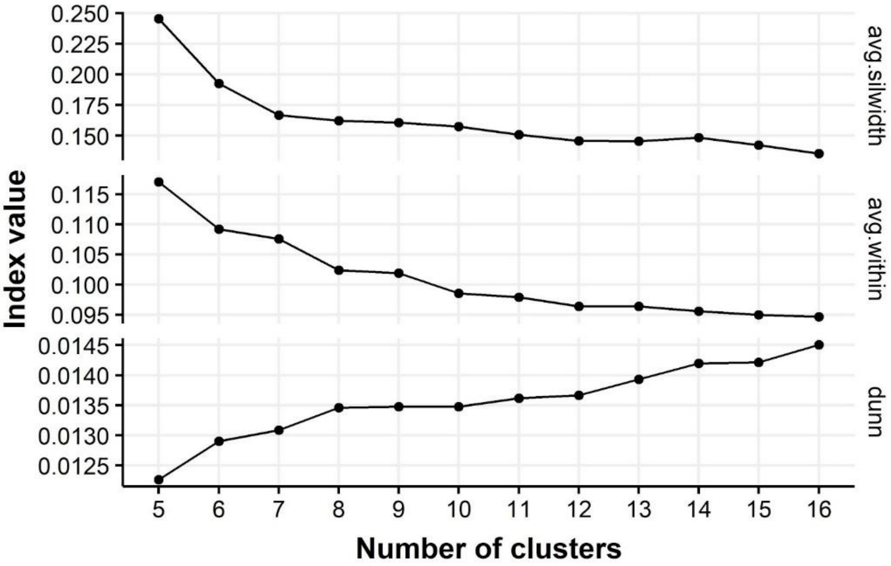


Figure A1.5. Identification of optimal number of clusters. Representation of three internal validation metrics performance: average silhouette width, average within distance, and dunn index (y axis) along the gradient of number of clusters (x axis).

5. Input variable’s relative contributions

To measure the input variable’s relative contribution we used Boosted Regression Trees. Regression trees are a regression/classification technique from machine learning where a model is trained to relate a response to their predictors by recursive binary splits. In boosted regression trees (BRT) the model accuracy is improved by repeating the regression tree algorithm adjusting the parameters in each iteration, similar to the “functional gradient descent” concept (Elith et al 2008). BRTs have very little restrictions, can handle different types of variables with no need of data transformation or outlier elimination, and can fit complex non-linear relationships. Through BRTs we can estimate the relative contribution of each input variable to the classification, measured as the number of times a variable is selected for splitting the tree, weighted by the model improvement by that split, and averaged across all trees (Elith et al. 2008).

We fitted 15 BRT different models in total, seeking to unravel the relative contribution of each variable in defining different target clusters: one multinomial for the 13 SELS simultaneously, one multinomial for the 5 SER simultaneously, and then individual binary models for each of the 13 SELS classes. Calculations were performed in R with the *gbm* function (*gmb* package, Greenwell et al. 2019) for the multinomial models and *gbm.step* function for the binomial models (*dismo* package, Hijmans et al. 2017). Model parameters are shown in Box A1.1. To evaluate how well the BRT models fit for each case we monitored the evolution of the holdout deviance along the iterations (Figure A1.6).

Box A1.1 BRT model parameters

Multinomial models:	**Binary models:**
learning rate (shrinkage) = 0.005	learning rate: 0.005
tree complexity (interaction depth) =1	tree complexity: 1 (default)
bag fraction: 0.5	bag fraction: 0.5
number of trees: 5000	number of trees: varies along the models

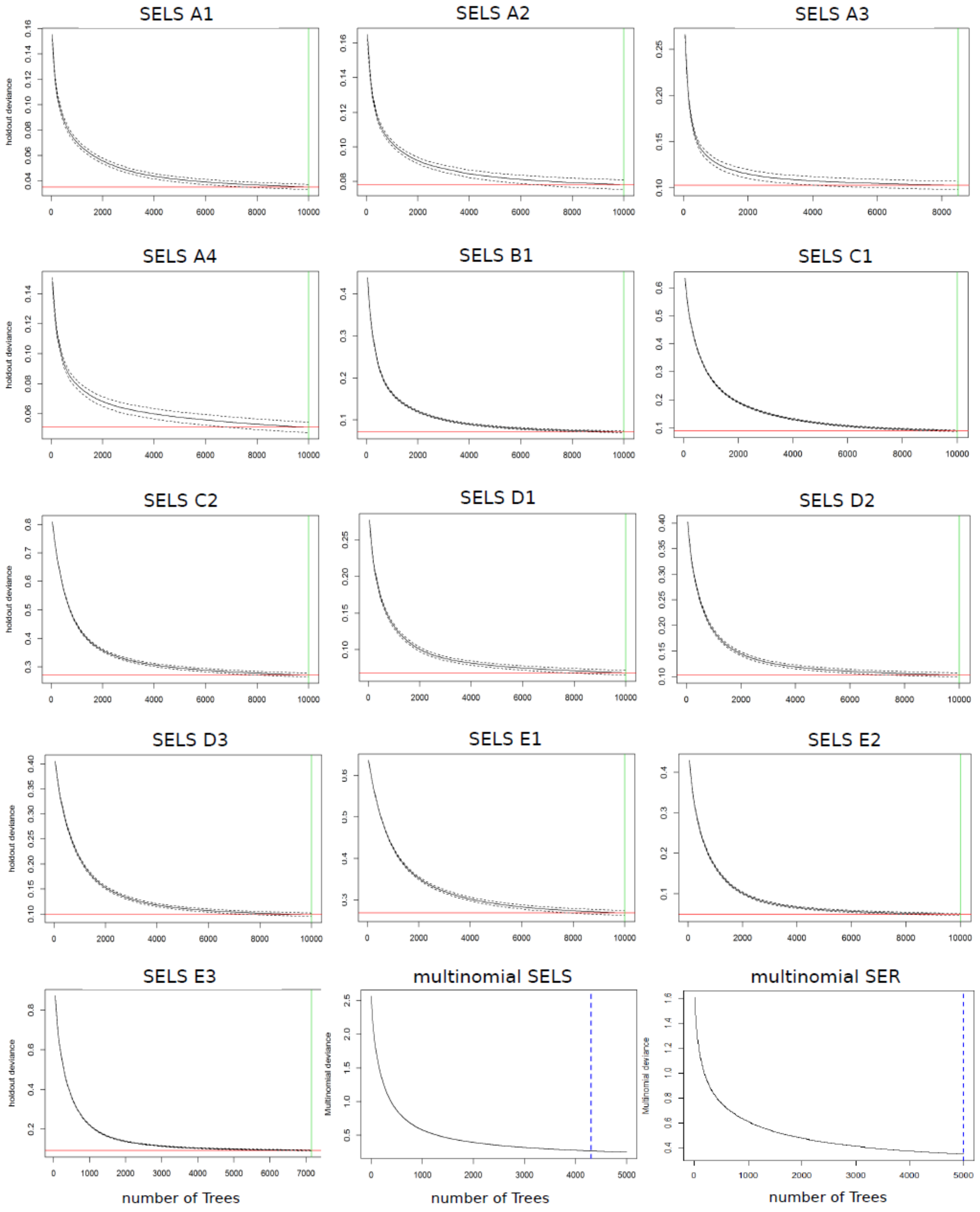


Figure A1.6. Holdout deviance along the iteration of the BRTs for the individual binomial models (SELS A1 to SELS E3) and the multinomial SELS and SER models.

LITERATURE CITED

Boehmke, B., and B. M. Greenwell. 2019. *Hands-On Machine Learning with R*. CRC Press.

Boillat, S., F. M. Scarpa, J. P. Robson, I. Gasparri, T. M. Aide, A. P. D. Aguiar, L. O. Anderson, M. Batistella, M. G. Fonseca, C. Futemma, H. R. Grau, S.-L. Mathez-Stiefel, J. P. Metzger, J. P. H. B. Ometto, M. A. Pedlowski, S. G. Perz, V. Robiglio, L. Soler, I. Vieira, and E. S. Brondizio. 2017. Land system science in Latin America: challenges and perspectives. *Current Opinion in Environmental Sustainability* 26–27:37–46.

Bright, E., P. Coleman, A. Rose, and M. Urban. 2012. LandScan 2011. Oak Ridge National Laboratory SE, Oak Ridge, TN.

Brock, G., V. Pihur, and S. Datta. 2008. clValid: An R Package for Cluster Validation. Journal of Statistical Software, 25(4), 1-22. URL https://www.jstatsoft.org/v25/i04/

ClimateSA: historical and projected climate data for Mexico, Central and South America. Climate data generated with the ClimateSA v1.0 softwarepackage, available at http://tinyurl.com/ClimateSA, based on methodology described by Hamann et al. (2013)

Debomit, D. 2019. Hierarchical clustering (Agglomerative and Divisive clustering). Blog entry at GeeksforGeeks org. https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/ (accessed February 2nd 2021)

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4):802–813.

Ellis, E. C., K. K. Goldewijk, S. Siebert, D. Lightman, and N. Ramankutty. 2010. Anthropogenic transformation of the biomes, 1700 to 2000. *Global Ecology and Biogeography* 19(5):589–606.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27(4):857–871.

Graesser, J., T. M. Aide, H. R. Grau, and N. Ramankutty. 2015. Cropland/pastureland dynamics and the slowdown of deforestation in Latin America. *Environmental Research Letters* 10(3):034017.

Greenwell, B., B. Boehmke, J. Cunningham, and GBM Developers, 2019. gbm: Generalized Boosted Regression Models. R package version 2.1.5. https://CRAN.R-project.org/package=gbm

Hennig C. 2020. fpc: Flexible Procedures for Clustering. R package version 2.2-7. https://CRAN.R-project.org/package=fpc

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith, J. 2017. dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo

Karagulle, D., C. Frye, R. Sayre, S. Breyer, P. Aniello, R. Vaughan, and D. Wright. 2017. Modeling global Hammond landform regions from 250-m elevation data. *Transactions in GIS* 21(5):1040–1060.

Kassambara, A. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kaufman, L., and P. J. Rousseeuw. 1990. Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, 344, 68-125.

Kulma, K. 2017. Cluster Validation In Unsupervised Machine Learning. https://kkulma.github.io/2017-05-10-cluster-validation-in-unsupervised-machine-learning/

Kreft, H., and W. Jetz. 2007. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences* 104(14):5925–5930.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

Monfreda, C., N. Ramankutty, and J. A. Foley. 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles* 22(1).

Nelson, A., 2008. Estimated travel time to the nearest city of 50,000 or more people in year 2000. Global Environment Monitoring Unit-Joint Research Centre of the European Commission, Ispra Italy. URL: http://bioval. jrc. ec. europa. eu/products/gam/.

Qian, H., and R. E. Ricklefs. 2008. Global concordance in diversity patterns of vascular plants and terrestrial vertebrates. *Ecology Letters* 11(6):547–553.

R Core Team 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Robinson, T. P., G. R. W. Wint, G. Conchedda, T. P. V. Boeckel, V. Ercoli, E. Palamara, G. Cinardi, L. D’Aietti, S. I. Hay, and M. Gilbert. 2014. Mapping the Global Distribution of Livestock. *PLOS ONE* 9(5):e96084.

Sayre, R., C. Frye, D. Karagulle, J. Krauer, S. Breyer, P. Aniello, D. J. Wright, D. Payne, C. Adler, H. Warner, D. P. VanSistine, and J. Cress. 2018. A New High-Resolution Map of World Mountains and an Online Tool for Visualizing and Comparing Characterizations of Global Mountain Distributions. *Mountain Research and Development* 38(3):240–249.

Siebert, S., P. Döll, J. Hoogeveen, J.-M. Faures, K. Frenken, and S. Feick. 2005. Development and validation of the global map of irrigation areas. *Hydrology and Earth System Sciences* 9(5):535–547.

Stein, A., K. Gerstner, and H. Kreft. 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecology Letters* 17(7):866–880.

UNEP-WCMC and IUCN, 2019. Protected Planet: The World Database on Protected Areas (WDPA [On-line, downloaded on May 2019], Cambridge, UK: UNEP-WCMC and IUCN. Available at: www.protectedplanet.net.

Wehrens R., and J. Kruisselbrink. 2018. Flexible Self-Organizing Maps in kohonen 3.0. Journal of Statistical Software, 87(7), 1–18. doi: 10.18637/jss.v087.i07

Weiss, D. J., A. Nelson, H. S. Gibson, W. Temperley, S. Peedell, A. Lieber, ... & P. W. Gething (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature, 553(7688), 333-336.

Winkler, K. J., M. W. Scown, and K. A. Nicholas. 2018. A classification to align social-ecological land systems research with policy in Europe. Land Use Policy, 79, 137-145.

Zikin M. Diversidad lingüística latinoamericana. Muturzikin.com © 2007. http://www.muturzikin.com/carteamerique.htm