# Geospatial data applied to the Brazilian Cerrado

**How to cite this document:**
Kennedy, C.M., Baumgarten, L., Matsumoto, M., Evans, J.S., and Miteva, D.A. (2014). Geospatial data applied to the Brazilian Cerrado. The Nature Conservancy. Available at: http://www.conservationgateway.org/ConservationPractices/EcosystemServices/tnc_dow_collaboration/brazil/Pages/default.aspx.

This document summarizes the geospatial data used and the methodology employed to generate data that were not readily available from external sources.

## Data sources

The primary geospatial layers used in the analyses are summarized in the GIS Data Sources Table provided on the The Nature Conservancy website (see link above). These data were used to derive the other geospatial data for our analyses in relation to the Brazilian Cerrado case study (e.g., production cost layers). In the data sources table, the ***Type*** column indicates whether the data were created by our research team (listed as "internal") or whether they were available through an outside source (listed as "external"). Below we provide further details on the creation of the internal GIS layers.

## Delineation of the study area

We spatially delineated the boundaries of the study area using two criteria: (1) the proposed sugarcane licensing area in our study region, which spans part of the Tijuco River watershed and other small rivers flowing directly into the São Simão Reservoir, and (2) the full extent of the sub-watershed that could be affected by potential commercial sugarcane expansion in the region. The final study area covers 373,043 ha and is bordered by a combination of watershed ridges and channels derived from a digital elevation model. This area covers the entire Ribeirão São Jerônimo watershed and follows the major stream channels found in that watershed (Fig. 1).

## Development of the current land use/land cover (LULC) layer

### Geometric correction, assessment, and image processing

Because the study area is not covered by a single satellite scene, we used a combination of satellite imagery from SPOT (2.5m resolution, June 2010), ALOS PRISM (2.5 m resolution, May and June 2009) and ALOS AVNIR2 (10 m resolution, June 2009) (Fig. 2). The images were verified and, when necessary, orthorectified using 31 training points collected from the study area in October 2012 via a dual frequency geodesic GPS (Fig. 3). We used a fusion function between PRISM and the multispectral AVNIR 2 images to obtain a color image of 2.5 m of spatial resolution. The final mosaic resolution meets the accuracy at the 1:25,000 scale required by Brazilian standards (PEC A, highest grade, Decree 89.817 06/20/1984). The layer was projected using UTM Zone 22S projection, and the SIRGAS 2000 datum.

### LULC mapping

To create a land use/land cover (LULC) dataset, we used supervised classification based on user-selected class training samples (forest, pasture, urban, etc) (Jensen 2004). The draft classification was subjected to expert review to reduce the errors and noise generated by the supervised process. We ground-truthed features with low level of certainty based on satellite imagery. The minimum mapping area was defined as one hectare, with smaller features

assigned to the dominant class within that spatial unit. We applied a boundary simplification process to reduce the total number of vertices of each feature. To supplement the land cover classification, we manually digitized the road networks, streams and water bodies.

The final LULC map spans the study area and a 1000-meter buffer and is available as a shapefile (in a vector format). It includes 12 LULC classes (Table 1). The delineation, description and definition of classes was based on that developed by IBGE (2006) and the Forestry Survey of Minas Gerais State (State Forestry Institute – IEF, Scolforo, J. R. and L. M. T. Carvalho. 2006), and adapted by TNC staff for the region. A map of the current LULC layer is presented in Fig. 4. The predominant LULC class is pasture (229,367 ha, 61.5%); natural vegetation classes comprise 73,051 ha (19.6% of the study area) (Table 2).

## Development of the precipitation layer

As an input into our water modeling, we created a precipitation layer from data from the National Institute of Meteorology (INMET, http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep, stations: 83514, 83521, 83565) and the National Water Agency (ANA, http://hidroweb.ana.gov.br/, stations: 1849002, 1849026, 180002, 1851000, 1949003, 1949004, 1950011, 1950012, 1950018). Using data for the past 20 years (1993-2012), we calculated the annual average total rainfall for each station and assigned to the geographical location of the station. To create a surface precipitation layer from these point data, we used a spline interpolation method, which populates values for the spaces between points on the basis of a mathematical function that minimizes overall surface curvature and generates a smooth surface that passes exactly through the input points (Mitas and Mitasova, 1988). The distribution of the interpolated precipitation showed a variation in the annual average rainfall (range: 1271 to 1403 mm), with precipitation increasing from west to east in our study area (Fig. 5).

## Development of the predictive vegetation layer

To predict the geographic distribution of the natural vegetation types across the study area, we used Random Forests, which is a weak-learning ensemble method (Cutler et al., 2007; Breiman 2001). Random Forests has been shown to be a very robust predictive model when the underlying ecological or statistical processes are unknown (Evans et al., 2011; Evans and Cushman 2009; Falkowski et al., 2009). Because of the nonparametric and hierarchal nature of the model, issues associated with autocorrelation, nonlinearity, overfitting and non-stationarity are minimized (Evans et al., 2011; Cutler et al., 2007).

Several independent variables, hypothesized to represent abiotic ecological characteristics in the area, were utilized in the model. Our covariates included soil types, distance to streams and several geomorphometric variables: wetness index (Gessler et al., 1995), roughness (Riley et al., 1999), slope*cosine (aspect) (Stage 1976), relative slope position (Pike et al., 2009), surface relief ratio (Pike and Wilson 1971), and topographic radiation index (Roberts and Cooper 1989) (Table 3). There are 12 soil types found in our region, with the most prevalent soil type being red latosols (latossolos vermelho) (Fig. 6, Table 4). The ecological rational behind these covariates are described in the associated citations and in Evans and Cushman (2009) and Murphy et al., (2010). The geomorphometric variables were calculated from a 90m SRTM digital elevation model (Rabus et al., 2003) using the Geomorphometric and Gradient modeling Toolbox in ArcGIS 10.0 (Evans et al., in prep).

Starting with the current LULC polygon dataset (n=6298), we removed human-modified cover classes (pasture, cropland, urban areas, and roads). The remaining polygons (n=5333)

represented discrete LULC units spanning the five native vegetation types: cerradão, cerrado, riparian forest, semi-deciduous forest, and wetland (for definitions see Table 1). Because of the discrete (categorical) nature of the dependent variable (presence/absence) and the soil type covariate, we implemented a conditional sampling procedure to generate a sample that represented spatial and statistical variation in both the discrete and continuous process in the covariates (Cressie 1996). Using the software R (R core team 2012) and associated libraries *sp*, *rgdal*, and *raster* (Pebesma and Bivand, 2005; Bivand et al., 2013; Hijmans and van Etten 2013), we developed a model that generated a 10% area-weighted random point sample for each discrete land unit (n=5333 polygons) resulting in 19,213 random samples. To avoid aggregation issues (Cressie 1996) and pseudoreplication (Hurlbert 1984), an isotropic kernel function (Warton and Shepherd 2010; Hengl et al., 2009) is applied within each polygon sample unit to ensure that the random sample is spatially balanced and independent. We then assigned the underlying raster value(s) for the associated land cover class (dependent variable) and the value from our candidate covariates (independent variables) to each point in the random sample.

Because recursive partitioning models perform better on binary data (Ham et al., 2005), we ran an independent model for each habitat class, assigning 1 to the class being modeled and 0 otherwise. For each model we predicted a 90m probability surface representing the positive [1] class. We then created a final class raster by assigning each pixel the class associated with the highest probability, using a $p$=0.45 threshold as the lower bound. Following Murphy et al. (2010), we applied the model selection procedure that uses permutated variable important measures and model optimization (i.e. minimizing both out-of-bag and within-class errors) to select covariates and the most parsimonious model(s); we found that all variables were contributing to the model. We also tested for class balance to avoid zero inflation problems (Evans et al., 2011; Evans and Cushman 2009; Jiménez-Valverde and Lobo 2006) and found all of our models were balanced.

We evaluated model performance using percent correctly classified (PCC), sensitivity, specificity, precision recall rate (Fleiss and Cohen 1973), and the area under the ROC curve (Fawcett 2006; Dodd and Pepe 2003). These indicators revealed the strongest support for the cerradao, cerrado, and wetland classes, with marginal support for riparian forest and semi-deciduous forest primarily due to interclass confusion between each other (Table 5). Given the inability to distinguish well between semi-deciduous and riparian forest types and the fact that they are similar in geographic distribution as well as vegetation communities for our region, these two habitat types were aggregated for our subsequent land use modeling. As a final step, we superimposed the natural vegetation classes from the current LULC layer onto the predicted vegetation and, in the case of discrepancies, changed the predicted vegetation class to the one in the current LULC. A mask was created using polygons, in the original LUCU data, classified as water. This mask was utilized to set background values (areas predicted as no data) to water in the resulting predictive model surface.

A map of the predicted natural vegetation for our study area is presented in Fig. 7. For a comparison with the current landscape see Fig. 4 and Table 2. The covariates that had the strongest influence on predicting the different habitat types were elevation and roughness, followed by soil types and distance to water, and to a lesser extent compound topographic index and stream distance (Table 6). Cerradão was modeled as the dominant vegetation type (213,288 ha, 67.9%), followed by semi-deciduous (riparian) forest (67,159 ha, 21.4%) and lastly cerrado (18,818 ha, 6.0%) and wetlands (15,098 ha, 4.8%) (Fig.7). Cerradão was predicted at lower elevations in moderately wet areas and more fertile cambisols and latosols

soils (Fig. 8); cerrado on drier areas, higher elevation slopes, on latosols and red-yellow ultisols soils in the southern range (Fig. 9); semi-deciduous (riparian) forest around waterways and intermediate slopes on latisols soils (Fig. 10, 11); wetlands on the lowest elevations on latosols and entisols soils in wet areas adjacent to streams and water bodies (Fig. 12).

## Tree mapping

To assess the costs of clearing and restoring existing vegetation on pastures, we created a layer of the current number of trees per pasture pixel. We used *the Detect and vectorize individual trees* function available for gvSIG under SEXTANTE algorithms (http://www.gvsig.org). Similar to the supervised imagery classification, which requires a user-assigned sample of LULC classes, the process requires a sample signature of an individual tree. For the tree mapping algorithm, the radiometric response of each pixel is considered along with the size and shape of pixels to be considered as an individual tree. Upon testing different bands, we selected band 2 (0.52-0.60µm-- green visible spectrum) of SPOT5 satellite due to the best response for this analysis at a 2.5 m of spatial resolution. To reduce the processing time and to exclude natural cerrado areas (where isolated trees can occur naturally), we used the existing pastures as a mask and delineated trees only in these agricultural areas. Our analysis was based on a total sample of 72 trees. The output was a point vector layer that contains a shape coefficient and the canopy surface area (Fig. 13). The shape coefficient depicts information about the tree canopy size and canopy shape, and relates to whether an individual tree or group of trees were delineated; a coefficient close to 1 referenced a single mapped tree and lower values pertained to features of more than one tree.

## Land tenure mapping

While some property boundaries are publicly available, full coverage was not available for our study area. Records of property boundaries are only available in paper format from notary offices. Even though some governmental agencies provide public records, the spatial coverage is usually poor. Therefore, we generated a spatial map of the farm boundaries for our study area based on 4 sources:

1. All farm polygons available from the National Institute of Colonization and Land Reform (INCRA).
2. A previous land tenure assessment ordered by Santa Vitória City Hall.
3. Two surveys in local notary offices where we accessed maps and property descriptions.
4. A field survey during which we collected spatial information of farm boundaries.

The information collected was digitized from paper maps, or converted from CAD files, to a single geodatabase. Eventual overlaps or gaps between neighbor farms were corrected by checking boundaries against the satellite imagery available.

Our efforts resulted in property boundaries for 74% of the study area or 1174 farms. Because some farms were only partially within our study area, we created artificial boundaries using roads, rivers, land use and households to delimit borders. The final land tenure layer contains 1304 properties. For confidentiality reasons, we do not provide a spatial map of farm boundaries for our study area.

## Assessment of Forest Code compliance

### Summary of Forest Code requirements

According to the Brazilian Forest Code (Federal law 12,651, May 25th, 2012), all agricultural producers should set aside land for Permanent Preservation Areas (PPAs), which include riparian forests, stream headwaters, vegetation on steep terrain, and other hydrologically sensitive areas. We assumed all deforestation in the study area occurred before July 2008, which is realistic given the historic land use trajectory. The PPA requirements for the properties in the study region are summarized in Table 7.

Depending on the size and location of properties, farmers are also required to maintain a portion of native vegetation in legal reserves (LRs). LR selection may be based on Watershed Planning, Ecological Economic Zoning, and/or mapped environmentally fragile areas and important areas for biodiversity. In the Cerrado and Atlantic Forest Biomes that make up most of central and southern Brazil, and where our study region lies, LR requirement is 20% for most farms. If a farm does not have the required natural vegetation and was deforested before 2008, PPAs can be counted towards the LR requirement. Farms smaller than 120 ha that lack sufficient remnants (and when conversion happened before 2008) are exempt from setting aside additional vegetation. Farms larger than 120 ha, and out of compliance, are required to have the necessary area of native vegetation restored. Alternatively, the requirements can be offset in available private or public areas elsewhere in the same watershed or biome. LRs are not required to form a single or continuous habitat patch. If the remnant area exceeds the requirements of the law, the surplus area can be legally converted or can be used to offset the deficit of other farms.

### Compliance options

Although degraded or converted PPAs need to be restored in the specific locations defined by the law, the placement of LRs is flexible and depends on the preferences of individual landowners, the characteristics of the farm (such as area of existing natural vegetation and profitability), and the agreements with state environmental officials. The only exception is when the clearing of natural vegetation on a property occurred after 2008. In such cases, a landowner is required to restore natural vegetation on the farm.

The Forest Code offers three main options to for compliance with LR requirements. Under the first option, usually viable in areas with more natural vegetation, compliance within a property is met by selecting and designating current natural remnants within its boundaries as LRs ("protection option"). Under the second option, a portion of the currently converted land is restored to natural vegetation ("restoration option"). Once an area within a farm is selected for a LR, the landowner can either actively plant natural vegetation or can allow natural succession to take place. If the landowner chooses to follow a planting routine, they can continue to use some of area designated as LR as long as they actively restore at least 10% of the LR area every 2 years. In contrast, opting for restoration through natural succession requires that the landowner completely remove the LR designated portion of land from agricultural production immediately. The third option for LR compliance allows for the establishment of LR allocation outside of a landowner's property by protecting existing natural fragments, as long as the selected land meets specific criteria established by the law ("offset option"). In particular, the offset area needs to match the LR size required, be located in the same biome, be on farms with surplus natural vegetation, and, in the case of offsetting on

another state, to be located in priority areas defined by state or federal government. Such ex-situ compensation can be economically viable, especially for properties with high quality soil on which highly profitable crops can be grown.

### Assessing Forest Code compliance

We assessed compliance for each farm within our study area using the current land cover/land use and land tenure maps (described above). Since all natural vegetation classes from the LULC map can count towards the PPAs or LRs requirements, we did not differentiate between natural vegetation classes.

First, we determined whether each farm met the PPA requirements summarized in Table 8. We treated all existing natural remnants in hydrologically sensitive areas as protected. If a farm had less than the required PPA area, we used Table 7 to calculate the additional area needed. This procedure allowed us to generate a map of existing and needed PPA areas (as illustrated in Fig. 14).

Second, we assessed whether each farm met the LR requirements. Farms between 120 ha and 300 ha, require up to 20% of their land (which may include PPA to reach this amount) to be restored. If a farm has more natural vegetation than is required by the Forest Code, it is considered to have a surplus of remnants, which can be used to offset the LR requirements of other farms. For example, if a large farm (>120 ha) is covered with more than 20% natural vegetation, the additional area is considered surplus. According to the revised FC, for farms smaller than 120 ha there is no deficit, so all additional remnant area over required 20% of their land is considered as surplus. Farms can also have the exact area needed or have a deficit of natural vegetation. In the latter case, smaller farms (up to 120 ha) are not required to take additional measures to comply and only need to protect their current remnants. Farms larger than 120 ha need to restore or offset the deficit LR area, and the sum of offset, restored, or existing natural vegetation within the property is no less than 20% of the property area.

Of the 1304 analyzed farms in our study area, only 254 farms (19.5%) were found to be in compliance with the Forest Code for both PPAs and LR requirements (Table 8). A total of 17,800 ha of PPAs and 51,148 ha of LRs were estimated to be required (Table 9). A total of 5,827 ha of PPAs were currently degraded or converted and need to be restored; for LRs, the total deficit amounted to 973 ha. The prevalent type of farm, properties between 120 and 300 ha, had the highest total deficit. In aggregate, however, the net difference between surplus and deficit areas was relatively small (Table 9); this finding suggests that offsets limited to the watershed are a viable option to meet FC requirements in the region.

## Acknowledgements

We thank TerraVision for digital image processing and ground truthing of the land cover and land use map and the interpretation/digitization of the stream networks and transportation system; Luciana Estevam, Milena Ribeiro and Jonathan Braga for their support in the development of the land tenure map; and Andy Liaw for statistical guidance on the analyses of the predictive vegetation map.

## References


Breiman L. (2001) Random forests. *Machine Learning* 45:5–32.

Bivand R., T. Keitt, and B. Rowlingson (2013). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.8-9. http://CRAN.R-project.org/package=rgdal

Cressie N (1996) Change of support and the modifiable areal unit problem. *Geographical Systems* 3:159–180.

Cutler D.R., T.C. Edwards Jr., K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, J. Lawler (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792.

Dodd, L.E and M.S. Pepe (2003). Partial AUC Estimation and Regression. Biometrics 59(3): 614–623. doi:10.1111/1541-0420.00071

Dow Chemical (2008). Environmental Impact Report, EIA-RIMA Usina Santa Vitória. Referencia 0078456.

Evans J.S., M.A. Murphy, Z.A. Holden, S.A. Cushman (2011). Modeling species distribution and change using Random Forests in *Predictive species and habitat modeling in landscape ecology: concepts and applications*. Eds Drew CA, YF Wiersma, F Huettmann. Springer, NY.

Evans, J.S. and S.A. Cushman (2009) Gradient Modeling of Conifer Species Using Random Forest. *Landscape Ecology* 5:673-683.

Evans, J.S., J. Oakleaf, S.A. Cushman, D. Theobald (In prep) An ArcGIS toolbox for Geomorphometric and Gradient modeling. *Ecography.* Available: http://evansmurphy.wix.com/evansspatial.

Falkowski M.J., J.S. Evans, S. Martinuzzi, P.E. Gessler, A.T. Hudak (2009) Characterizing forest succession with lidar data: an evaluation for the inland Northwest, USA. *Remote Sensing of the Environment* 113:946–956.

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27:861–874.

Fleiss, J. L., & J. Cohen (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33:613-619.

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. (1995) Soil-landscape modeling and spatial prediction of soil attributes. *International Journal of GIS* 9(4):421-432.

Ham, J., Y. Chen, M.M. Crawford, J. Ghosh (2005) Investigation of the Random Forest Framework for Classification of Hyperspectral Data. IEEE Transactions on Geoscience and Remote Sensing 43(3):492-501.

Hengl, T., H. Sierdsema, A. Radovic, and A. Dilo (2009) Spatial prediction of species distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling*, 220(24):3499-3511.

Hijmans R.J.. and J. van Etten (2013). raster: raster: Geographic data analysis and modeling. R package version 2.1-25. http://CRAN.R-project.org/package=raster.

Hurlbert, S.H. (1984) Pseudoreplication and the design in ecological field experiments. *Ecological Monographs* 54(2):187-211.

IBGE (2006) *Manual Técnico de Uso da Terra*, 2. ed., Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, pp. 91.

Jensen, J.R. (2004) *Introductory Digital Image Processing*, 3rd Ed., Prentice Hall, Upper Saddle River, NJ, 526 pages.

Jiménez-Valverde A., and J.M. Lobo (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distribution* 12:521–524.

Mitas, L., and H. Mitasova. 1988. General Variational Approach to the Interpolation Problem. Computer and Mathematics with Appllications. Vol. 16. No. 12. pp. 983–992. Great Britain.

Murphy M, J.S. Evans, and A. Storfer (2010) Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. Ecology 91:252-261.

Oliveira-Filho, A. T.; Scolforo, J. R. S.; Oliveira, A. D. & Carvalho, L. M. T. (2006) Definição e delimitação de domínios e subdomínios das paisagens naturais do Estado de Minas Gerais in *Mapeamento e inventário da flora nativa e dos reflorestamentos de Minas Gerais.* Eds. Scolforo, J. R. S. & Carvalho, L. M. T. Editora UFLA, Lavras, pp. 21-35.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews/.

Pike, R.J., I.S. Evans and T. Hengl (2009) Geomorphometry: A Brief Guide. Developments in Soil Science Volume 33.

Pike, R.J., and S.E. Wilson (1971) Elevation relief ratio, hypsometric integral, and geomorphic area altitude analysis. *Bulletin of the Geological Society of America* 82:1079-1084

Rabus, B., M. Eineder, A. Roth, R. Bamler (2003) The shuttle radar topography mission- a new class of digital elevation models acquired by spaceborne radar. *Photogrammetry and Remote Sensing*. 57:241-262.

Riley, S. J., S. D. DeGloria and R. Elliot (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5:1-4

Roberts. D. W., and Cooper, S. V., 1989. Concepts and techniques of vegetation mapping. In Land Classifications Based on Vegetation: Applications for Resource Management. USDA Forest Service GTR INT-257, Ogden, UT, pp 90-96.

R Core Team (2012). R: A language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Scolforo, J. R. & L. M. T Carvalho. (2006) Mapeamento e inventário da flora nativa e dos reflorestamentos de Minas Gerais, UFLA – Lavras, pp. 288.

Silva, A. M., Alvares, C. A., & Watanave, C. H. (2011). Natural Potential for Erosion for Brazilian Territory. In Soil Erosion Studies.

Stage, A. R. 1976. An Expression of the Effects of Aspect, Slope, and Habitat Type on Tree Growth. *Forest Science* 22(3):457- 460.

Warton, D.I., and L.C. Shepherd (2010) Poisson Point Process Models Solve the Pseudo-Absence Problem for Presence-only Data in Ecology. *The Annals of Applied Statistics* 4(3):1383-1402.

UFV - CETEC - UFLA - FEAM. 2010. Mapa de solos do Estado de Minas Gerais. Belo Horizonte, Fundação Estadual do Meio Ambiente. 49p. Available on: http://www.feam.br/noticias/1/949-mapas-de-solo-do-estado-de-minas-gerais.

## Figures

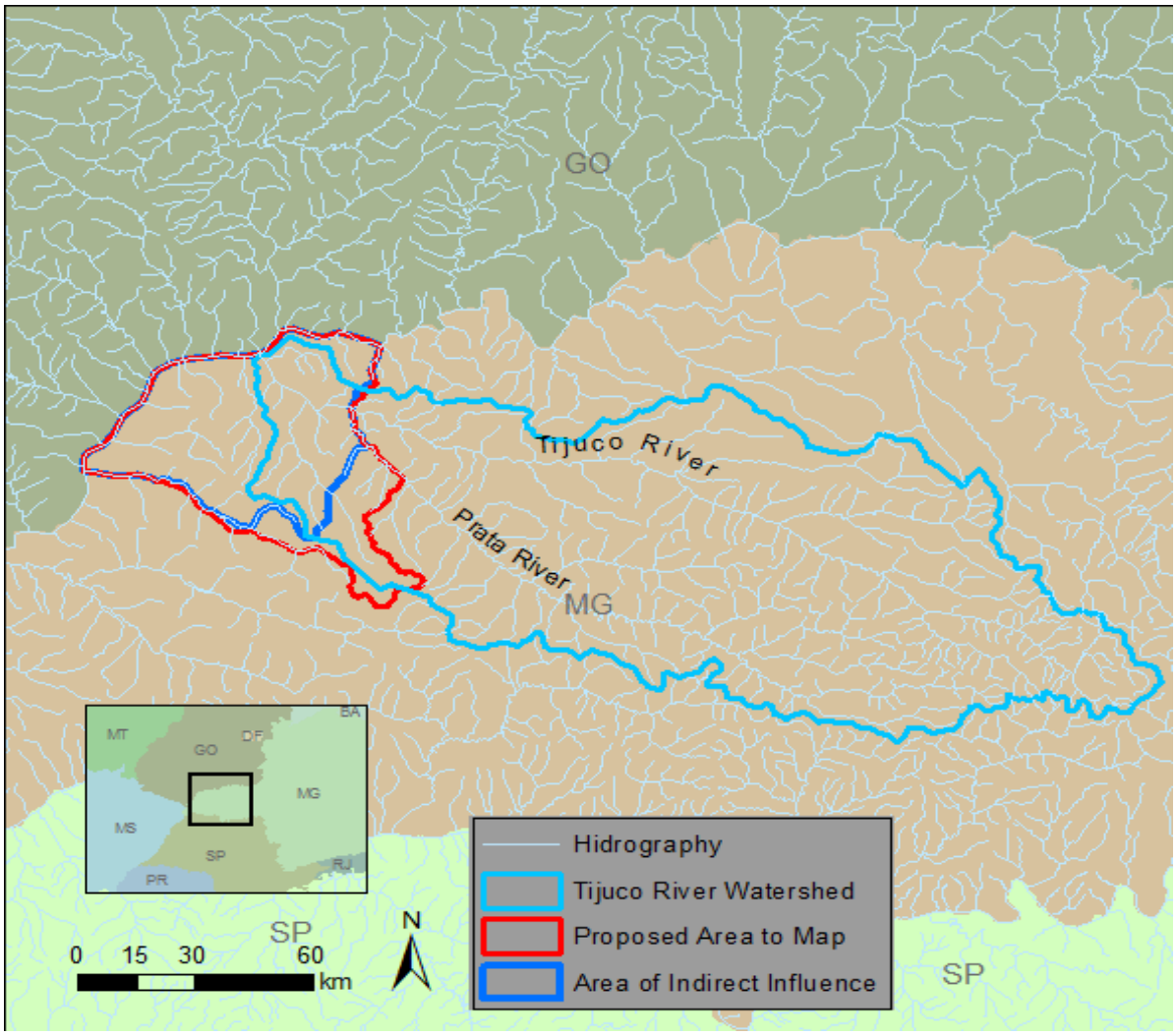


**Figure 1.** Map of Tijuco River watershed (in light blue) in relation to our study area as delineated by the Ribeirão São Jerônimo watershed (in red). Area of Indirect Influence is the region potentially affected by the expansion of sugarcane production (Dow Chemical 2008).

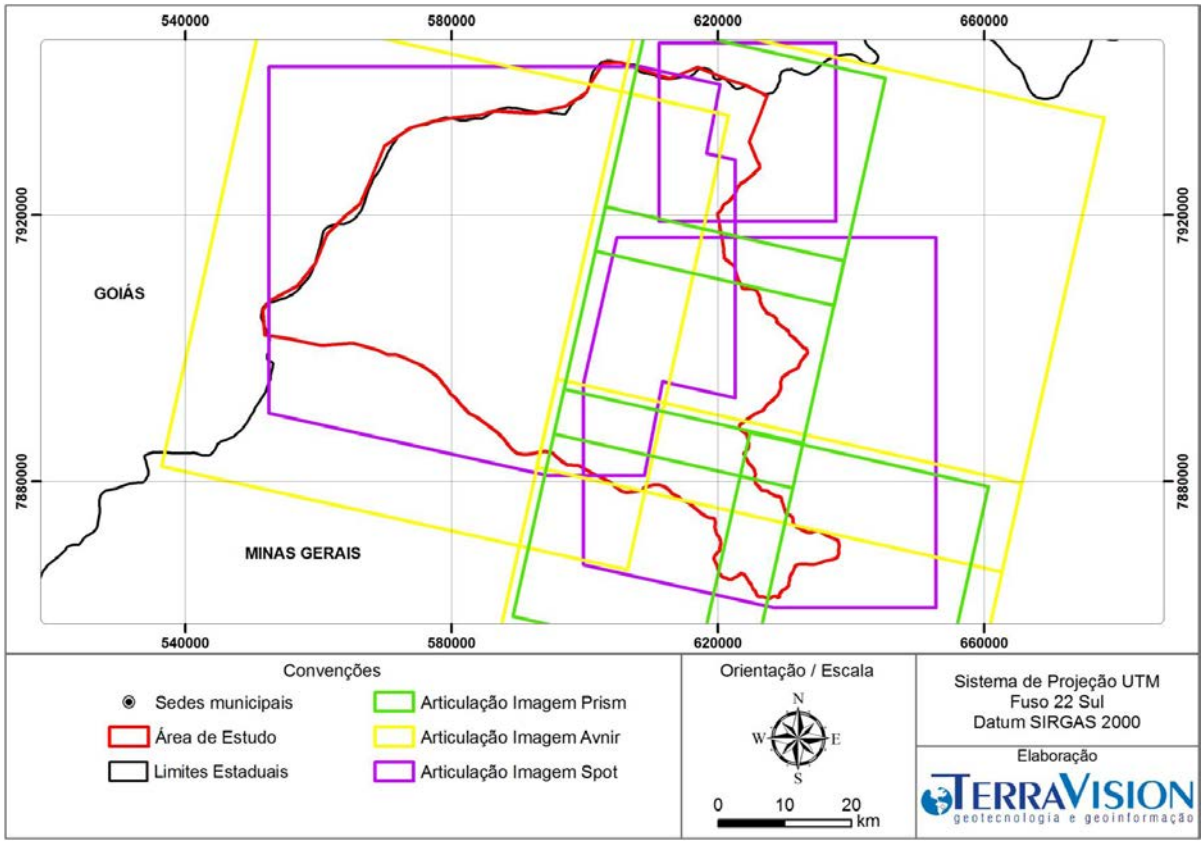


**Figure 2.** Satellite imagery coverage in the study area.

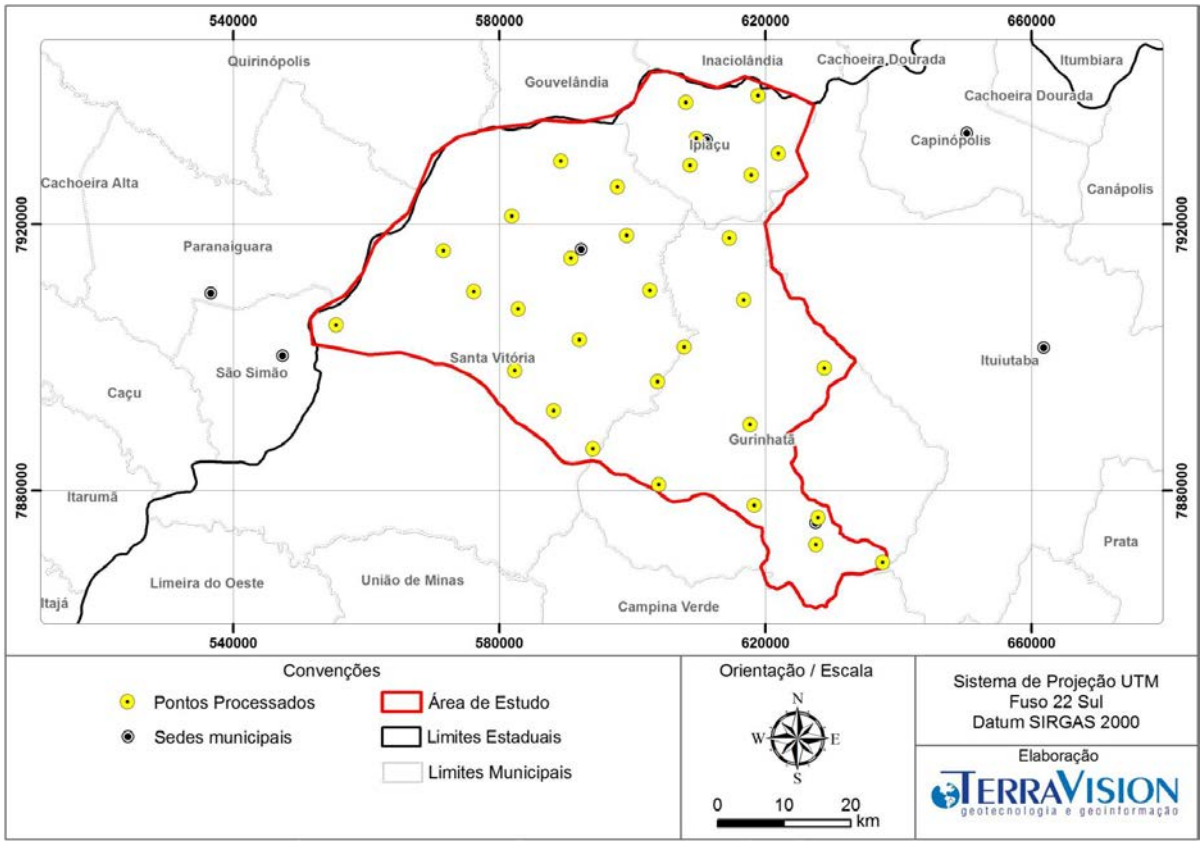


**Figure 3.** Ground point control distribution within the study area.

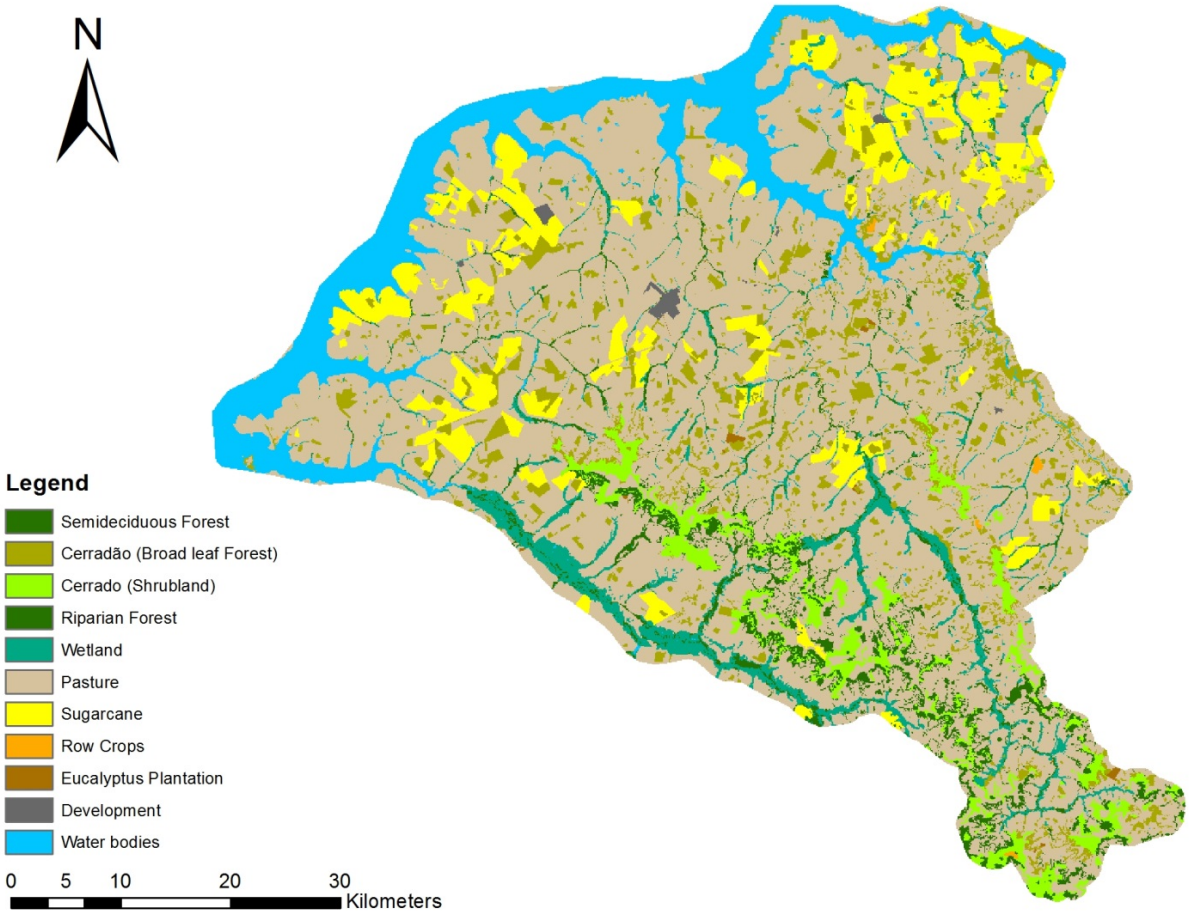


**Figure 4.** Land use/land cover map created for the study area (based on the Ribeirão São Jerônimo watershed) located in the state of Minas Gerais, Brazil.

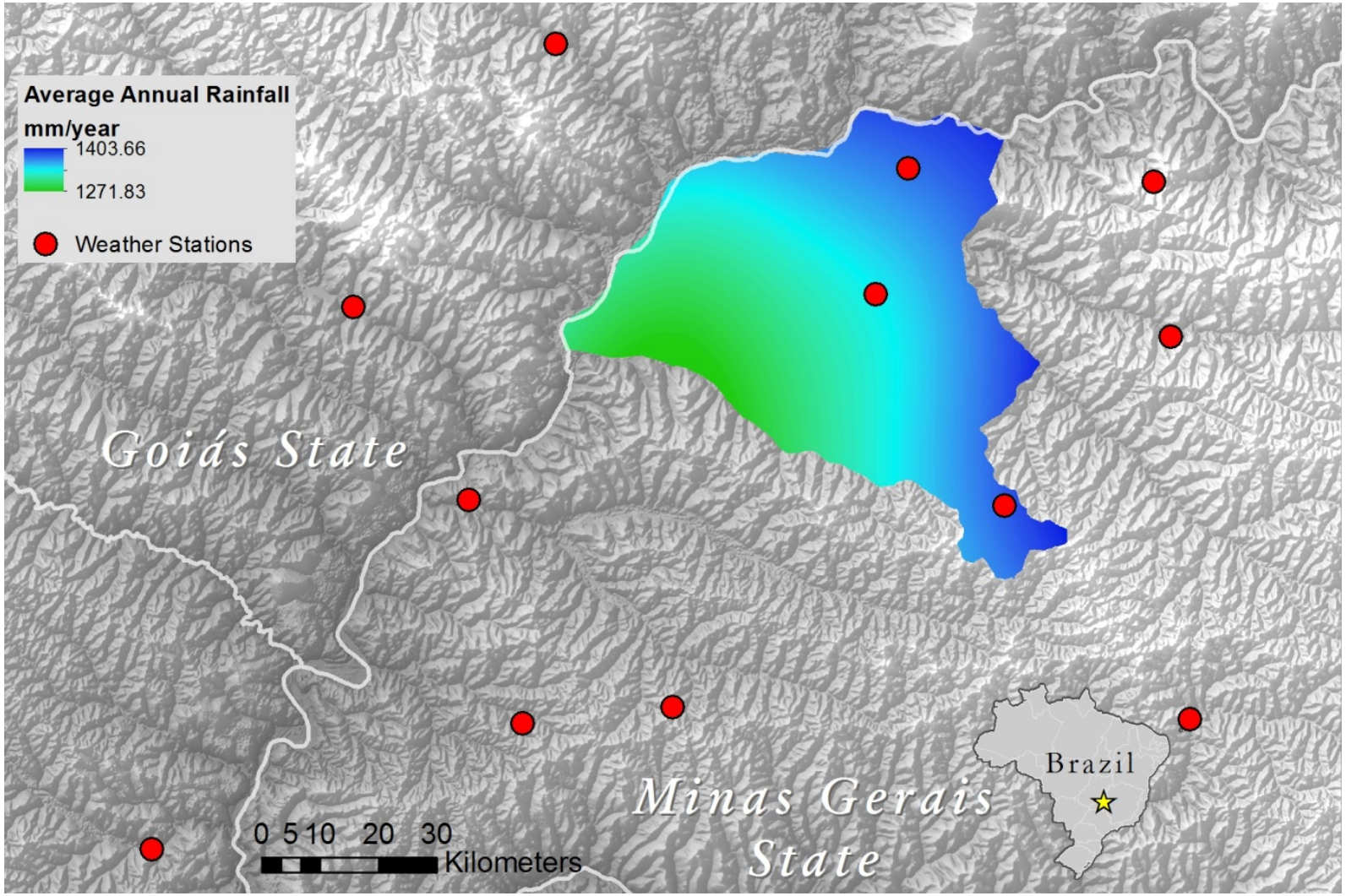


**Figure 5.** Annual average rainfall (mm/year) interpolated based on the weather stations in the study area.

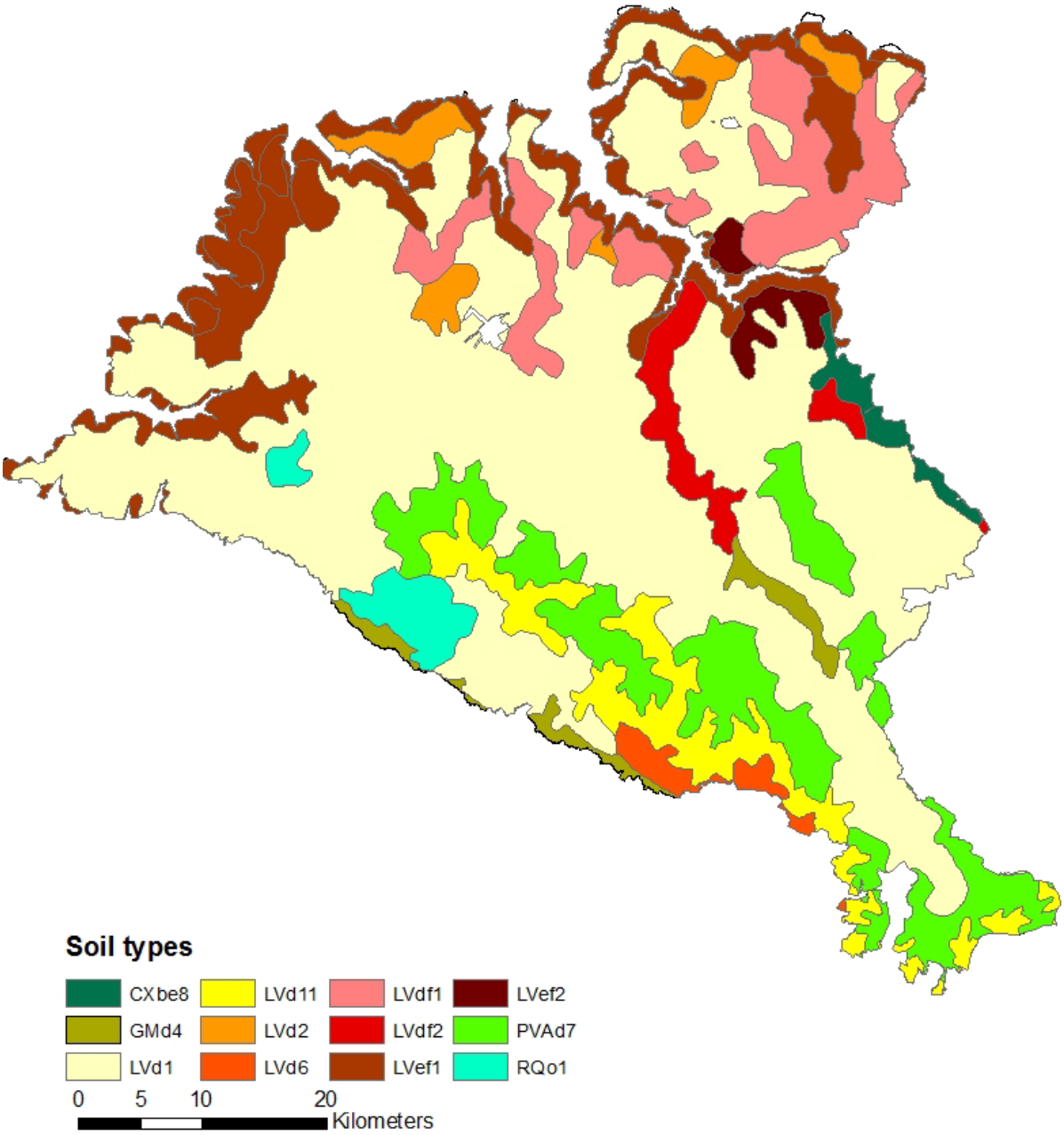


**Figure 6.** Distribution of soil types within the study area. Type CXbe corresponds to inceptisols; type GMd4-to gleisols, types LV-to red latosols, type PVAd7-to red-yellow ultisols, and type RQo1-to quartzsipsamment (Table 4). The most common soil type is red latosols (oxisols, although they vary in fertility).

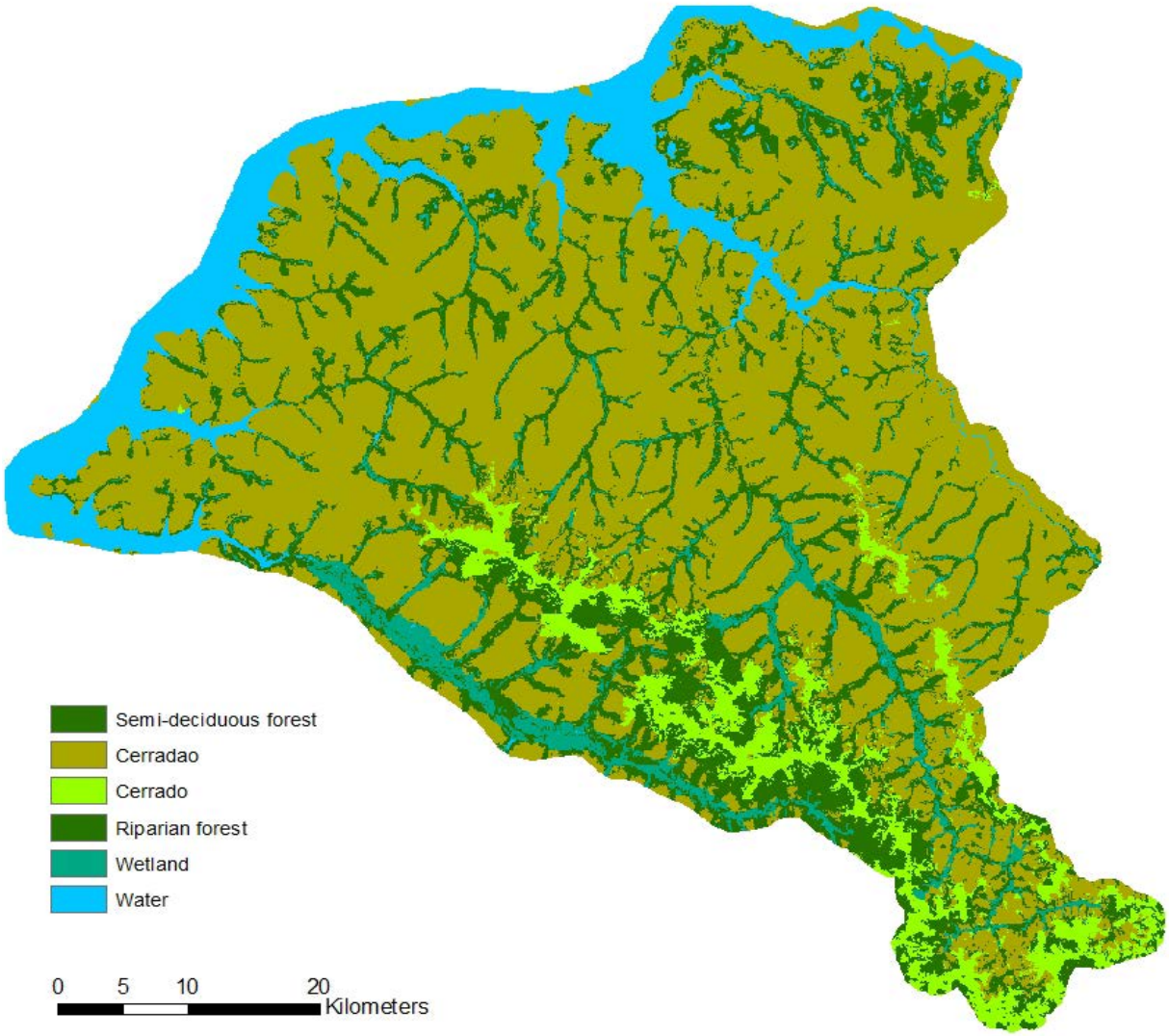


**Figure 7.** Predicted distributions of natural vegetation types for our study area modeled in the absence of human disturbance.

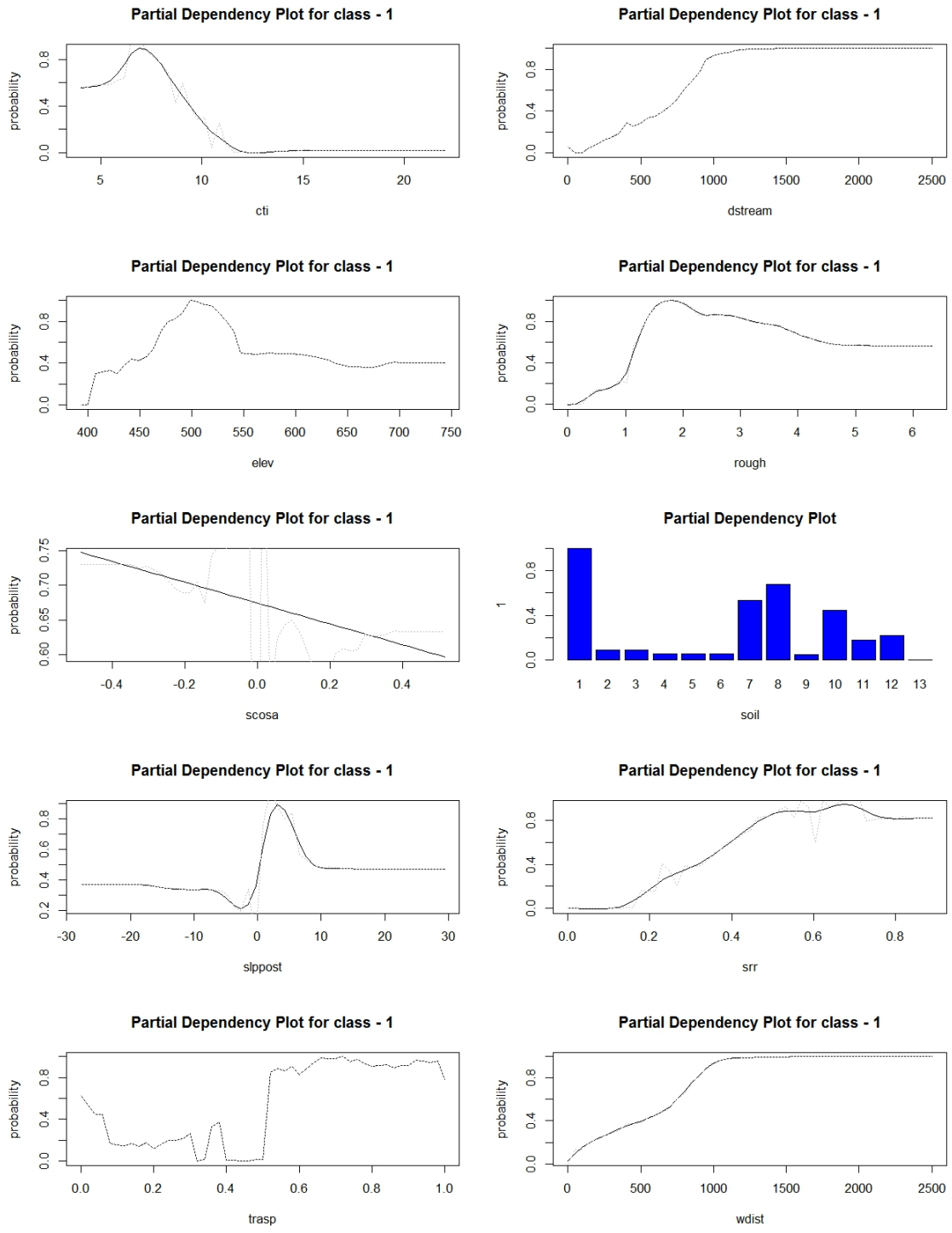


**Figure 8.** Partial plot of the covariates and the presence of cerradão (in dotted grey) and the conditional density plot of presence/absence of cerradao based on Bicubic spline (in black) (with exception of soil, which is show in binned median distance for categorical classes).

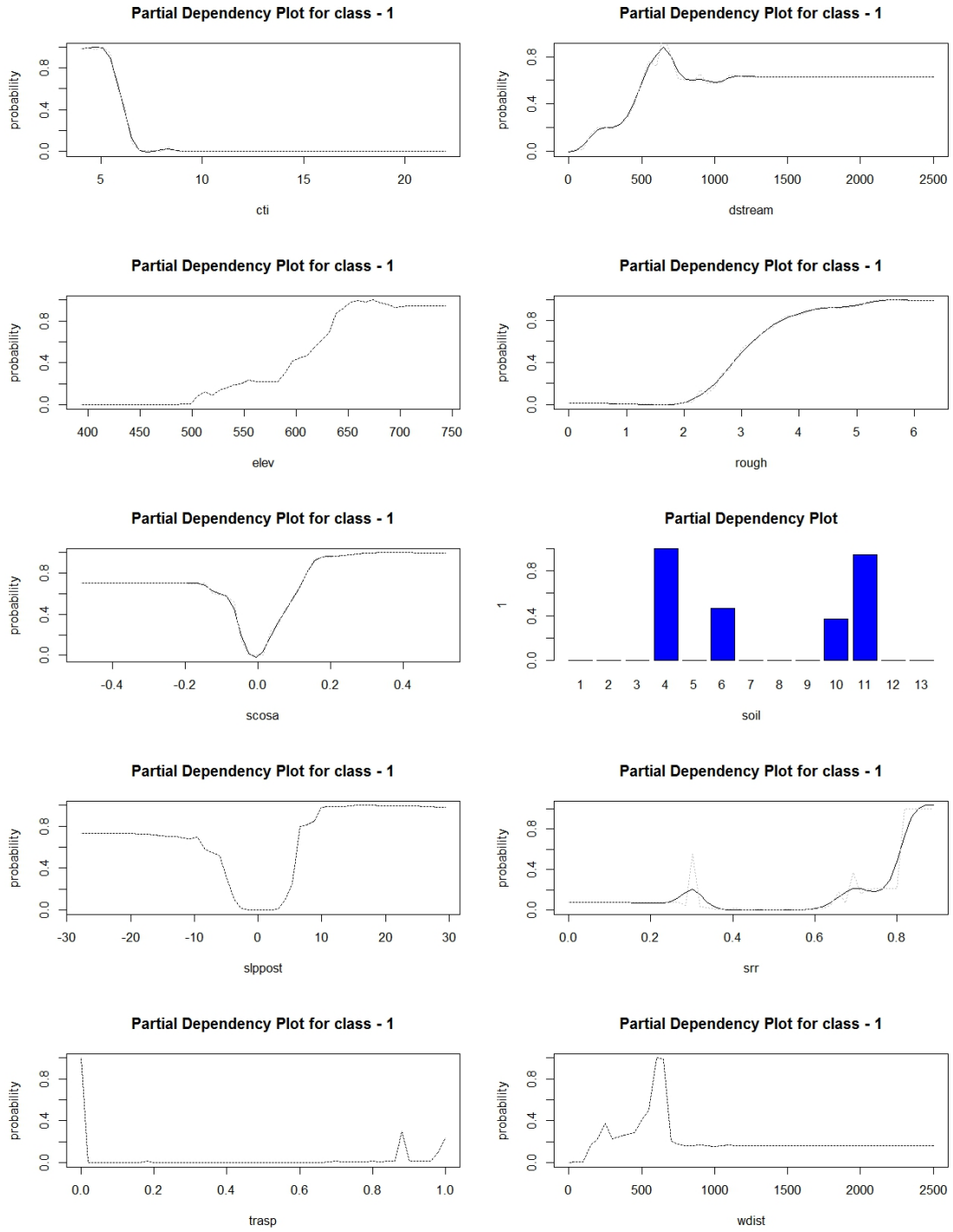


**Figure 9.** Partial plot of the covariates and the presence of cerrado (in dotted grey) and the conditional density plot of presence/absence of cerrado based on Bicubic spline (in black) (with exception of soil, which is show in binned median distance for categorical classes).

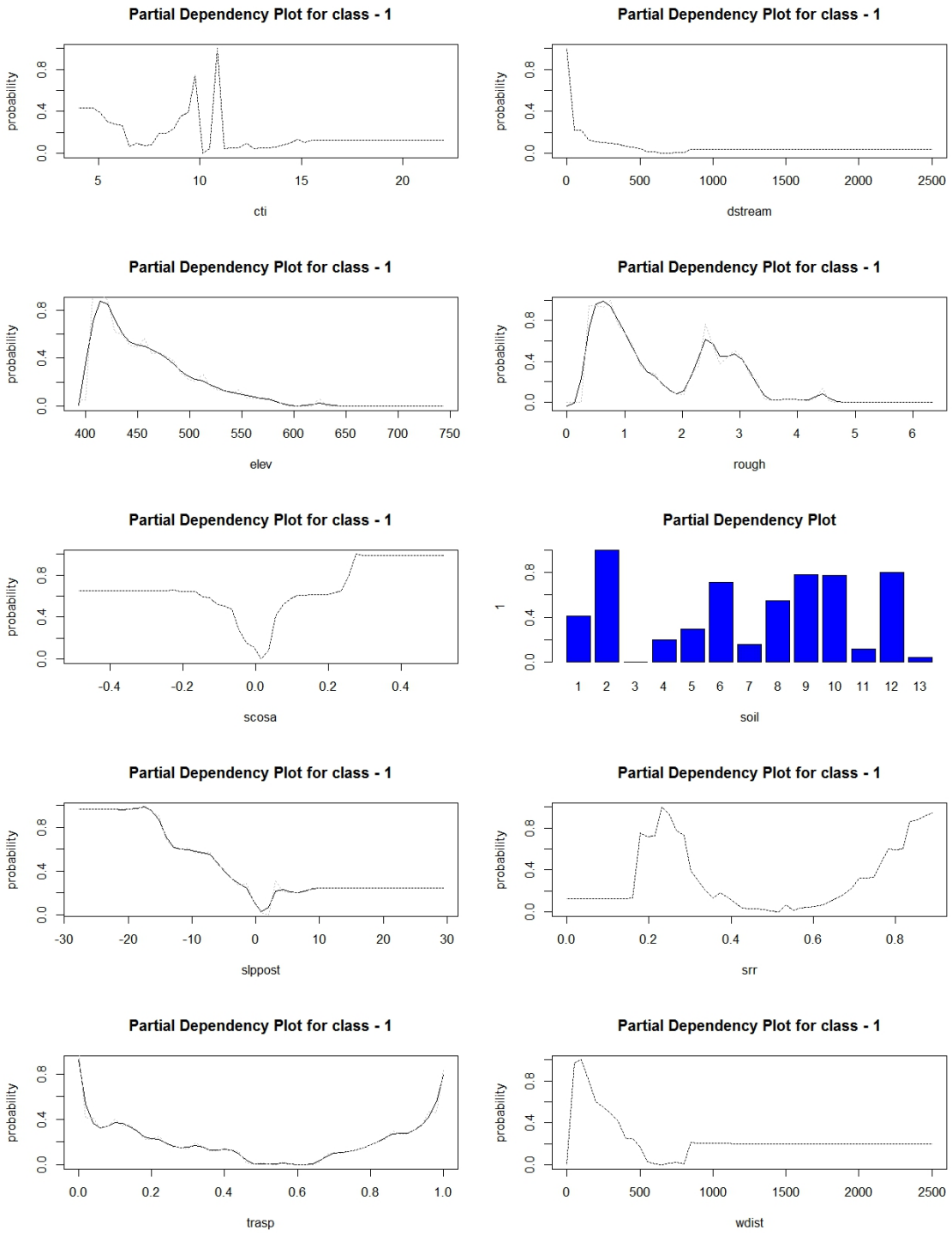


**Figure 10.** Partial plot of the covariates and the presence of riparian forest (in dotted grey) and the conditional density plot of presence/absence of riparian forest based on Bicubic spline (in black) (with exception of soil, which is show in binned median distance for categorical classes).

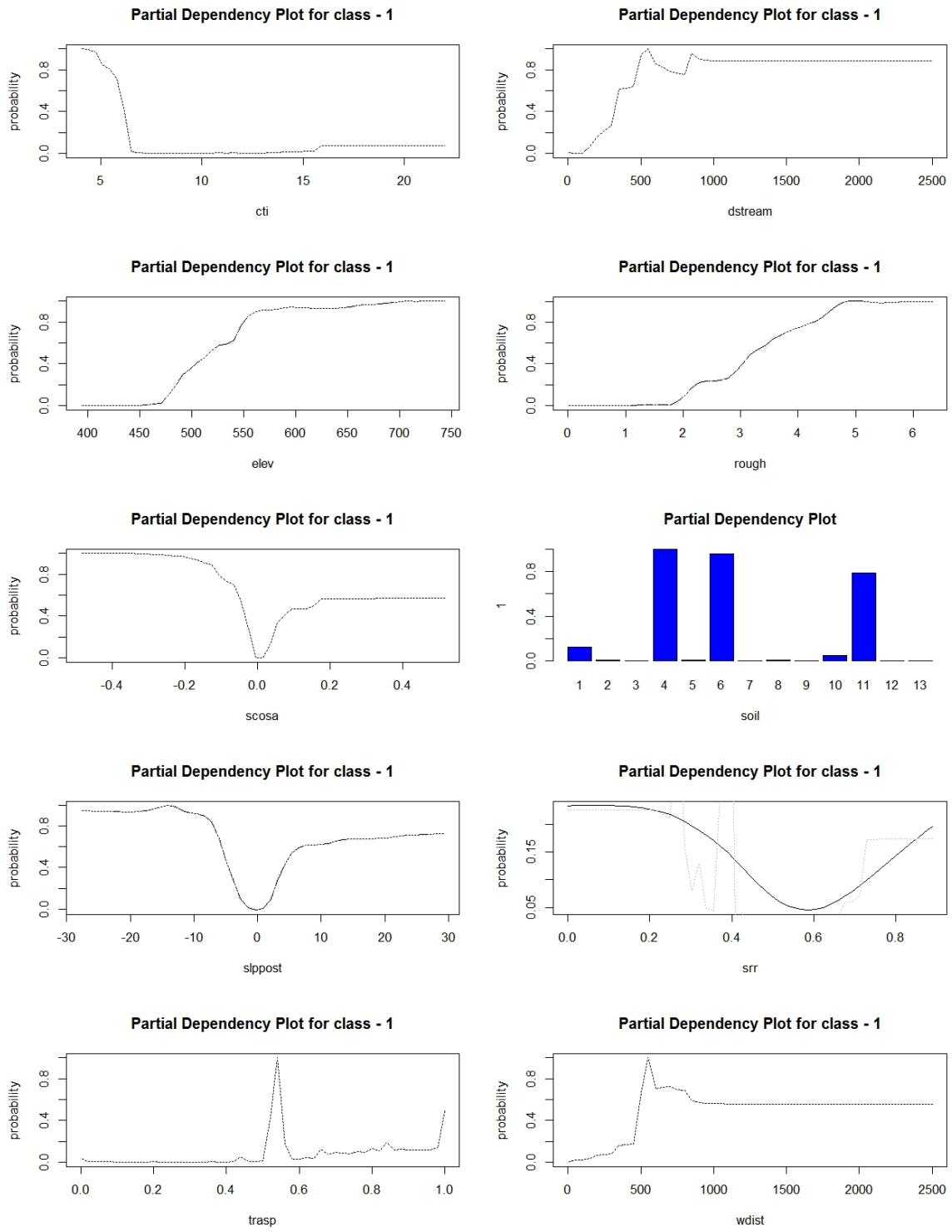


**Figure 11**. Partial plot of the covariates and the presence of semi-deciduous forest (in dotted grey) and the conditional density plot of presence/absence of semi-deciduous forest based on Bicubic spline (in black) (with exception of soil, which is show in binned median distance for categorical classes).

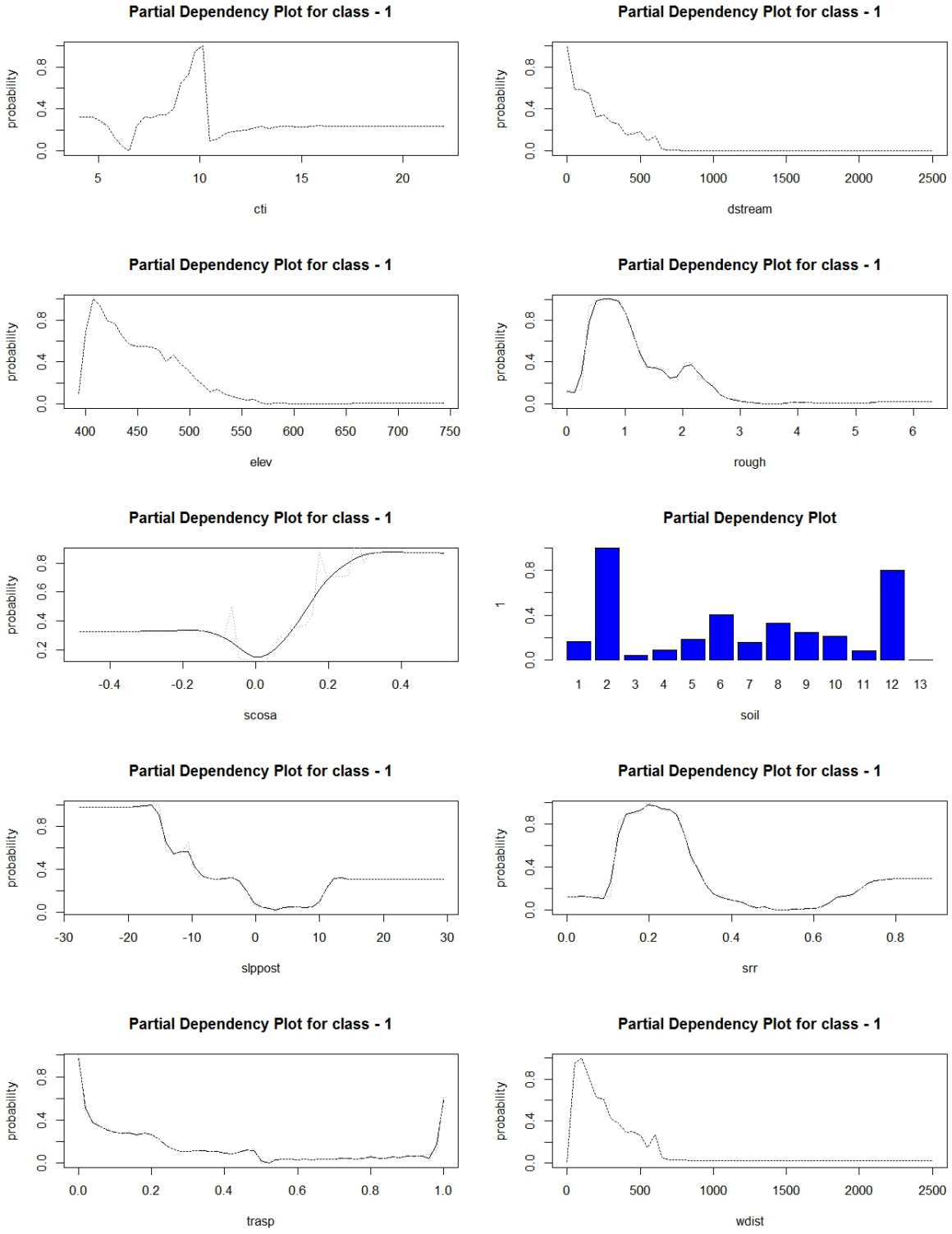


**Figure 12.** Partial plot of the covariates and the presence of wetlands (in dotted grey) and the conditional density plot of presence/absence of wetlands based on Bicubic spline (in black) (with exception of soil, which is show in binned median distance for categorical classes).

**Figure 13.** Sample map of the individual trees mapped within a pasture parcel.

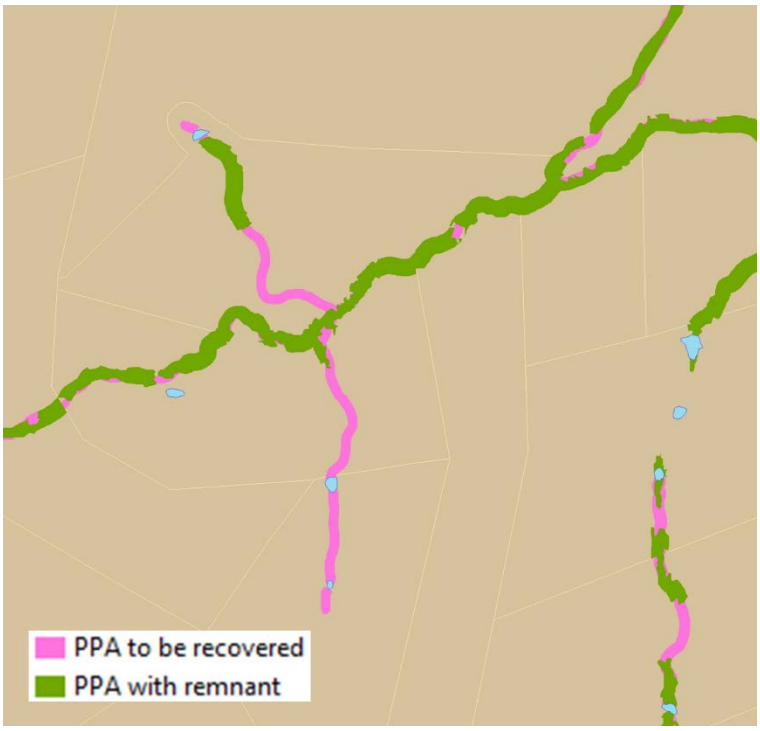


**Figure 14.** Sample map of farms showing Permanent Preservation Areas (PPAs) with habitat remnants and areas to be restored (or recovered) to achieve compliance with the Brazilian Forest Code.

## Tables

**Table 1.** Definitions of the land cover/land use (LULC) classes.

| Domain | LULC Class | Land Cover Description |
|---|---|---|
| Natural | Cerradão (Broad leaf Forest) | Cerrado forest physiognomy with a continuous canopy at least 7 meters high and understory with few grasses. Differ from other forest because it presents a species composition typical of cerrado vegetation (Oliveira-Filho, 2006). |
| Natural | Cerrado (Shrubland) | Cerrado physiognomies with two strata, one herbaceous and other shrub-arboreal. The percentage of each stratum defines the physiognomy. The category includes three Cerrado subtypes: Campo limpo (open grasslands, with less than 10% of shrubs), Campo sujo (grasslands and shrubs, no higher than 2m) and Cerrado senso stricto (20 - 50% of shrubs and trees, no higher than 7 meters) (Oliveira-Filho, 2006). |
| Natural | Semi-deciduous Forest | Forest physiognomy with 15-25 meter high trees and dense understory. Semi-deciduous (20 to 70%). The vegetation included in this category has lower density of epiphytes and ferns when compared with ombrophylous forest and variable density of lianas and bamboo. It can occur along rivers (Riparian Forest) or in isolated patches. It includes also arboreal formation between ecotones, with different types of vegetation, and forest in advanced stage of regeneration. |
| Natural | Wetland | Grassland vegetation on wet soils may include wetlands with one species of buriti palms (*Mauritia flexuosa*). |
| Pasture | Pasture | Areas of grasses planted with African grasses (*Brachiaria* spp) that are intensively managed for livestock grazing. It includes “pasto limpo” (pasture) and “pasto sujo” (abandoned/degraded pasture). |
| Sugarcane | Sugarcane | Sugarcane crops. |
| Other Cultivated | Row Crops | Crops under annual rotation system. This category may include rice, cassava, corn, sorghum, and soybean as well as bananas and coconuts and other permanent crops that are harvested seasonally but do not require replanting each period.* |
| Other Cultivated | Eucalyptus Plantation | Primarily Eucalyptus forest, but may include rubber trees.* |
| Development | Urban Development | Areas covered by buildings and road system around villages and cities, with predominance of non-agricultural artificial surfaces. |
| Development | Industrial/Commercial | Industrial or commercial units, covered with artificial surfaces with no vegetation and located outside urban developments. This category includes airports, the SVAA plant, and farm stockyards. |
| Water | Water bodies | Areas of open water, including inland rivers, streams, and ponds. |

* Occurrence and area of the crops is an estimate based on secondary data (IBGE, agricultural data 2011) (http://www.ibge.gov.br/).

**Table 2.** Areal and percentage coverage of land use /land cover (LULC) classes.

| LULC class | Area (ha) | % Area |
|---|---|---|
| Cerrado (Shrubland) | 13,682 | 3.7 |
| Cerradão (Broad leaf Forest) | 27,120 | 7.3 |
| Semi-deciduous Forest | 9,247 | 2.5 |
| Riparian forest | 7,057 | 1.9 |
| Wetland | 15,945 | 4.3 |
| Pasture | 229,367 | 61.5 |
| Sugarcane | 31,255 | 8.4 |
| Row Crops | 268 | 0.1 |
| Eucalyptus Plantation | 345 | 0.1 |
| Urban Development | 684 | 0.2 |
| Industrial/Commercial | 282 | 0.1 |
| Water bodies | 37,790 | 10.1 |
| **All land cover classes** | 373,043 | 100.0 |

**Table 3.** Variables used in the Random Forest models to predict natural vegetation types across the study area.

| Variable | Name | Description | Citation |
|---|---|---|---|
| **cti** | Compound topographic index | Compound topographic index (deterministic wetness index) | Moore et al. (1993) |
| **dstream** | Stream distance | Euclidean distance to nearest stream | |
| **elev** | Elevation | Elevation (shuttle topographic radar mission, STRM) | Rabus et al. (2003) |
| **rough** | Roughness | Roughness of elevation (3x3 window size) | Murphy et al. (2009) |
| **scosa** | Slope*cosine (aspect) | Slope percent x cosine (rad(aspect)) | Stage (1976) |
| **slppost** | Slope position | Relative slope position. Standardized difference between elevation and mean elevation in 5x5 window. | Murphy et al. (2009) |
| **soil** | Soil | Soil type | UFV/CETEC/UFLA/FEAM (2010) |
| **srr** | Surface relief ratio | Surface relief ratio or rugosity in raster surface (3x3 window size) | Pike et al. (1971) |
| **trasp** | Topographic radaiation index | Topographic radiation index | Roberts & Cooper (1989) |
| **wdist** | Distance to water | Euclidean distance to all water bodies (streams, ponds, reservoir) | |

**Table 4.** Standard productivity by soil type (soil type classification synchronized according to Silva et al, 2011).

| Soil code | Soil type | Brazilian soil category | USDA soil category | Soil fertility category |
|---|---|---|---|---|
| 1 | CXbe8 | Cambisols (CAMBISSOLO HÁPLICO) | Inceptisols | High |
| 2 | GMd4 | Gleisols (GLEISSOLO MELÂNICO) | Entisols | High |
| 3 | LVd1 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 4 | LVd11 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 5 | LVd2 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 6 | LVd6 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 7 | LVdf1 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 8 | LVdf2 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | Medium/low |
| 9 | LVef1 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | High |
| 10 | LVef2 | Latisols (LATOSSOLOS VERMELHO) | Red latosols (oxisols) | High |
| 11 | PVAd7 | ARGISSOLO VERMELHO-AMARELO | Red-yellow ultisols | Medium/low |
| 12 | RQo1 | NEOSSOLO QUARTZARÊNICO | Quartzipsamment | Medium/low |

**Table 5.** Model validation statistics including: percent correctly classified (PCC),[1] Bootstrap error (mean percent error across n=999 model randomizations)[2], sensitivity,[3] specificity[4], precision,[5] and the area under the receiver operator curve (AUC/ROC).[6]

| Class | PCC | Bootstrap error | Sensitivity | Specificity | Precision | ROC/AUC |
|---|---|---|---|---|---|---|
| Cerradao | 0.91 | 0.09 | 0.76 | 0.96 | 0.89 | 0.86 |
| Cerrado | 0.92 | 0.07 | 0.56 | 0.96 | 0.69 | 0.77 |
| Riparian Forest | 0.88 | 0.12 | 0.26 | 0.98 | 0.68 | 0.62 |
| Semi-deciduous Forest | 0.91 | 0.09 | 0.35 | 0.98 | 0.67 | 0.67 |
| Wetland | 0.86 | 0.14 | 0.57 | 0.93 | 0.68 | 0.75 |

[1] PCC is the percent of observations that are correctly classified for a given habitat type.
[2] Bootstrap error (or out-of-bag error) is the total percent incorrectly classified for each habitat type based on the number of randomizations.
[3] Sensitivity is the percent actual positive values correctly predicted to be positive, and is complimentary to the false negative rate. It is calculated as TP/(TP + FN), where TP = true positive, FN = false negative.
[4] Specificity is the percent true negatives (0) values correctly predicted as such by the model, and is complimentary to the false positive rate. It is calculated as TN/(TN + FP), where TN = true negative, FP = false positive.
[5] Precision is the percent of retrieved positives that are relevant. It is calculated as TP/(TP + TN), where TP = true positive, TN = true negative.
[6] AUC/ROC is the area under the curve when the true positive rate is plotted against the false positive rate, or alternatively sensitivity against specificity. More information is available here: http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf.

**Table 6.** Variable Importance Measures (VIM) for each covariate in the model. Covariates with VIMs greater than 0.50 are highlighted as they had greatest influence on the discernment of each natural vegetation type.

| | Cerrado | Cerradao | Riparian Forest | Semideciduous Forest | Wetland |
|---|---|---|---|---|---|
| **cti** | 0.34 | 0.46 | 0.51 | 0.33 | 0.43 |
| **dstream** | 0.20 | 0.62 | 0.41 | 0.22 | 0.40 |
| **elev** | 1.00 | 0.84 | 1.00 | 0.91 | 1.00 |
| **rough** | 0.61 | 0.90 | 0.67 | 0.50 | 0.72 |
| **scosa** | 0.39 | 0.28 | 0.30 | 0.31 | 0.21 |
| **slppost** | 0.22 | 0.23 | 0.23 | 0.23 | 0.16 |
| **soil** | 0.59 | 0.87 | 0.47 | 1.00 | 0.50 |
| **srr** | 0.13 | 0.30 | 0.30 | 0.22 | 0.23 |
| **trasp** | 0.15 | 0.17 | 0.22 | 0.20 | 0.19 |
| **wdist** | 0.21 | 1.00 | 0.53 | 0.23 | 0.54 |

**Table 7.** Spatial requirements for Permanent Preservation Areas (PPAs) based on the Forest Code, and guidelines for compliance when lands were deforested prior to 2008.

| PPA Type | Requirement | Farm Size | Minimum Requirement |
|---|---|---|---|
| **Riparian Buffers** | Depends on river width:<br>Up to 10m = 30m<br>10 to 50m = 50m<br>50 to 200m = 100m<br>200 to 600m = 200m<br>Over 600m = 500m | Up to 30 ha<br>30 to 60 ha<br>60 to 120 ha<br>120 to 300 ha<br>Over 300 ha | 5m<br>8m<br>15m<br>20m to 100m<br>30m to 100m |
| **Spring Buffer** | 50 m | All | 15m |
| **Reservoir Buffer** | Depending on reservoir area:<br>Up to 1ha = no PPA<br>Over 1 ha = 30 a 100m | All | Area between maximum operational contour and maximum potential contour |
| **Slope > 45 degrees** | All area | All | All area (except planted forests) |
| **Mesa border Buffer**[7] | 100 m | Up to 30 ha<br>30 to 60 ha<br>60 to 120 ha<br>120 to 300 ha<br>over 300 ha | No requirement<br>No requirement<br>No requirement<br>100m (except planted forests) |
| **Wetlands** | 50 m | Up to 30 ha<br>30 to 60 ha<br>60 to 120 ha<br>120 to 300 ha<br>over 300 ha | 30 m<br>30 m<br>30 m<br>50 m<br>50 m |

**Table 8.** Summary of Forest Code compliance for the study area (in number of farms).

| | | ***Permanent Preservation Area (PPA) Compliance*** | |
|---|---|---|---|
| | | **Yes** | **No** |
| ***Legal Reserve (LR) Compliance*** | **Yes** | 254 | 658 |
| | **No** | 45 | 347 |

[7] A mesa is a term for tableland, an elevated area of land with flat top and steep slopes on sides (see http://en.wikipedia.org/wiki/Mesa).

**Table 9.** Forest Code requirements (in ha) for the study area by farm size.

| *Rural Parcel Size* | *# Parcels* | *Parcel Area (Ha)* | *Remnants (Ha)* | *PPA required(Ha)* | *Final Required LR (Ha)* | *Surplus/ Deficit (LR)* |
|---|---|---|---|---|---|---|
| ***Up to 30 ha*** | 128 | 2404.22 | 304.05 | 131.15 | 175.54 | 128.51 |
| ***30 to 60 ha*** | 205 | 9393.55 | 1090.22 | 434.57 | 719.48 | 370.74 |
| ***60 to 120 ha*** | 273 | 23742.24 | 3175.03 | 1224.73 | 2273.54 | 901.49 |
| ***120 to 300 ha*** | 398 | 78471.88 | 12384.44 | 4508.89 | 13705.07 | -1320.63 |
| ***> 300 ha*** | 300 | 190754.9 | 33221.43 | 11501.45 | 34275.32 | -1053.89 |
| ***Total*** | **1304** | **304766.79** | **50175.17** | **17800.79** | **51148.95** | **-973.78** |