A tool for prioritizing deforestation hotspots in the Brazilian Amazon

Alber Sanchez[1], Guilherme Mataveli[1], Gabriel de Oliveira[2], Michel E. D. Chaves[1,5], Ricardo Dalagnol[3], Fabien H. Wagner[3], Celso H. L. Silva-Junior[4], Luiz E. O. C. Aragão[1]

[1] Earth Observation and Geoinformatics Division
National Institute for Space Research - Brazil

[2]Department of Earth Sciences
University of South Alabama - USA

[3]Institute of Environment and Sustainability
University of California - USA

[4]Program in Biodiversity Conservation
State University of Maranhão - Brazil

[5]School of Sciences and Engineering
São Paulo State University in Tupã - Brazil

{alber.ipia,guilherme.mataveli,luiz.aragao,michel.chaves}@inpe.br

{wagner.h.fabien,celsohlsj}@gmail.com

deoliveira@southalabama.edu,ricds@hotmail.com

Abstract. *Deforestation monitoring and control require scientific tools to increase the impact of official policies under limited resources for environmental law enforcement. In our paper "Science-based Planning Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon" we proposed an index for prioritizing deforestation areas for law enforcement. In this paper, we present an R package that contains both the data and software required to estimate our index. We expect the public and scientific community to check our proposal along with our tool or to use it as a starting point for improving or proposing creative ways to prioritize areas in the Brazilian Amazon for policy or law enforcement actions.*

Resumo. *O monitoramento e controle do desmatamento precisam de ferramentas científicas para aumentar o impacto dos recursos previstos pelas políticas oficiais, para a aplicação da lei ambiental. Em nosso artigo "Science-based Planning Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon" propusemos um índice para priorizar áreas de desmatamento, visando subsidiar as ações para aplicação da lei. Neste artigo, apresentamos um pacote do R que contém os dados e o software necessários para estimar o índice proposto. Esperamos que o público e a comunidade científica verifiquem nossa proposta, juntamente com nossa ferramenta, ou que a utilizem como ponto de partida para melhorar ou propor formas criativas de* *priorizar áreas na Amazônia brasileira para ações políticas ou para a aplicação da lei.*

1. Introduction

The deforested area in the Brazilian Amazon is still increasing and has had a positive trend since 2012 [1]. Despite deforestation reduction promises by different administrations, no administration has achieved zero illegal deforestation [Pereira et al. 2019].

Deforestation policies and their enforcement are subject to government changes, challenging the establishment of long-term environmental planning. Such policies should be subject to public scrutiny and preferably based on scientific principles. Therefore, we published a paper entitled *Science-based planning can support law enforcement actions to curb deforestation in the Brazilian Amazon* [Mataveli et al. 2022]. In our paper, we proposed an index for prioritizing areas in the Brazilian Amazon for law enforcement actions. This index is based on a set of variables observed during the previous years and aggregated into a regular grid of 25x25 km. This index was estimated from 2019 to 2022 using the Random Forest algorithm and has been updated for 2023 [Mataveli et al. 2023].

To ensure the openness and transparency of our proposal, we prepared a data and software bundle using the *R* language (hereby called package) that allows the public and other research teams to reproduce our methods and findings. In this paper, we introduce the computational details of the development of the software used to produce our original paper [Mataveli et al. 2022].

2. Computing environment

R is a programing (scripting) language for statistical computing and graphics [R Core Team 2022]. Its source code is open, it runs on the most popular operating systems (GNU/Linux, MacOS, Windows), and it has native support for matrices, linear algebra, and statistical analysis methods [Ihaka and Gentleman 1996]. *R* is extensible through packages, which enable *R* to load and run code (C, C++, Fortran, Java, Python, or *R*), data, demos, examples, documentation, tests, and consistency checks [Wickham 2015]. *R* also counts with a centralized package repository called CRAN (The Comprehensive R Archive Network) which ensures package availability and a minimal quality level through automated testing and checking. CRAN counts with almost 20,000 packages, organized in task views, covering topics from actuarial science to Web technologies, including spatial and spatio-temporal analysis of vector and raster geographic data [Pebesma et al. 2012]. In addition to CRAN, the *R* development community is organized around scientific journals (The *R* Journal, Journal of Statistical Software), blogs (e.g. R-bloggers, The R Blog), and other organizations besides the *R* foundation (Why R foundation, Posit software, rOpenSci, among others).

3. Scientific reproducibility with R

The ability to consistently run an experiment setup and obtain similar results has been proposed for some time now and along different areas, causing

[1] Deforestation rate in Brazil's Legal Amazon `http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/rates`

some confusion regarding the wording used [Plesser 2018]. We adhere to the definitions used by the Association for Computing Machinery (ACM) badging system, which considers three definitions: repeatability, reproducibility, and replicability [Association for Computing Machinery 2022]. Repeatability refers to the ability of research teams to reliably repeat their own computations. Reproducibility means that independent research teams can obtain the same experimental results using the authors' software artifacts. Finally, replicability implies that independent research teams can obtain the same results using their own artifacts.

Given the definitions above, we argue that by using *R* packages, we achieve both repeatability and reproducibility. For example, we developed the *R* package (see Section 4) during the entire development of our manuscript, thus achieving repeatability and, by making it available online, we achieved reproducibility. Since replicability depends on other research groups collecting their own data and writing their own software, it cannot be achieved by us writing *R* packages.

4. Package description

As mentioned earlier, our *R* package allows users to reproduce the results presented in our paper [Mataveli et al. 2022] and its update [Mataveli et al. 2023]. Our package bundles both the code and the data required to prioritize deforestation areas for 2022 and 2023. Our package is available at Github[2] and has an approximate size of 6 MB (zipped), which unfortunately, disqualifies it from submission to CRAN as it rejects packages larger than 5 MB [The Comprenhensive R Archive Network 2023].

Installing our package requires, in addition to R, the package *devtools*, which allows the installation of packages from GitHub (see Code snippet 1).

```
devtools::install_github(”albhasan/prioritizedeforestationhotspots”,
                         dependencies = TRUE)
```

Code snippet 1. Install the package in R. Note that the package *devtools* is required before installation.

This package includes a function to fit the model presented in our paper (*fit_model*) and a function to estimate its accuracy (*estimate_accuracy*), which is achieved by adjusting 100 models to the data and then cross-validating them. An additional function (*results_to_shp*) applies thresholds to the results of our model into categories (e.g. low, average, and high) and exports them to a vector file compatible with Geographic Information System software. Calling these functions reproduces the results presented in our paper (see Code snippet 2).

These functions take only one parameter, the output directory (*out_dir*). After running, the functions store R data files containing the model used to estimate the prioritization index (*final_model.rds*), the model generated during each iteration of the accuracy estimation (e.g. *param_final_100.rds*), and their metrics (e.g. *performance_test_100.rds*). Comma-separated files are also generated containing a summary of the models' root mean square error (*crossvalidation_tb.csv*), the estimation produced by the final model (*new_data_tb.csv*), and an estimation of the importance of

[2] Prioritize deforestation hotspots https://github.com/albhasan/prioritizedeforestationhotspots

each variable in the model (*variable_importance*.csv). In addition, a GeoPackage file is produced (*priority_classes.gpkg*) containing the prioritization index stored as geographic data compatible with Geographic Information System software (e.g. QGIS or ArcGIS).

```
library(sf); library(prioritizedeforestationhotspots)
out_dir <- ”~/Documents/prioritize_res”
estimate_accuracy(out_dir)# NOTE: This takes long to run!
fit_model(out_dir)
results_to_shp(out_dir)
```

Code snippet 2. Reproduce the updated results presented in [Mataveli et al. 2022]. The resulting files are stored in the directory specified by the variable *out_dir*.

We ran Code snippet 2 using R 4.3.1 running GNU/Linux Ubuntu 20.04.6 (Kernel 5.15.90.1) LTS on top of Windows 10 Subsystem for Linux 1.2.5.0 using 16 of the 32 available cores in a processor Intel Xeon E5-2640 v3 2.593GHz with 32 GB of memory. The function *fit_model* took 13 minutes to run (user 9535.76, system 129.55, elapsed 752.90), *results_to_shp* took 14 hours (user 640670.17, system 10554.66, elapsed 50190.13), and *results_to_shp* took a second (user 0.91, system 0.02, elapsed 0.972).

Our package also includes the data required to run our model: *deforestation_data* and *deforestation_grid*. The former contains the model variables aggregated at 25 km resolution; the latter is the grid itself stored using R's vector format (an object of the *sf* package [Pebesma 2018]). In addition, our pre-computed results are available as variables. Code snippet 3 shows how to format and plot these results, as shown in Figure 1.

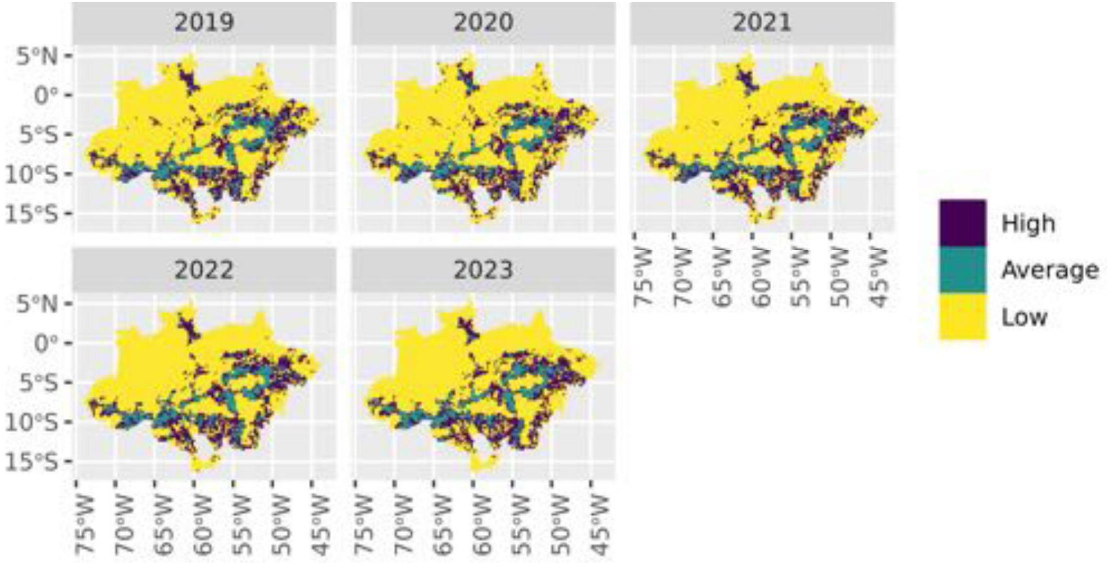

Figure 1. Plotting prioritization data stored in the package. This figure is the result of running Code snippet 3.

Additional variables for each cell in the grid include an identifier (*id*); a reference year (*ref_year*); the number of fires the year before the reference year (*active_fires_ly*; the area of deforestation during the reference year (*deforestation* in km2) and during one, two, and four years before the reference year (*def_1_ly*, *def_2_ly*, and *def_4_ly* in km2); and indigenous or protected areas (*area_PA* in km2). In addition, the distances to the closest waterway (*dist_hydro* in km), closest highway (*dist_road* in km), closest highway or waterway (*dist_road_hidro* in km), and the closest grid centroid with more than 1% and

2% deforestation one year before the reference (*dist_1_percent_ly* and *dist_2_percent_ly*, both in km) are made available. Code snippet 4 shows how to plot one of these variables.

```r
library(prioritizedeforestationhotspots)
library(tidyverse); library(sf)

# Read the result data from the package.
priority_sf<-system.file("extdata", "results", "priority_classes.shp",
                         package = "prioritizedeforestationhotspots") %>%
    read_sf()

# Format the data.
priority_tb <- priority_sf %>%
    st_drop_geometry() %>%
    pivot_longer(cols = starts_with("pri"), names_prefix = "pri",
                 names_to = "ref_year", values_to = "priority")

# Arrange data into a sf object.
priority_sf <- priority_sf %>%
    select(id) %>%
    right_join(priority_tb, by = "id", multiple = "all") %>%
    mutate(priority = factor(priority, ordered = TRUE),
                        labels = c("High", "Average", "Low"))

# Plot.
priority_sf %>%
    ggplot() +
    geom_sf(aes(fill = priority), lwd = 0) +
    facet_wrap(~ref_year) +
    theme(axis.text.x = element_text(angle = 90)) +
    theme(legend.title=element_blank())
```

Code snippet 3. Plot the results already stored in the package.

```r
library(prioritizedeforestationhotspots)
library(tidyverse); library(sf)

deforestation_grid %>%
    right_join(deforestation_data, by = "id") %>%
    ggplot() +
    geom_sf(aes(fill = area_PA), lwd = 0) +
    scale_fill_gradient(name = "Area (km2)", trans = "log",
                        breaks = c(1, 10, 100, 600),
                        low = "green", high = "red") +
    theme(axis.text.x = element_text(angle = 90))
```

Code snippet 4. Plot the extent of protected areas or indigenous lands in each cell in the grid.

5. Final remarks

We presented the R package *prioritizedeforestationhotspots* [3], which enables users to reproduce the results presented in the paper “*Science-based Planning*

[3] Prioritize deforestation hotspots https://github.com/albhasan/prioritizedeforestationhotspots

Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon" [Mataveli et al. 2022, Mataveli et al. 2023]. This tool comprises not only the software but also the data used during the writing and analysis stages of the aforementioned paper. In this way, we provide other research teams with the opportunity to check our conclusions and the potential to start extending our research to cover new hypotheses.

References

Association for Computing Machinery (2022). Artifact review and badging. `https://www.acm.org/publications/policies/artifact-review-and-badging-current`. Accessed: 2023-09-18.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299.

Mataveli, G., de Oliveira, G., Chaves, M. E. D., Dalagnol, R., Wagner, F. H., Ipia, A. H. S., Silva-Junior, C. H. L., and Aragão, L. E. O. C. (2022). Science-based planning can support law enforcement actions to curb deforestation in the Brazilian Amazon. *Conservation Letters*.

Mataveli, G. A. V., Oliveira, G. d., Chaves, M. E. D., Silva, R. D. d., Wagner, F. H., Sanchez Ipia, A. H., Silva-Junior, C. H. L., Dutra, D. J., and Aragão, L. E. O. e. C. d. (2023). *Determinação de áreas prioritárias para o combate ao desmatamento na Amazônia em 2023*. Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439.

Pebesma, E., Nüst, D., and Bivand, R. (2012). The R software environment in reproducible geoscientific research. *Eos. Commentarii Societatis philologae Polonorum*, 93(16):163.

Pereira, E. J. d. A. L., Silveira Ferreira, P. J., De Santana Ribeiro, L. C., Sabadini Carvalho, T., and De Barros Pereira, H. B. (2019). Policy in Brazil (2016–2019) threaten conservation of the Amazon rainforest. *Environmental Science & Policy*, 100:8–12.

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11:76.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

The Comprenhensive R Archive Network (2023). Cran repository policy. `https://cran.r-project.org/web/packages/policies.html`. Accessed: 2023-09-18.

Wickham, H. (2015). *R Packages*. O'Reilly Media, Sebastopol, CA, first edition edition.