ISSN 2763-6860

# Boletim

## TÉCNICO SIF

Número 06 - Volume 03
Junho 2023

# A CONTRIBUIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO R PARA O AVANÇO DO SETOR FLORESTAL

*Écio Souza Diniz et. al.*

SIF UFV Universidade Federal de Viçosa

**ISSN 2763-6860**

# A CONTRIBUIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO R PARA O AVANÇO DO SETOR FLORESTAL

Écio Souza Diniz[2*], Alexandre Simões Lorenzon[3] , Ernani Lopes Possato[3] e Diogo Nepomuceno Cosenza[3]

[2] Universidade Federal de Viçosa, Programa de Pós-Graduação em Ciência Florestal, Viçosa, MG - Brasil. E-mail: <eciodiniz@gmail.com>.
[3] Universidade Federal de Viçosa, Departamento de Engenharia Florestal, Viçosa, MG - Brasil.. E-mail: <alexandre.lorenzon@ufv.br> , <ernani.possato@ufv.br> e <diogo.cosenza@ufv.br>.
*Corresponding author.

RESUMO – *O avanço das tecnologias digitais nas últimas décadas, em grande parte fomentado pelo aprimoramento de linguagens de programação, nos tem permitido o acesso a um grande volume de dados, contribuindo para tomadas de decisões mais eficientes e precisas. A linguagem computacional do software R é a mais utilizada em variadas aplicações em análises de dados, especialmente para geração de modelos de previsão. Dentre os principais fatores responsáveis pela popularidade da linguagem R estão o oferecimento de todos os seus recursos disponibilizados gratuitamente, interface de usuário amigável com o RStudio, fornecimento de código aberto e de livre acesso e uma sólida e ampla rede de suporte fornecida em fóruns digitais. A tendência do uso do R no setor florestal também tem aumentado exponencialmente, o que é comprovado pelo crescente número de pacotes em R voltados para pesquisas florestais. Esses pacotes incluem desde funções para análises ecológicas de comunidades florestais até o sensoriamento remoto para fins de inventário florestal. Neste texto, nós abordamos a importância do R para o avanço do setor florestal e fornecemos exemplos de aplicações práticas.*

*Palavras-Chave*: *Linguagem R; Análise de dados; Modelos preditivos; Pesquisa Florestal.*

## 1. INTRODUÇÃO

As últimas décadas foram marcadas por avanços surpreendentes das tecnologias digitais, que possibilitaram a dinamização da produção de conhecimento e o acesso a um enorme volume de dados para apoiar as tomadas de decisões em diferentes setores econômicos. Essa revolução digital é conhecida como a era dos grandes dados (*big data*). Nas Ciências Florestais, essas mudanças ficaram evidentes pela forte demanda por informações cada vez mais precisas. Atualmente, os analistas florestais necessitam combinar dados de diversas naturezas para permitir o correto diagnóstico dos fenômenos. Alguns exemplos destes dados são aqueles de natureza ecológica, climática, geográfica, biométrica, social e econômica. Entretanto, a manipulação destes dados requer ferramentas robustas de análise, que permitam combinar as múltiplas informações de maneira rápida, automatizada e intuitiva. Muitos programas computacionais foram desenvolvidos para este propósito. Entretanto, as situações enfrentadas pelos analistas florestais são diversas, requerendo programas que lhes garantam maior flexibilidade de análise. Assim, programas tradicionais como as famosas planilhas eletrônicas e outros pacotes estatísticos consolidados no mercado passaram a ser substituídos por programas gratuitos operados por linguagens de programação.

Dentre os programas mais utilizados no mundo para a análise de dados está o software R, um ambiente de programação criado nos anos 1990 que implementa a linguagem de programação S, desenvolvida nos anos 1970 pelo estatístico e cientista de dados Dr. John Chambers (IHAKA 2017; BECKER 2018), atualmente da Universidade de Stanford, EUA. Inicialmente o R foi criado para ser uma ferramenta estatística, mas há quase 20.000 pacotes destinados a analisar dados de diversos tipos, como informações geográficas e de sensoriamento remoto, dados florestais, econômicos, etc. Estes pacotes são disponibilizados gratuitamente no CRAN (*The Comprehensive R Archive Network*), o repositório oficial do R, e representam conjuntos integrados de códigos, dados e documentações de uso.

Até o início de 2022, já havia 105 pacotes baseados na linguagem de programação do software R destinados às aplicações florestais (ATKINS et al. 2022). Dentre os pacotes mais famosos estão o *vegan* para análise ecológica de florestas, e *lidR* para análise de dados de sensoriamento remoto, especificamente do LiDAR, e *forestmangr* e *forestinventory* para análises de dados de inventário florestal. Outros pacotes relevantes são relacionados à manipulação de tabelas (ex., *data.table*), organização e combinação de dados (ex., *dplyr*), e modelagem de dados com base em inteligência artificial (ex., *caret*). Além disso, existe a possibilidade de usar pacotes como o *doParallel* e *future* que permitem realizar processamento em paralelo, isto é, que aproveitam todo o potencial dos computadores para acelerar a execução das análises. Assim, o R permite ao usuário solucionar problemas corriqueiros e específicos com muita agilidade e flexibilidade, unindo uma gama de ferramentas modernas (MOUNT e ZUMEL 2019).

Alguns fatores podem explicar a popularidade do R quando comparada a outras linguagens, como o C++ e o Java, por exemplo, a interface de usuário amigável Rstudio (CAMPBELL 2019). Há também uma rede ampla de usuários ativos e engajados que se comunicam em fóruns digitais para trocar informações e tirar dúvidas, facilitando o aprendizado dos iniciantes. Exemplos desses fóruns são o Stack Exchange, Stack Overflow, R-bloggers, R-community e Posit Community. Outra grande vantagem do R frente aos demais programas de análise de dados é a possibilidade de verificar o código dos pacotes para saber como os dados são processados. Assim, não existem as famosas "caixas pretas", comuns em muitos programas pagos.

Neste texto é abordada uma visão geral da importância do uso do R e como ele pode ser utilizado para a análise de dados florestais e, consequentemente, para o aumento da democratização do conhecimento. São apresentados também exemplos de aplicações práticas envolvendo o R no setor florestal.

## 2. PACOTES EM R PARA PESQUISAS FLORESTAIS

O primeiro pacote em R que um usuário pode empregar para atender suas necessidades em pesquisas florestais é o *ForestAnalysisInR* (Atkins et al. 2022), que fornece recomendações de pacotes e tutoriais que atendam às necessidades do usuário. Na Tabela 1 há uma lista composta de informações extraídas de Atkins et al. (2022) com alguns dos pacotes mais utilizados

para análises de dados florestais e, adicionalmente, nós indicamos as áreas de pesquisa com as quais os usos desses pacotes são mais comumente relacionados.

Dentre os pacotes citados na Tabela 1, os pacotes "forestmangr", "ForestFit", "ForestGapR", "lidR", "phenocamr", "raster", "rgdal" e "vegan" são rotineiramente utilizados por usuários do R do setor florestal. O pacote "vegan" fornece funções para calcular índices de diversidade (Ex: Simpson, Shannon) e estimar curvas espécie-área para extrapolação do aumento do número de espécies com o aumento do esforço amostral, além de diversos métodos de análise multivariada (Ex: Clustering, PCA, CCA, RDA), utilizados para avaliar a formação de grupos florestais, gradientes ambientais e suas relações com distribuições de espécies de plantas. O "forestmangr" oferece funções para calcular parâmetros comuns em inventário florestal (Ex: classes de diâmetro das árvores, área basal, dominância, volume de madeira, estratificação vertical) derivado de diferentes tipos de amostragem (Ex: método de parcelas). Já o pacote "ForestFit" fornece funcionalidades para simular modelos de crescimento florestal a partir de estimativas para distribuições de tamanhos (Ex: crescimento em diâmetro) de populações de plantas.

Os dados que retratam mudanças fenológicas no dossel dos estratos superiores das florestas registrados temporalmente em diferentes estações por ferramentas de sensoriamento remoto podem ser analisados usando o pacote "phenocamr". Esse pacote possui funcionalidades que retornam com precisão a avaliação estatística dos períodos exatos de transição que geram mudanças fenológicas entre as estações. Nesse contexto de sensoriamento remoto, o pacote "raster" fornece funções eficientes para importação, leitura e manipulação de arquivos matriciais (raster) georreferenciados, enquanto o pacote "rgdal" provém o usuário com ferramentas de geoprocessamento. Quando os dados georreferenciados representam nuvens de pontos obtidas com técnicas de sensoriamento remoto, como o Lidar, o pacote "lidr" oferece as funções necessárias para ler e manusear (Ex: normalizar dados) tais dados. Assim, o "lidr" é um dos pacotes fundamentais para acessar e usar dados Lidar de inventário florestal remoto no ambiente R. Também utilizando dados obtidos via Lidar, o pacote "ForestGapR" permite avaliar a dinâmica de formação de clareiras florestais (antrópicas ou naturais) através da mensuração de suas dimensões e padrões de distribuição espacial.

***Tabela 1*** *– Exemplos de pacotes do R comumente utilizados em pesquisas florestais.*

| Tópico de pesquisa | Área de esquisa | Pacotes em R |
|---|---|---|
| Análise de comunidades | Ecologia | vegan, CommEcol, FD, BiodiversityR, ecolTest |
| Dendrologia | Manejo, Incêndios, Dendrocronologia | burnr, dplR |
| Inventário, Mensuração e fitossociologia | Manejo, Ecologia, Restauração | forestmangr, rFIA, fgeo, BIEN, |
| Modelagens e simulações | Manejo, Ecologia, Economia, Restauração | forestfit, Fgmutils, forestecology, modEva |
| Fenologia | Manejo, Ecologia, Monitoramento e Restauração | phenopix, phenocamr, phenor |
| Hidrologia | Hidrologia | ecohydrology |
| Geoprocessamento (SIG e sensoriamento remoto) | Manejo, Ecologia, Economia, Hidrologia, Incêndios , Monitoramento e Restauração | npphen, phenopix, phenomap,AMAPVox,canopyLazR, ecochange, forestr, ForestGapR, foster, rGEDI, lidR, rLiDAR, Sky, hyperspec, hsad, raster, rgdal |
| Sistemática e Taxonomia | Ecologia e Restauração | taxa, Taxonstand |

*Fontes: CRAN R e adaptação de Atkins et al. 2022.*

Além dos pacotes mencionados anteriormente, outros não exclusivamente desenvolvidos para pesquisas florestais, mas amplamente utilizados para esse propósito, também devem ser citados: "caret", "randomForest", "dplyr" e "ggplot2". O pacote "caret" é o mais utilizado no R para criação e execução de modelos gerados por algoritmos de aprendizagem de máquina (supervisionada e não supervisionada). Assim, o "caret" oferece uma imensa gama de algoritmos para treinar algoritmos de aprendizado supervisionado (Ex: Redes Neurais Artificiais, Random Forest - RF, Máquina de vetores de suporte – SVM) em modelos que gerem a extração de previsões a partir dos dados fornecidos (KUHN e JOHNSON 2013). Esses modelos supervisionados podem ser utilizados tanto para a abordagem estatística preditiva de modelagem de regressão quanto para uma classificação.

Na regressão, é usado um conjunto de dados que representam fatores que potencialmente preveem o comportamento de uma variável numérica (DINIZ e THIELE 2021) como, por exemplo, para estimar uma previsão de quais fatores ambientais (Ex: clima, solo, relevo) têm mais importância no crescimento horizontal de uma floresta. Já a classificação treina um conjunto de dados avaliando quais os fatores mais importantes na previsão de uma condição, fenômeno ou evento. Para isso, o usuário comumente combina dados georreferenciados, manuseados com os pacotes mencionados acima, com os algoritmos para modelos de classificação aplicados com "caret". Desta forma, é possível obter resultados de classificações como, por exemplo, classes de uso e cobertura do solo, estágios de sucessão florestal e zoneamentos (Ex: climático, produtividade, ocorrência de pragas). O pacote "randomForest" também fornece funcionalidades para calcular modelos de regressão ou classificação, mas voltado somente para a abordagem estatística de algoritmo do tipo Random Forest (RF). Embora o "randomForest" somente treine modelos do tipo RF, o que também pode ser calculado com o "caret", ele é ágil e robusto para esse propósito.

A manipulação dos diferentes tipos de dados (Ex: tabulares como as planilhas em formato data frame, matriciais, vetoriais), visando sua organização e padronização é uma etapa fundamental para inseri-los em modelos preditivos. Esses ajustes podem ser realizados em softwares de planilha de dados, como o Excel, no entanto em alguns casos essa tarefa pode-se tornar um trabalho manual, repetitivo e pouco produtivo. Além disso, softwares como o Excel comumente impõem limitação da dimensão da base de dados possível de manipulação. Para facilitar esse trabalho, o pacote "dplyr" fornece uma ampla gramática de códigos que simplifica a organização e manuseio de dados por meio de funções que executam múltiplas tarefas de manipulação da base de dados: filtragem específica de dados; geração de novas variáveis a partir de variáveis existentes nos dados; estimativas médias de variáveis por grupos ou classes em que os dados são distribuídos; seleção específica de variáveis num banco de dados; e ordenação de valores.

Outra funcionalidade utilizada pelos usuários do R está associada à visualização dos dados e resultados das análises por meio de gráficos e figuras. Para essa finalidade, o pacote "ggplot2" possui uma excelente gramática de códigos, inclusive utilizada por outros pacotes no R, que permite ao usuário gerar e ajustar gráficos e figuras de distintos formatos e customizados de acordo com a sua necessidade. Inclusive, pacotes como o "ggmap" que integra a gramática do "ggplot2" possuem estilos de gráficos que são usados para geração de mapas (KAHLE e WICKHAM, 2013), algo essencial em pesquisas florestais.

Por fim, além do "caret", "randomForest", "dplyr" e "ggplot2", outro pacote não exclusivamente desenvolvido para pesquisas florestais, mas ainda sim com bastante potencial para ser utilizado para essas finalizadas, é o "Shiny". O "Shiny" permite de forma consideravelmente simples criar aplicativos web direto do RStudio, o que fornece amplas possibilidades em pesquisas florestais para criar, por exemplo, *dashboards* e mapas interativos para demonstração de mudanças em cobertura florestal, aumento ou redução de biomassa, delimitações temporais de talhões em plantios florestais, entre outras.

## 3. EXEMPLO DE CASO EM PESQUISA FLORESTAL USANDO O R

### 3. 1 Zoneamento de risco de geada em plantios florestais

Num projeto desenvolvido pelo Departamento de Engenharia Florestal, da Universidade Federal de Viçosa (UFV), coordenado pelo professor Alexandre Simões Lorenzon (UFV) e Drª. Cibele Hummel do Amaral (University of Colorado Boulder) e

executado pelo pesquisador Dr. Écio Souza Diniz (UFV), o objetivo foi gerar um modelo de previsão automatizado a partir de aprendizado supervisionado de máquina para prever riscos de geada (Figura 1) em áreas de plantio de eucalipto (DINIZ et al., 2021). Para isso, foram utilizados dados de históricos de ocorrência de geada fornecidos juntamente com dados ambientais espacializados (altitude, declividade, latitude e longitude e distância de hidrografias) de cada área de plantio. Esses dados foram utilizados para treinar modelos de classificação supervisionado usando algoritmos do tipo Random Forest (RF), Redes Neurais e Máquina de vetores de suporte (SVM) (DINIZ et al. 2021). Então, usando o algoritmo RF, que demonstrou melhor desempenho de previsão/ classificação de ocorrência e não ocorrência de geada, foram geradas as imagens geoespacializadas classificadas em formato matricial (raster), as quais posteriormente foram reclassificadas no software ArcGIS para gerar mapas com classes de riscos (muito baixo a muito alto) de geada (Figura 1) para cada área de plantio de eucalipto.

Portanto, com a criação desse modelo automatizado via algoritmo Random Forest, una empresa do setor florestal pode mapear de forma mais precisa e eficaz as áreas com maior risco de geada e selecionar clones mais resistentes para plantio, enquanto reservando os clones menos resistentes para plantio em áreas de menor risco de geada. Ainda, nos mapas gerados para cada área de plantio é possível avaliar em quais partes do terreno o risco de ocorrência de geada é maior (Figura 1).

O automatismo de previsão de geada provido pela modelagem preditiva em R desenvolvida por Diniz et al. (2021) contribui para a maximização de produtividade de plantios e minimização de perdas financeiras na cadeia produtiva. Além disso, essa abordagem automatizada também pode ser utilizada em diversas outras situações nas Ciências Agrárias

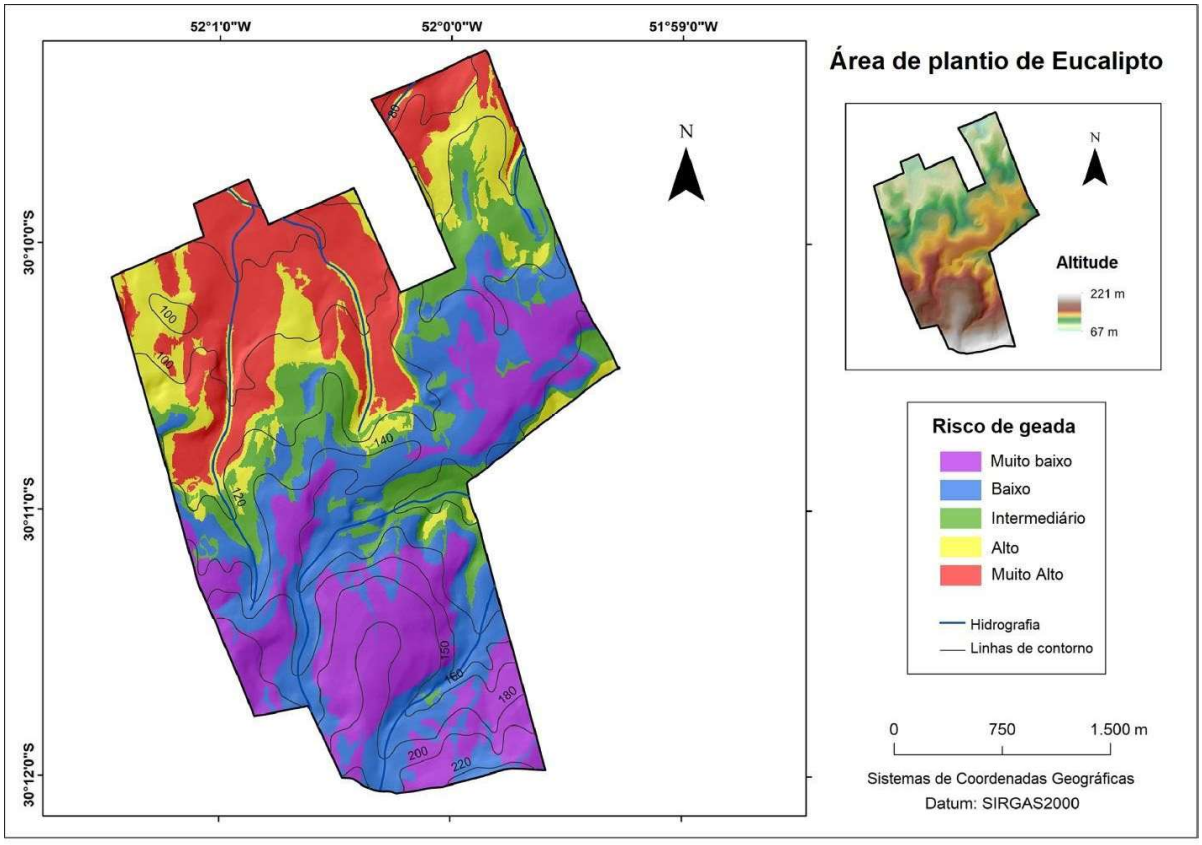


***Figura 1*** *– Classificação de probabilidade de risco de geada para área de plantio de. Risco de geada: Muito baixo (0–20%); Baixo (21–40%); Médio (41–60%), Alto (61–80%), Muito Alto (81–100%). Fonte: Adaptado de Diniz et al. (2021).*

como, por exemplo, para zoneamento climático em culturas agrícolas. Mais detalhes sobre todo o processo de coleta, organização e análise dos dados utilizados para o zoneamento de geada pode ser conferido em Diniz et al. (2021).

## 4. CONSIDERAÇÕES FINAIS

O crescente aprimoramento e constante desenvolvimento de novos pacotes em R fazem dessa linguagem uma ferramenta de grande relevância em pesquisas florestais. Portanto, nós apoiamos, incentivamos e recomendamos o aumento do investimento em capacitações profissionais no uso dessa linguagem de programação, tanto no setor acadêmico, nos cursos de Engenharia e Ciências Florestais, quanto no setor privado, por meio do suporte de empresas em treinamentos do seu corpo técnico. Na era que vivemos a necessidade de produção de conhecimento livre, aberto e replicável e a automatização de processos diversos não é só uma tendência, mas uma realidade já em ocorrência. Assim, setores de grande importância econômica como o florestal, no qual o Brasil é uma referência mundial, não devem ficar atrás, mas se alinhar às tecnologias e facilidades providas com a programação, como no caso do R, para aumentar o seu potencial produtivo.

## 5. REFERÊNCIAS BIBLIOGRÁFICAS

ATKINS, J.F.; STOVALL, A.E.L.; SILVA, C.A. Open-Source tools in R for forestry and forest ecology. Forest Ecology and Management, v. 503, n.1, p. 119813, 2022.

BECKER, R. The new S language. Boca Raton: CRC Press, 2018.

CAMPBELL, M. Learn RStudio IDE: Quick, Effective, and Productive Data Science. New York: Springer, 2019.

DINIZ, É. S. et al. Forecasting frost risk in forest plantations by the combination of spatial data and machine learning algorithms. Agricultural and Forest Meteorology, v. 306, p. 108450, 2021.

DINIZ, E.S.; THIELE, J. Modelos de regressão em R. Joinville: Clube de Autores, 2021.

IHAKA, R. The r project: A brief history and thoughts about the future. University of Auckland, v. 4, p. 22, 2017.

KAHLE, D.J.; WICKHAM, H. ggmap: spatial visualization with ggplot2. The R Journal, v. 5, n. 1, p. 144, 2013.

MOUNT, J.; ZUMEL, N. Practical data science with R. New York: Manning Publications Co, 2019.

KUHN, M.; JOHNSON, K. Applied Predictive Modeling. New York: Springer-Verlag, 2013.