Enhancing Precipitation Gap Filling in SWAT: A Comparative Analysis with Cubist Methods and Artificial Intelligence

- Josicleda Galvincio, Rodrigo Miranda, Gabrielly Luz, Verissimo Pinheiro, Magna Moura, Werônica Souza, Suzana Montenegro

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Introduction

- The quality of input data plays a fundamental role in hydrological modeling.
- In Brazil, precipitation data often present gaps and inconsistencies.
- Objective: Evaluate the accuracy of SWAT's precipitation gap-filling method compared to the Cubist method.

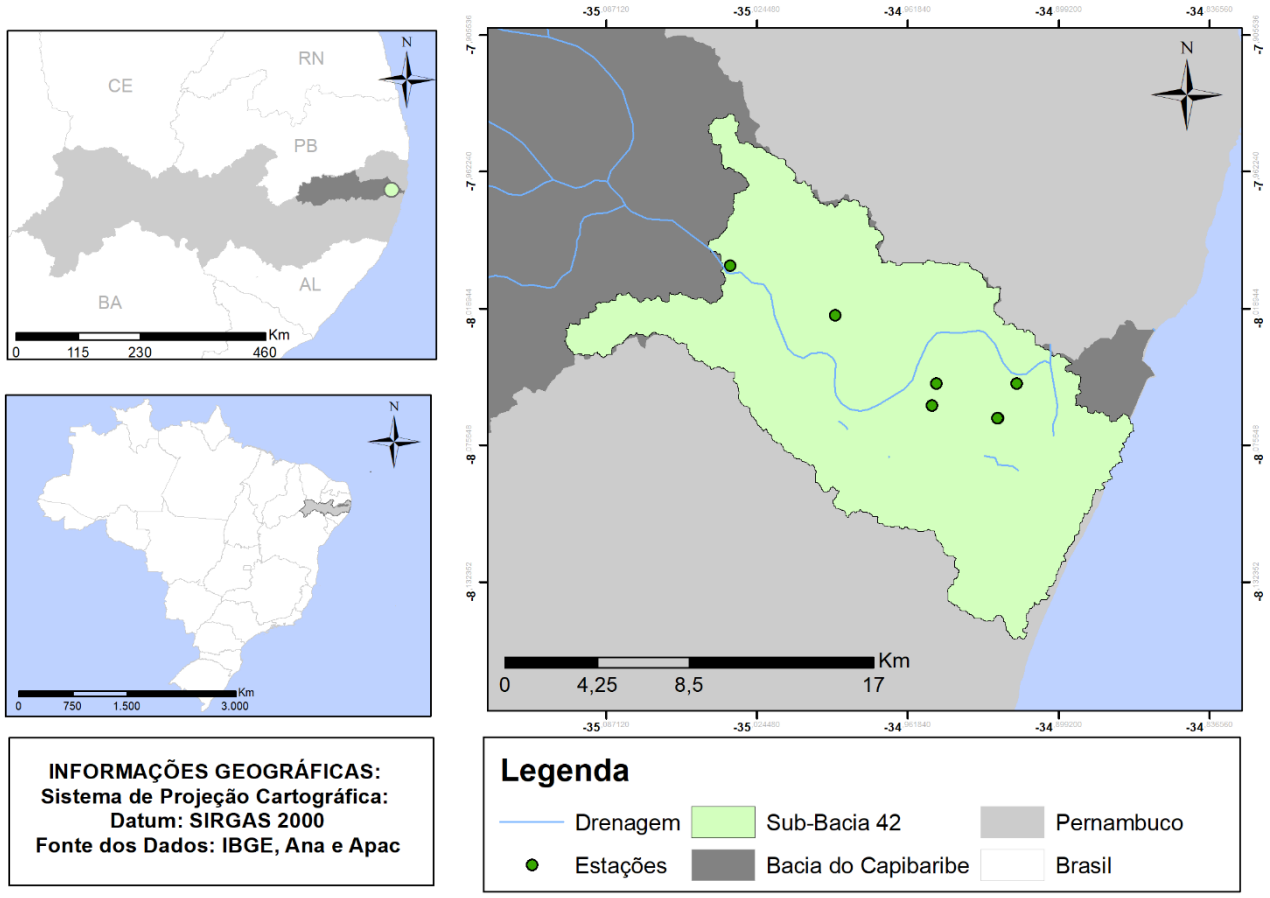
Importance of Precipitation Data

Influences accuracy and reliability of hydrological models.

Data gaps pose a challenge for climate change studies.

Essential for robust hydrological models.

Study area



The Capibaribe River basin is located at coordinates 07º 41' 20" and 08º 19' 30" S, and 34º 51' 00" and 36º 41' 58" W. The basin covers an area of 7,454.88 km^2 and encompasses 42 municipalities in the state of Pernambuco, with 15 of them entirely within its boundaries, (Apac, 2023). Due to its size, the area has various distinct physical characteristics, both hydrological and geological. Figure shows the location map of the basin and the sub-basin where the precipitation data from the stations were analyzed.

Study area

- The basin includes landscape Atlantic Forest biomes, with significant variations in relief and land use. Historically, in the Zona da Mata region, there has been intensive use for sugarcane cultivation, which has led to the suppression of Atlantic Forest vegetation in the area.

Precipitation Data Acquisition

- The rainfall data from all stations are provided on the agency's monitoring website (APAC). The information is made available through daily bulletins with rainfall values in millimeters, organized in databases by months and years of historical series. The historical series for this study covers the years 1961 to 2021. Additionally, the sub-basin selected for model calibration has seven stations near the coast. Next Table shows the code and coordinates of the studied stations. The data were organized into spreadsheets, where the amount of missing data per station within the historical series was calculated.

Precipitation Data Acquisition

Código	Latitude	Longitude
p82900	-8,05	-34,95
p30	-8,05	-34,9167
p129	-8,0011	-35,0358
p201	-8,0217	-34,9922
p263	-8,0644	-34,9244
p344	-8,0592	-34,9519
p480	-8,0644	-34,9247

Methodology

- Evaluation of SWAT and Cubist gap-filling methods.
- Data collection and analysis.
- Accuracy and effectiveness assessment.

Hydrological Modeling Tools

- SWAT: Soil Water Assessment Tool.
- Simulates water balance in a watershed.
- Internal gap-filling method for precipitation data.

The Cubist Method

- Cubist method for data gap-filling.
- Captures complex, non-linear relationships.
- Potentially more accurate than SWAT’s internal method.

Organization and Interpolation/Spatialization of the Data

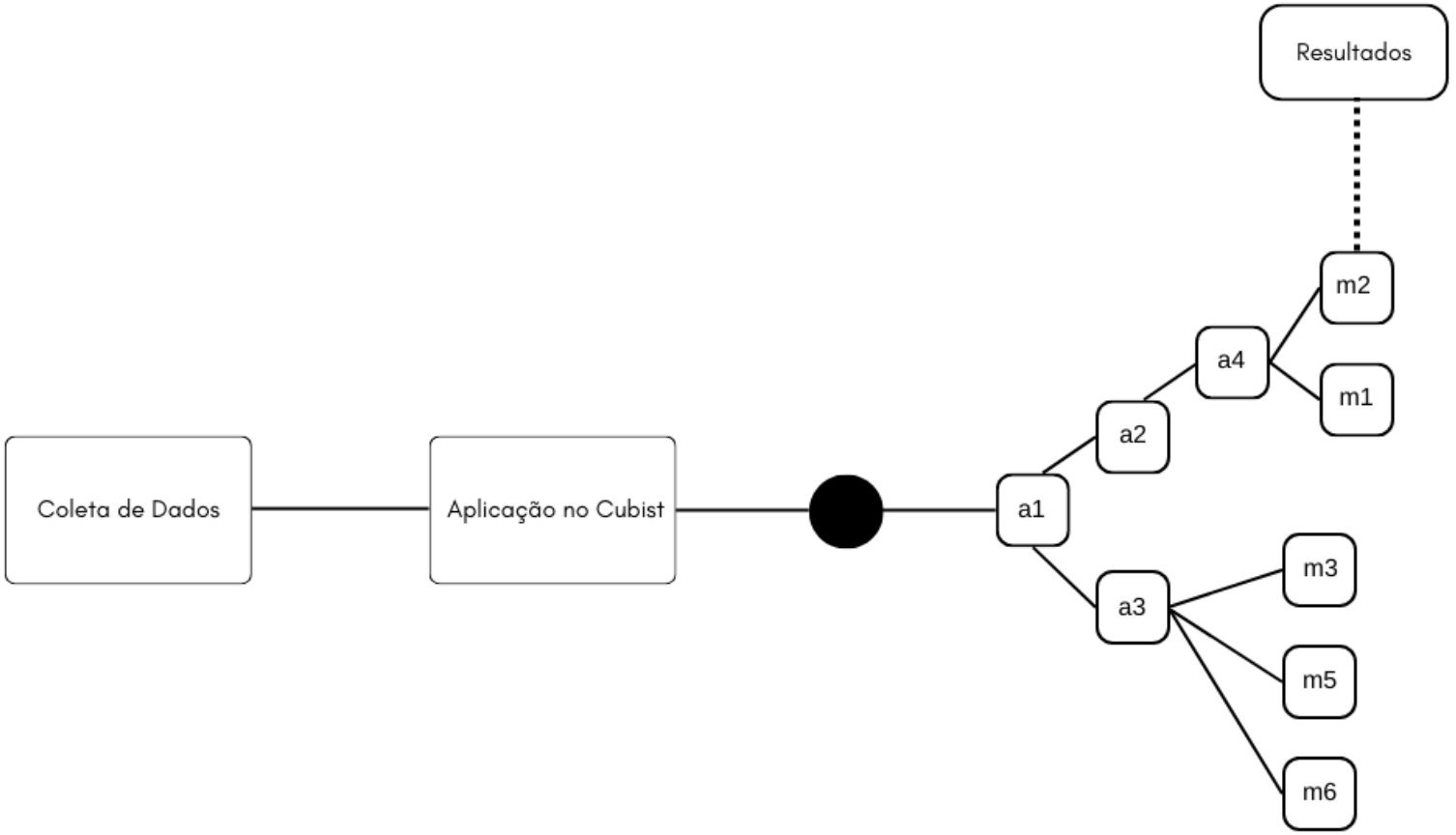
- Spatial data interpolation can be defined as the statistical projection of georeferenced datasets to obtain data representations and distributions. Data interpolation is based on Tobler's Law, which states that all factors relate to each other, but closer factors have stronger relationships than distant ones. Equation next shows the calculation used for IDW interpolation (Shepard, 1968):

$$w(x) = \sum_{i=0}^{\mathbb{N}} \frac{w_i(x)u_i}{\sum_{j=0}^{N} w}$$

Cubist and Model Calibration

- Cubist is a rule-based model where each conditional or branch of the tree constitutes a rule, and the leaves are linear regressions (Figure next). The R language package is an improvement of the M5 model published by Quinlan (1992, 1993), and it was chosen for its transparency, ease of interpretation, and high predictive power. The model allows for boosting in filling gaps by performing double iterations.
- Additionally, another innovation of Cubist is the use of the nearest neighbors methodology to adjust the predictions of the decision tree model. Thus, the seven stations of the sub-basin were used to create the model, with each station having its model generated individually to evaluate the performance of Cubist in filling gaps in each of the historical series.

Representative Flowchart of the Cubist Decision Tree



Evaluation criteria

- The series was divided into two parts: 80% for calibration and 20% for verification.
- Additionally, the Leave-one-out cross-validation method was used, in which the dataset is divided in a 1/sampling ratio. In other words, for each calibration of the model for one station, the remainder was used for calibration.
- Then, the model performances were calculated by evaluating the Correlation Coefficient, Mean Error, and Relative Error.

MAE- *Mean Absolute Error*

$$MAE(Y, Y^1) = \frac{1}{n}\sum_{i=1}^{n} |y_i - y^1{}_i|$$

Relative error

$$RE = \frac{MAE}{VR}$$

Preliminary Results-Application and Modeling of Cubist in the Sub-Basin

- Table shows the statistical performance values of the Cubist model at each station in sub-basin 42, used. Additionally, the percentages of missing data in the historical series were included to provide a complete view of the study sample. All databases contain 22,005 daily precipitation values between the years 1961 and 2021.

Not always

fewer data gaps

Estações	Coef. Correlação	MAE	RE	% de Falhas
p82900	0.64	4.27	0.5	1,19%
p129	0.51	33.57	0.69	38,44%
p201	0.86	13.83	0.27	58,48%
p263	0.56	8.60	0.58	92,39%
p344	0.75	7.37	0.34	88,27%
p480	0.92	4.67	0.12	74,71%
p30	0.56	25.45	0.65	24,98%

Tabela 4: Performance Estatística do Modelo

Preliminary Results-Application and Modeling of Cubist in the Sub-Basin

- The results obtained from the machine learning modeling on the database were promising. The correlation coefficient (r) of the model showed moderate to strong values at all stations where it was applied. The evaluation of absolute and relative errors also proved to be promising given the amount of data with which the model correlated. The use of Cubist as a methodology for data imputation appears promising in the field of data research..

Comparison of Weather Generator-SWAT and Cubist

Correlations

Control Variables			Valores_Simulados_cubist	Capibaribe PRECIP_SUPER	Valores_Observados
-none-[a]	Valores_Simulados_cubist	Correlation	1.000	-.020	.946
		Significance (2-tailed)	.	.336	.000
		df	0	2295	2295
	Capibaribe PRECIP_SUPER	Correlation	-.020	1.000	-.021
		Significance (2-tailed)	.336	.	.315
		df	2295	0	2295
	Valores_Observados	Correlation	.946	-.021	1.000
		Significance (2-tailed)	.000	.315	.
		df	2295	2295	0
Valores_Observados	Valores_Simulados_cubist	Correlation	1.000	-.001	
		Significance (2-tailed)	.	.970	
		df	0	2294	
	Capibaribe PRECIP_SUPER	Correlation	-.001	1.000	
		Significance (2-tailed)	.970	.	
		df	2294	0	

a. Cells contain zero-order (Pearson) correlations.

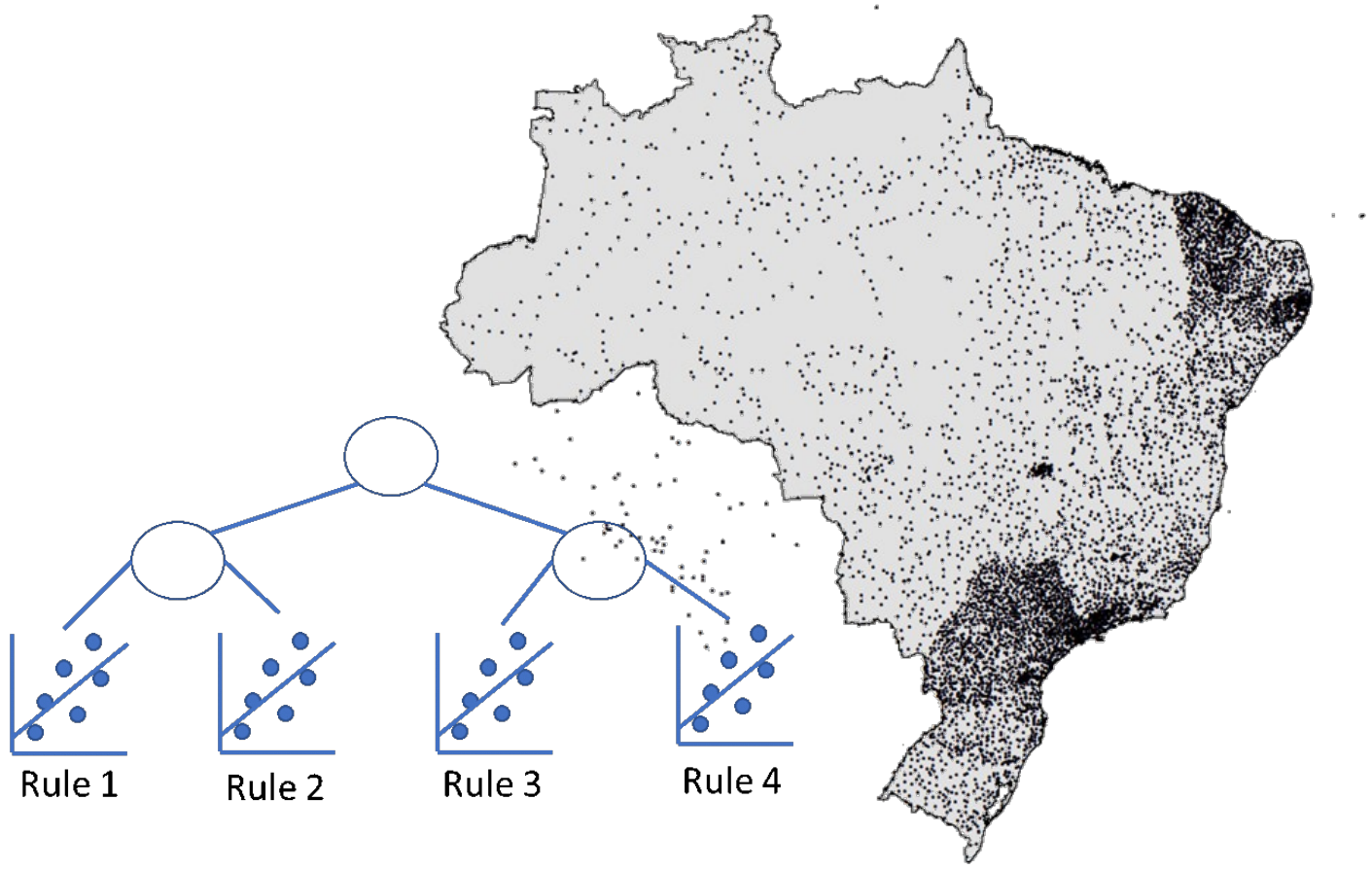
Preliminary Results-Application and Modeling of Cubist in the Sub-Basin

- Cubist method shows lower average error.
- Higher correlation coefficient (0.95) for Cubist.
- **Weather Generator-SWAT** correlation coefficient: -0.02.

Implications for SUPer System

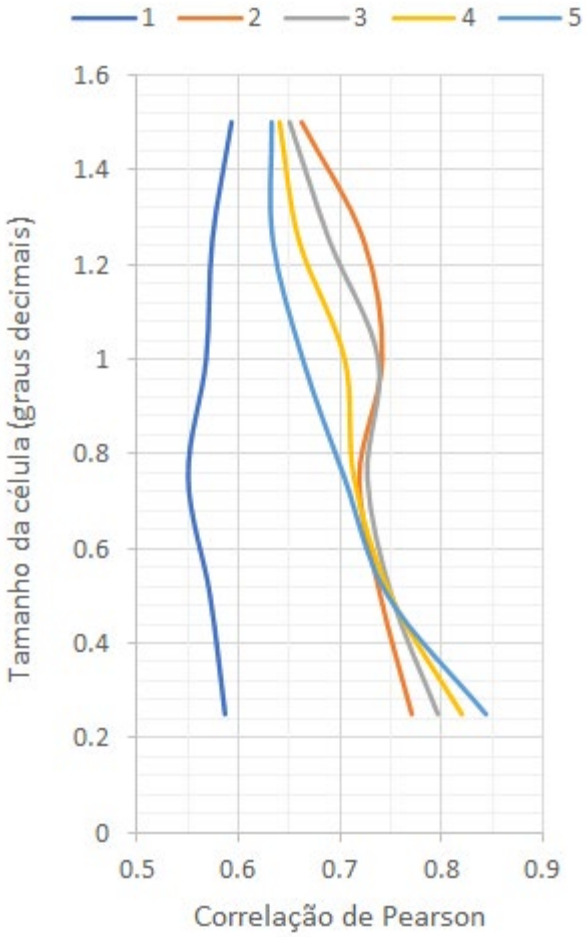
- SUPer: Hydrological Response Units System for Pernambuco.
- Benefits of implementing Cubist method.
- Improved water resource management and decision-making.

Brazil- Algorithm (*Cubist model*)

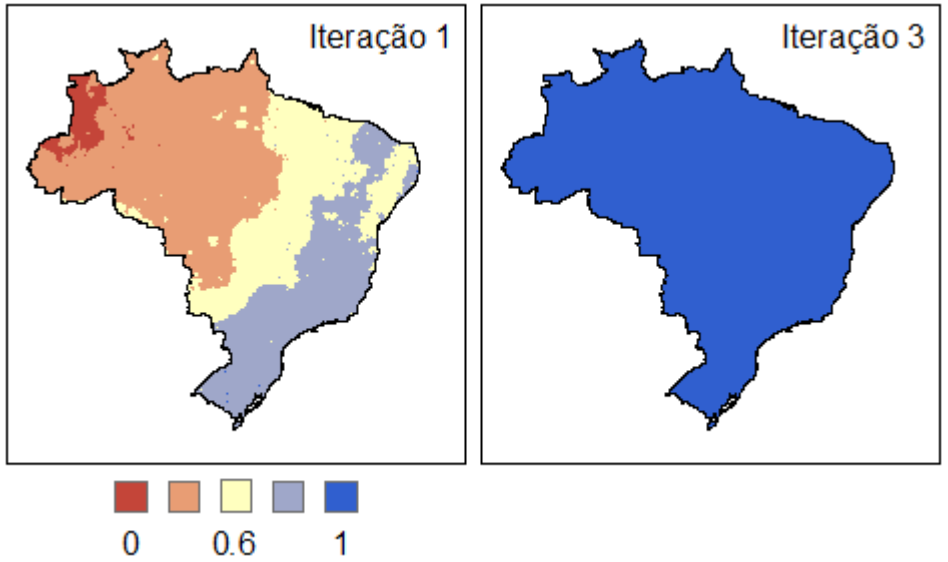


Methodological steps

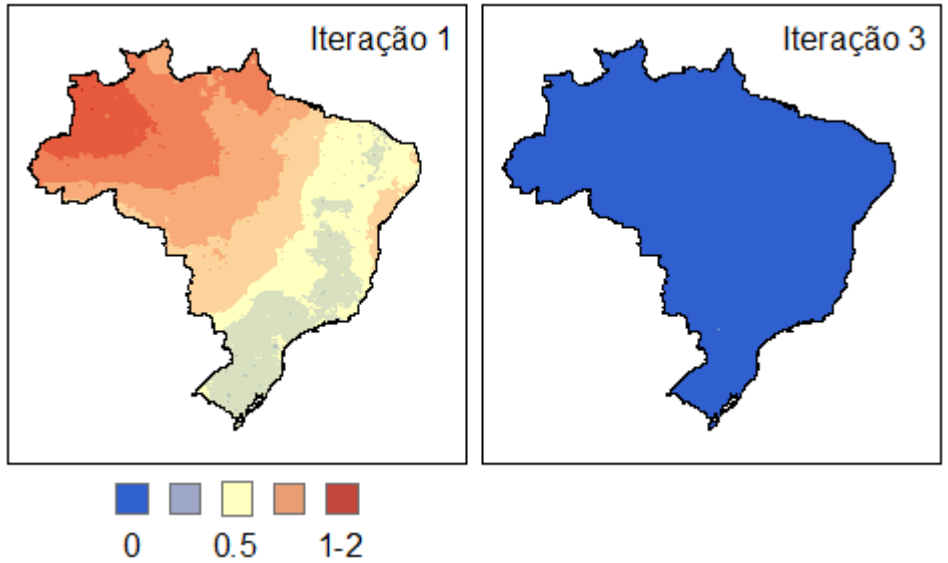
1. Iterations with Cubist
2. Model calibration
3. Trend correction
4. Aggregation Tests
5. Different cell sizes
6. Model validation
7. Formatting and availability



Correlation coefficient



Relative error in relation to the mean(%)



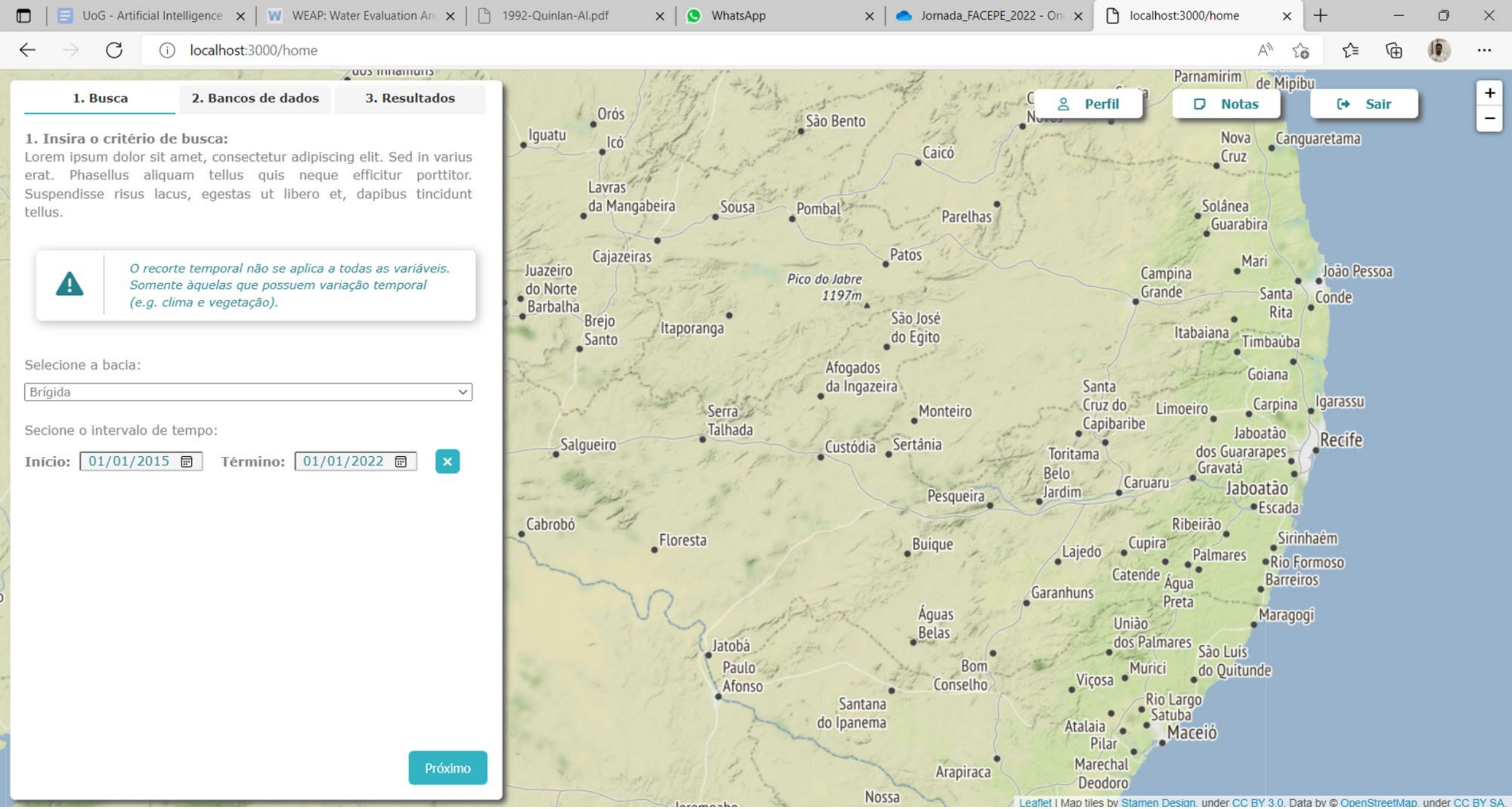
SWATech

Sobre Equipe Parceiros

E-Mail

Senha

Entrar


UoG - Artificial Intelligence
WEAP: Water Evaluation An
1992-Quinlan-AI.pdf
WhatsApp
Jornada_FACEPE_2022 - On
localhost:3000/home
localhost:3000/home
Perfil
Notas
Sair
1. Busca
2. Bancos de dados
3. Resultados
1. Insira o critério de busca:
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed in varius erat. Phasellus aliquam tellus quis neque efficitur porttitor. Suspendisse risus lacus, egestas ut libero et, dapibus tincidunt tellus.
O recorte temporal não se aplica a todas as variáveis. Somente àquelas que possuem variação temporal (e.g. clima e vegetação).
Selecione a bacia:
Brígida
Secione o intervalo de tempo:
Início: 01/01/2015
Término: 01/01/2022
Próximo
Parnamirim
de Mipibu
Orós
Iguatu
Icó
São Bento
Caicó
Nova Cruz
Canguaretama
Lavras da Mangabeira
Sousa
Pombal
Parelhas
Solânea
Guarabira
Cajazeiras
Patos
Mari
João Pessoa
Juazeiro do Norte
Barbalha
Pico do Jabre 1197m
Campina Grande
Santa Rita
Conde
Brejo Santo
Itaporanga
São José do Egito
Itabaiana
Timbaúba
Goiana
Afogados da Ingazeira
Santa Cruz do Capibaribe
Limoeiro
Carpina
Igarassu
Serra Talhada
Monteiro
Jaboatão dos Guararapes
Recife
Salgueiro
Custódia
Sertânia
Toritama
Gravatá
Belo Jardim
Caruaru
Jaboatão
Pesqueira
Escada
Cabrobó
Ribeirão
Floresta
Buíque
Lajedo
Cupira
Palmares
Sirinhaém
Rio Formoso
Catende
Água Preta
Barreiros
Garanhuns
Maragogi
Águas Belas
União dos Palmares
São Luís do Quitunde
Jatobá
Paulo Afonso
Bom Conselho
Murici
Viçosa
Santana do Ipanema
Rio Largo
Satuba
Atalaia
Maceió
Pilar
Arapiraca
Marechal Deodoro
Nossa
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under CC BY SA.

Conclusion

- Summary of key findings.
- Benefits of Cubist method for gap-filling.
- Informing water resource policies and strategies.

References

•APAC - Agência Pernambucana de Águas e Clima - Institucional. Disponível em: <https://www.apac.pe.gov.br/institucional>. Acesso em: 01 mar. 2024.

•Model Trees. Disponível em: <https://cran.r-project.org/web/packages/Cubist/vignettes/cubist.html>. Acesso em: 28 fev. 2024.

• AYOADE, J. O. **Introdução à Climatologia para os Trópicos.** 11ª edição. Rio de Janeiro: Bertrand Brasil, 2006.

• AB' SABER, Aziz Nacib. **Os domínios de natureza no Brasil: potencialidades paisagísticas.** 7ª edição. São Paulo: Ateliê Editorial, 2012.

•REBOITA, M. S.; *et al*. Entendendo o Tempo e o Clima na América do Sul. **TERRAE Didática**, online, n.8, p. 34-50, 2012.

•GARSON, G. David. (2009), **Statnotes: Topics in Multivariate Analysis.** Disponível em: http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm .Acesso em 15 de fev 2024.

•USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global | U.S. Geological Survey. Disponível em: <https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1>. 01 de mar de 2024.

•RESUMO EXECUTIVO. [s.l: s.n.]. Disponível em: <https://www.apac.pe.gov.br/images/media/1649787904_TOMO%20IV%20-%20Resumo%20Executivo.pdf>. Acesso em: 15 mar. 2024.

•Kousky E.V. 1980. Diurnal rainfall variation in northeast Brazil. Monthly Weather Review, 108:488-498.

•Kousky V.E. 1979. Frontal Influences on Northeast Brazil. Monthly Weather Review, 107:1140-1153

•ALAN CÉZAR BEZERRA et al. Annual Rainfall in Pernambuco, Brazil: Regionalities, Regimes, and Time Trends. **Revista Brasileira De Meteorologia**, v. 36, n. 3, p. 403–414, 1 set. 2021.

• COSTA, A.S.; COSTA, W.L.B.; BRAGA, C.C.; DANTAS, M.P. Temporal space variability for precipitation in the state of Pernambuco. **Journal of Hyperspectral Remote Sensing**, v. 7, n. 1, p. 1-7, 2017

•DOURADO, C.S.; OLIVEIRA, S.R.M.; AVILA, A.M.H. Análise de zonas homogêneas em séries temporais de precipitação no Estado da Bahia. **Bragantia**, v. 72, n. 2, p. 192-198, 2013.

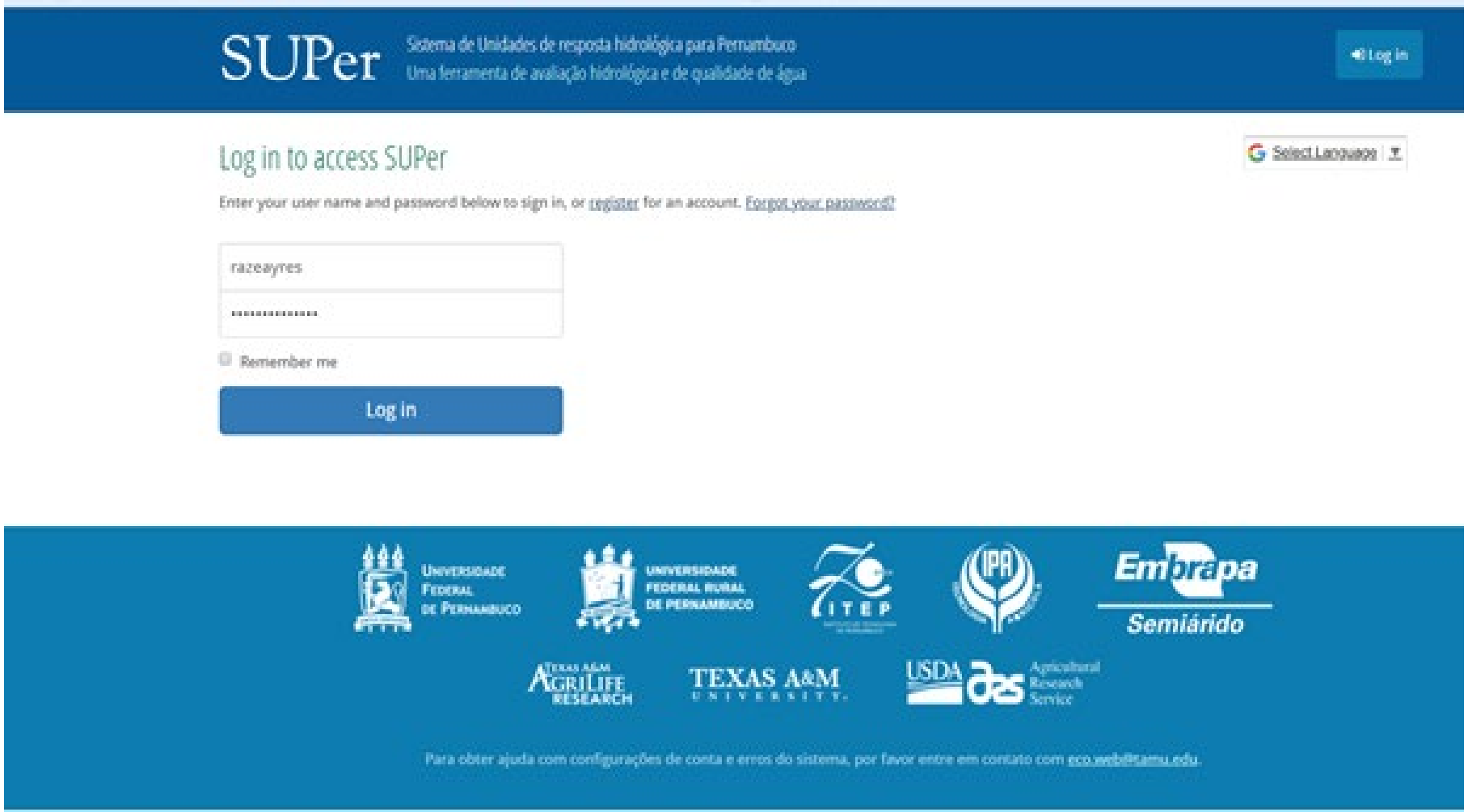
• ZHOU, J. et al. Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill **Materials. Applied sciences**, v. 9, n. 8, p. 1621–1621, 18 abr. 2019.

Thank you so much!

PROJETO:

405853/2022-0 - Desenvolvimento de um sistema automático, em tempo real, para assimilação, identificação e preenchimento de falhas da precipitação no SUPER e avaliação dos seus impactos na gestão de recursos hídricos de Pernambuco

PROJECT: 405853/2022-0 - Development of an automatic, real-time system for assimilation, identification, and filling of precipitation gaps in the SUPER system and evaluation of its impacts on water resource management in Pernambuco


SUPer
Sistema de Unidades de resposta hidrológica para Pernambuco
Uma ferramenta de avaliação hidrológica e de qualidade de água
Log in
Log in to access SUPer
Select Language
Enter your user name and password below to sign in, or register for an account. Forgot your password?
razeayres
Remember me
Log in
Universidade Federal de Pernambuco
Universidade Federal Rural de Pernambuco
ITEP
IPA
Embrapa
Semiárido
Texas A&M AgriLife Research
Texas A&M University
USDA
aes
Agricultural Research Service
Para obter ajuda com configurações de conta e erros do sistema, por favor entre em contato com eco.web@tamu.edu.

Agradecimentos

E-mail Contact

- Rodrigo Miranda e-mail: rodrigo.qmiranda@gmail.com Process Cubist model
- Verissimo e-mail: verissimo.pinheiro@ufpe.br
- Josicleda Galvincio : josicleda.galvincio@ufpe.br informations the results