ISSN: 2447-3359

REVISTA DE GEOCIÊNCIAS DO NORDESTE

Northeast Geosciences Journal

v. 11, nº 2 (2025)

https://doi.org/10.21680/2447-3359.2025v11n2ID39777

O uso de *machine learning* para detecção de descontinuidades nas séries temporais das coordenadas da RBMC

The use of machine learning for detecting descontinuities in RBMC station coordinate time series.

Alberto Luis da Silva[1]; Julio Cesar de Oliveira[2]; William Rodrigo Dal Poz[3]

[1] Universidade Federal de Viçosa/Departamento de Engenharia Civil, Viçosa/MG, Brasil. Email: alberto.silva@ufv.br
ORCID**:** https://orcid.org/0009-0008-4172-1265

[2] Universidade Federal de Viçosa/Departamento de Engenharia Civil, Viçosa/MG, Brasil. Email: oliveirajc@ufv.br
ORCID**:** https://orcid.org/0000-0003-0894-5597

[3] Universidade Federal de Viçosa/Departamento de Engenharia Civil, Viçosa/MG, Brasil. Email: william.dalpoz@ufv.br
ORCID**:** https://orcid.org/0000-0001-9532-3643

Resumo: As observações GNSS coletadas pela RBMC permitem estimar, com precisão milimétrica, as coordenadas diárias de suas estações. Como resultado, é possível gerar séries temporais robustas, capazes de descrever deslocamentos geodinâmicos, como os movimentos das placas litosféricas, bem como os efeitos das marés terrestres, oceânicas e atmosféricas. Entretanto, descontinuidades nessas séries podem indicar alterações nas coordenadas de referência das estações, e precisam ser consideradas quando significativas. Eventos como trocas de antenas e terremotos, são as principais fontes causadoras de descontinuidades nas séries temporais. Neste trabalho, avaliou-se a capacidade de algoritmos de *machine learning* na identificação automática de descontinuidades relacionadas a trocas de antenas. Dos cinco métodos avaliados, o *Random Forest* apresentou o melhor desempenho, com um F1-Score de 0,78 e uma taxa de acerto de 77,5% para descontinuidades iguais ou superiores a 1 cm. Os resultados demonstram o potencial dos métodos de *machine learning* na classificação de padrões em séries temporais de coordenadas da RBMC. Contudo, o desempenho depende de um tratamento adequado dos dados e da representatividade dos eventos modelados.

Palavras-chave: RBMC; GNSS; machine learning.

Abstract: The GNSS observations collected by RBMC allow estimating, with millimeter precision, the daily coordinates of its stations. As a result, it is possible to generate robust time series, capable of describing geodynamic displacements, such as the movements of lithospheric plates, as well as the effects of terrestrial, oceanic and atmospheric tides. However, discontinuities in these series indicate changes in the reference coordinates of the stations, and need to be considered when significant. Events such as antenna changes and earthquakes are the main sources of discontinuities in time series. In this work, we evaluated the capacity of machine learning algorithms in the automatic identification of discontinuities related to antenna changes. Of the five methods evaluated, Random Forest presented the best performance, with an F1-Score of 0.78 and an accuracy rate of 77.5% for discontinuities equal to or greater than 1 cm. The results demonstrate the potential of machine learning methods in classifying patterns in time series of RBMC coordinates. However, performance depends on adequate data processing and the representativeness of the modeled events.

Keywords: RBMC; GNSS; machine learning**.**

Recebido: 07/04/2025; Aceito: 13/08/2025; Publicado: 26/08/2025.

1. Introdução

Desde a implantação de sua primeira estação, em 1995, na cidade de Brasília, a Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS – RBMC vem disponibilizando dados de observações de pontos localizados na superfície terrestre (COSTA et al., 2012). Ao todo, são mais de 200 estações que servem de apoio para trabalhos que necessitam de uma estrutura primária de referência. A continuidade de sua operação e o processamento GNSS sistemático permitem determinar, com precisão milimétrica, coordenadas diárias para cada uma das estações, estabelecendo séries temporais capazes de monitorar o referencial geodésico no Brasil. Porém, fenômenos naturais, como terremotos, e manutenções de equipamentos, como a troca de antenas, podem provocar descontinuidades (saltos) nas séries (DAWIDOWICZ et al., 2023; LE et al., 2024; LAHTINEN et al., 2022).

A identificação de descontinuidades em séries temporais de coordenadas resulta na segmentação da série original em múltiplas soluções, conforme discutido por Altamimi et al. (2023). A estação BRAZ, localizada em Brasília, teve sua série segmentada em oito soluções, cada uma com coordenadas distintas no ITRF2020 (ALTAMIMI et al., 2022). Ferramentas como o FODITS, do programa Bernese (DACH et al., 2015), permitem a análise de descontinuidades. No entanto, a maioria dessas ferramentas requer informações prévias sobre os eventos, tornando o processo limitado e semiautomático.

O objetivo deste trabalho é avaliar a capacidade de identificar descontinuidades geradas por trocas de antenas por meio de técnicas de *machine learning*. Foram utilizadas séries temporais das coordenadas de 198 estações da RBMC, estimadas pelo *Nevada Geodetic Laboratory – NGL*, *da University of Nevada, Reno* (BLEWITT et al., 2018), além de soluções relativas diárias estimadas pelo IBGE no referencial IGS20. Dos algoritmos de *machine learning* já utilizados em GNSS (SIEMURI et al., 2022) cinco foram testados: *Random Forest*, *Linear Suporte Vector Machine*, *K-Nearest Neighbor*, *Desicion Trees* e *Naive Bayes*. Considerando que as séries apresentam desvios padrão que podem atingir até 0,7 cm, tomamos como significativas as descontinuidades iguais ou superiores a 1 cm, conforme adotado em Crocetti et al. (2021).

Um dos principais desafios na identificação de descontinuidades provocadas por trocas de antenas é a ausência de dados ou a presença de dados ruidosos. Além disso, os saltos não ocorrem de forma homogênea, pois dependem de fatores, como o modelo das antenas (DAWIDOWICZ et al., 2023), a localização da estação, a qualidade dos dados, o tipo de dispositivo de centragem e orientação da antena. A seção 2 deste trabalho apresenta algumas características das séries temporais da RBMC, as descontinuidades provocadas por terremotos e por manutenções realizadas, além dos critérios para a escolha das séries analisadas. A metodologia empregada para a formação das amostras, o pré-processamento dos dados e os critérios de avaliação da classificação são descritos na seção 3. Na seção 4, são apresentados os índices de qualidade da classificação obtidos para cada um dos algoritmos de *machine learning*, a capacidade de detecção de descontinuidades para diferentes amplitudes de salto e sua aplicação em séries temporais da RBMC produzidas pelo IBGE. As conclusões e recomendações finais são discutidas na seção 5.

2. Séries temporais das coordenadas da RBMC

A RBMC tem os dados de suas estações processados continuamente por diversos centros de processamento, como aqueles vinculados ao SIRGAS, ao IGS, pelo próprio IBGE e por institutos geodésicos, como o NGL. Este último disponibiliza coordenadas diárias determinadas com o programa GipsyX, além de um arquivo denominado *steps.txt* (http://geodesy.unr.edu/NGLStationPages/steps.txt) que contém informações sobre eventos considerados potenciais causadores de descontinuidades, como terremotos e trocas de antenas (BLEWITT et al., 2018). Informações sobre os equipamentos também podem ser encontradas nos *logfiles* das estações (IBGE, 2024).

Trocas de antenas nas estações, mesmo quando substituídas por modelos idênticos, podem provocar descontinuidades devido ao fato de que as correções dos desvios e variações do centro de fase são estabelecidas em laboratório, e não no local da estação, sendo assim influenciadas por fenômenos locais, como o multicaminhamento (DAWIDOWICZ et al., 2023). A orientação e a mudança na altura da antena, quando não estabelecidas corretamente, são outras causas relacionadas à manutenção que também podem gerar descontinuidades (HUANG et al., 2025).

A Figura 1 apresenta a série temporal das coordenadas diárias da estação POAL. As linhas pontilhadas azuis representam os terremotos registrados segundo BLEWITT et al. (2018), enquanto as vermelhas indicam as diferentes trocas de antenas.

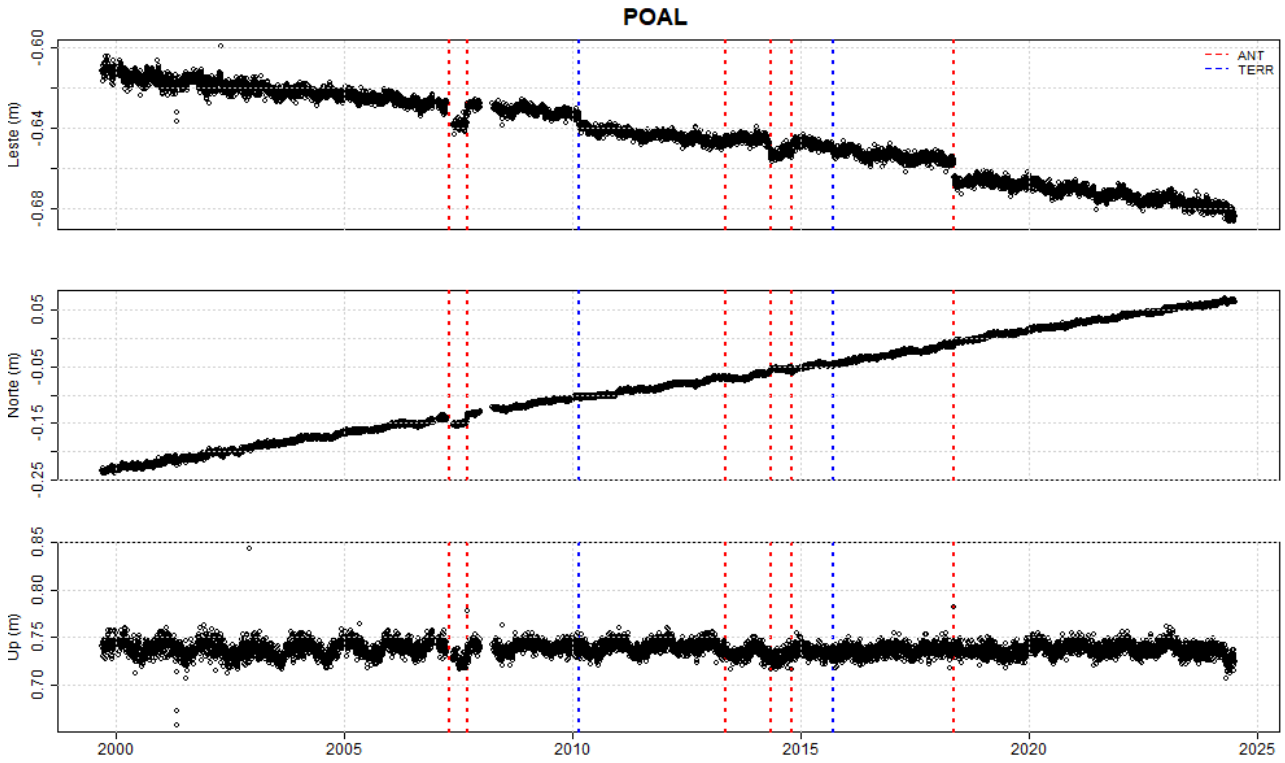


Figura 1 – Série temporal da estação POAL (Porto Alegre-RS) determinada a partir das soluções do NGL. Nota-se descontinuidades decorrentes de múltiplas trocas de antenas e do terremoto ocorrido em janeiro de 2010 em Concepción, Chile. Fonte: Os autores (2024).

3. Metodologia

Para o treinamento de um algoritmo de *machine learning*, é fundamental que os dados sejam representativos das classes desejadas e estejam disponíveis em quantidade suficiente para uma classificação eficaz (SIEMERS et al., 2022; RAJPUT et al., 2023). No caso das descontinuidades em séries temporais causadas por trocas de antenas, nem sempre as observações estão disponíveis, pois muitas dessas trocas ocorrem devido a falhas nos equipamentos. Em outros casos, os dados estão disponíveis, mas são ruidosos, ou seja, de baixa qualidade, justamente a razão pela qual o equipamento está sendo substituído (Figura 2). Em ambas as situações, não é recomendável utilizar esses períodos no treinamento do modelo, uma vez que a descontinuidade pode não estar presente na amostra temporal ou não apresentar um padrão representativo (GAO et al., 2024).

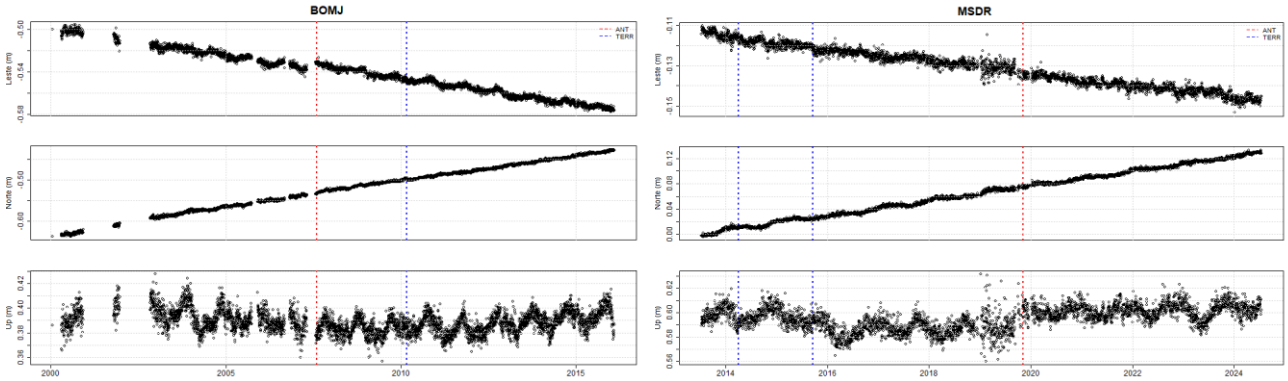


Figura 2 – Descontinuidades provenientes de trocas de antenas, porém de difícil modelagem uma vez que possuem falta de dados antes da troca (BOMJ - 01/08/2007) e com dados altamente ruidosos (MSDR - 04/11/2019). Fonte: Os autores (2024).

A amplitude mínima de um salto caracterizado como descontinuidade depende da variabilidade natural da série temporal, a qual resulta da precisão das coordenadas, dos efeitos de marés terrestres, oceânicas e atmosféricas, além de movimentos geodinâmicos locais (SAVCHUK et al., 2023). Nas séries das estações da RBMC, estimadas pelo NGL, o desvio padrão móvel médio das componentes Leste, Norte e Up, considerando uma janela temporal de três semanas, foi de 1 mm, 1 mm e 5 mm respectivamente, com valores máximos chegando a 7 mm na componente Up. Como a troca de antena pode gerar descontinuidade em todas as componentes, mas com maior probabilidade de ocorrência na componente altimétrica, valores inferiores ao desvio padrão médio de Up não devem ser considerados significativos para fins de classificação. Neste trabalho, adotou-se o valor de 1 cm como limite para descontinuidades significativas, o mesmo critério utilizado por Crocetti et al. (2021). As descontinuidades foram calculadas a partir da diferença entre a mediana dos últimos dez dias antes da troca e a mediana dos primeiros dez dias após a troca.

Das 233 trocas de antenas registradas nas 198 séries temporais das estações da RBMC analisadas, apenas 33 estações apresentaram descontinuidades superiores a 1 cm. Considerando-se que 80% das séries são utilizadas para o desenvolvimento do modelo, apenas 26 estações seriam selecionadas para o treinamento. No entanto, devido às particularidades de cada série temporal, o uso de apenas 13% dos dados das estações da RBMC revela-se insuficiente para o estabelecimento de padrões que permitam classificar, com precisão aceitável, as descontinuidades em toda a rede. Para contornar essa baixa ocorrência de eventos, adotaram-se neste trabalho alguns procedimentos com o objetivo de obter uma classificação mais representativa e eficiente.

3.1. Seleção dos dados para o treinamento

Considerando o baixo número de descontinuidades conhecidas, significativas e adequadas para o treinamento de um modelo de *machine learning*, e, ao mesmo tempo, a grande quantidade de dados disponíveis, selecionamos períodos das séries temporais que apresentassem continuidade nos dados, contemplassem as variações sazonais de cada estação e estivessem livres de vazios que pudessem induzir à identificação de falsas descontinuidades, bem como de trocas de antenas e terremotos potencialmente perturbadores. As descontinuidades foram, então, intencionalmente introduzidas, com o objetivo de avaliar a eficiência dos métodos de identificação e tratamento dessas interrupções (Seção 3.3). Optamos apenas por períodos com, no mínimo, 95% de dados disponíveis, no máximo uma semana consecutiva sem coordenadas, e duração mínima de quatro anos, o que assegura a captura das variações sazonais, além de permitir uma estimativa mais confiável das velocidades (TUNINI et al., 2024). Com base nesses critérios, das 198 soluções disponibilizadas, identificamos séries temporais de 108 estações da RBMC, das quais utilizamos 87 para o treinamento e 21 para os testes (Figura 3).

Figura 3 – Estações da RBMC escolhidas aleatoriamente para treinarem (80%) e testarem (20%) os modelos de machine learning. Fonte: Os autores (2024).

3.2. Remoção de *outliers* e áreas ruidosas

Os dados utilizados para o treinamento do algoritmo foram preprocessados com o objetivo de remover observações indesejadas, como aquelas provenientes de medições de baixa qualidade ou resultantes de um processamento inadequado. Inicialmente, determinou-se uma precisão de referência para cada série temporal das componentes, por meio do cálculo da mediana do desvio padrão móvel estimado em uma janela temporal de três semanas, conforme apresentado em Crocetti et al. (2021). Em seguida, a mediana móvel da série temporal, estimada para a mesma janela de tempo, foi subtraída dos valores da série. Observações com diferenças superiores a três vezes o desvio padrão de referência foram consideradas *outliers* e, portanto, removidas. Após a remoção dos *outliers*, procedeu-se à identificação de áreas ruidosas. Para isso, os valores de referência foram novamente estimados. As observações foram classificadas como ruidosas quando a mediana do desvio padrão móvel, estimada para uma janela de três semanas, foi superior a duas vezes a precisão de referência da respectiva componente (CROCETTI et al., 2021).

3.3. Inserindo descontinuidades nas séries de dados reais

Para permitir que o algoritmo de *machine learning* seja treinado com amostras contendo saltos bem definidos, optou-se pela simulação de descontinuidades em dados reais das séries temporais. Assim, as únicas descontinuidades presentes nas séries selecionadas foram aquelas inseridas propositalmente. Para isso, foi incluído um salto de 1 cm em cada série temporal, aplicando-se às componentes Leste, Norte e Up os valores 0, +1 ou -1, onde 0 indica ausência de salto, +1 indica um salto positivo e -1 um salto negativo. Gerou-se combinações entre essas três componentes aleatoriamente, com exceção

da combinação [0, 0, 0], que foi descartada por não representar nenhuma descontinuidade. Ressalta-se que saltos provocados por trocas de antenas podem afetar qualquer uma das componentes das coordenadas, embora ocorram com maior frequência e intensidade na componente altimétrica (WANNINGER, 2009). A definição da época para a inserção do salto foi realizada por meio de uma varredura iniciada a partir do ponto médio da série temporal, buscando-se um intervalo contínuo de dados que correspondesse ao dobro da janela amostral utilizada no treinamento do modelo, no caso deste trabalho, 21 dias. Realizamos a varredura em ambas as direções temporais, e o período escolhido foi aquele cuja distância temporal entre o seu ponto médio e o centro da série fosse a menor (Figura 4).

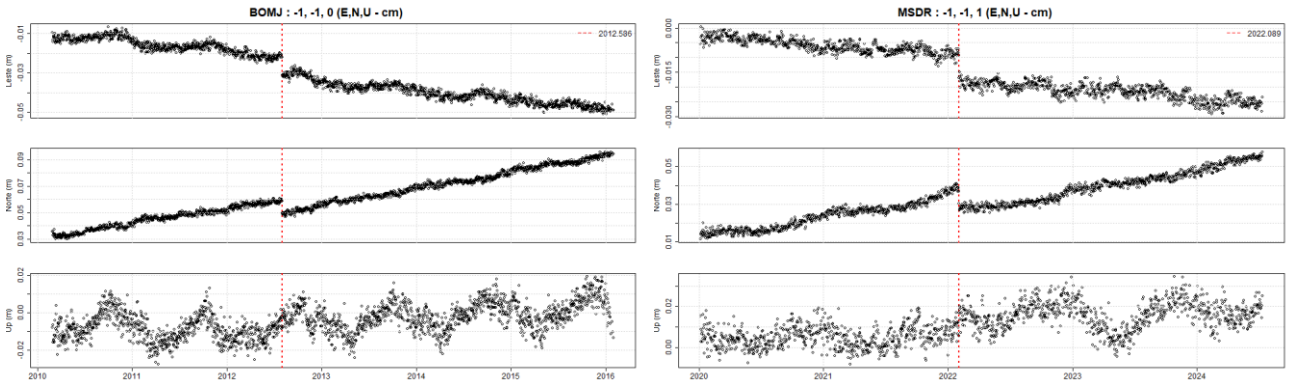


Figura 4 – Descontinuidades inseridas aleatoriamente nas séries temporais de BOMJ (Bom Jesus da Lapa-BA) em 01/08/2012 (-1 cm em Leste e -1 cm em Norte) e MSDR (Dracena-MS) em 30/03/2022 (-1 cm em Leste, -1 cm em Norte e 1 cm em Up). Fonte: Os autores (2024).

3.4. Montando as amostras para o treinamento

Essa etapa consistiu em transformar uma série temporal com *n* épocas em um conjunto de *n* - *m* amostras. Cada amostra contém *m* épocas, extraídas a partir de uma janela móvel que percorre toda a série. Cada uma dessas amostras representa um padrão que, associado à informação sobre a existência ou não de um salto e sua posição (quando houver), permite o treinamento do algoritmo. Como as coordenadas são compostas por três componentes (Leste, Norte e Up), e todas podem ser afetadas por descontinuidades, o algoritmo foi treinado utilizando o conjunto de amostras das três componentes simultaneamente. Isso resultou em uma matriz de estudo com dimensão (*n* - *m*) x 3*m*. De acordo com Crocetti et al. (2021), os tamanhos de amostra que apresentam melhor desempenho na detecção de descontinuidades em séries temporais de coordenadas são aqueles compostos por 21 ou 28 épocas. Neste trabalho, optou-se pela configuração com 21 épocas, o que, considerando as três componentes, gera para cada amostra 63 elementos, representando uma janela de 21 dias contendo dados das componentes Leste, Norte e Up. Essa configuração permite não apenas identificar a existência de um salto, mas também sua posição dentro da janela. O treinamento do algoritmo foi conduzido por meio de um vetor denominado *vetor alvo*, no qual os valores 0 indicam ausência de descontinuidade na amostra, e os valores de 1 a 20 indicam a ocorrência de uma descontinuidade e sua posição relativa dentro da janela amostral. Vale destacar que, se o salto ocorrer justamente na primeira época da amostra, o vetor alvo receberá o valor 0, uma vez que, nesse caso, a amostra conterá apenas dados posteriores ao salto (Figura 5).

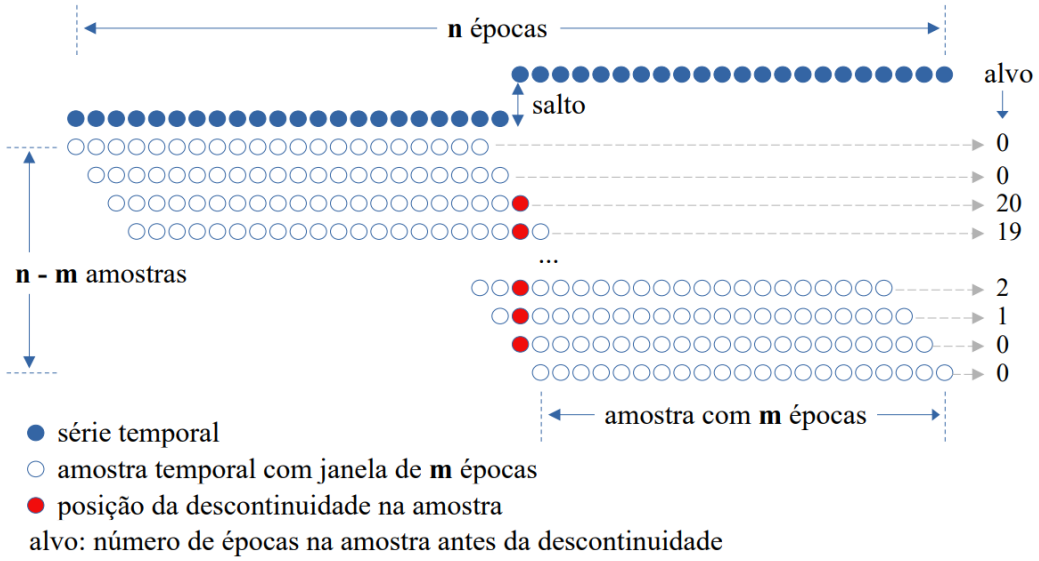


Figura 5 – Esquema com a montagem da matriz de estudo formada a partir de uma janela móvel e o vetor alvo correspondente. Fonte: Os autores (2024).

Além da construção da matriz de estudo, outra estratégia recomendada por Crocetti et al. (2021) e adotada neste trabalho foi a normalização individual dos valores de cada amostra. Essa normalização visa padronizar a escala dos dados, facilitando o aprendizado do algoritmo. Adicionalmente, foram inseridas três colunas extras na matriz de estudo, correspondentes ao intervalo (range) original de cada amostra para as componentes Leste, Norte e Up, respectivamente. Espera-se que, na presença de um salto, o respectivo range sofra uma alteração proporcional à magnitude do salto, quando comparado a amostras sem descontinuidades naquela mesma componente.

3.5. Avaliação de desempenho da classificação

Com o modelo treinado, torna-se necessário avaliar seu desempenho em relação aos dados reservados para teste. Para isso, realiza-se a classificação conforme a montagem das amostras e compara-se os resultados com aqueles previamente conhecidos, utilizando a matriz de confusão (ÖZBEY et al., 2024; HEYDARIAN et al., 2022). As amostras com descontinuidades corretamente classificadas são definidas como Verdadeiros Positivos – VP, enquanto as amostras sem descontinuidades corretamente identificadas são os Verdadeiros Negativos – VN. Já as amostras incorretamente classificadas como contendo descontinuidades, correspondem aos Falsos Positivos – FP, e aquelas que possuíam descontinuidades, mas foram classificadas como sem saltos, são os Falsos Negativos – FN. Adicionalmente, há casos em que as descontinuidades foram corretamente identificadas, mas com deslocamento temporal em relação à época real do salto. Nesses casos, a amostra também é classificada como Verdadeiro Positivo – VP*, porém não considerada nos indicadores de qualidade (CROCETTI, et al., 2021). A partir dos valores da matriz de confusão, são determinados os principais índices de desempenho da classificação: Precisão (Pr), indicando a proporção de amostras classificadas com descontinuidades que de fato ocorreram; Revocação (Re), representando a capacidade do modelo de identificar corretamente as descontinuidades presentes; e F1-score, que é a média harmônica entre precisão e revocação, sendo o indicador principal adotado neste trabalho para selecionar o algoritmo que melhor se ajusta às amostras analisadas. As equações 1, 2 e 3 apresentam as expressões matemáticas utilizadas para o cálculo desses indicadores (DE DIEGO et al., 2022). Um fluxograma resumindo a metodologia adotada é apresentado na Figura 6.

$$Pr = \frac{VP}{VP+FP} \quad (1)$$

$$Re = \frac{VP}{VP+FN} \quad (2)$$

$$F1score = \frac{2xPrxRe}{Pr+R} = \frac{2VP}{2VP+FP+FN} \quad (3)$$

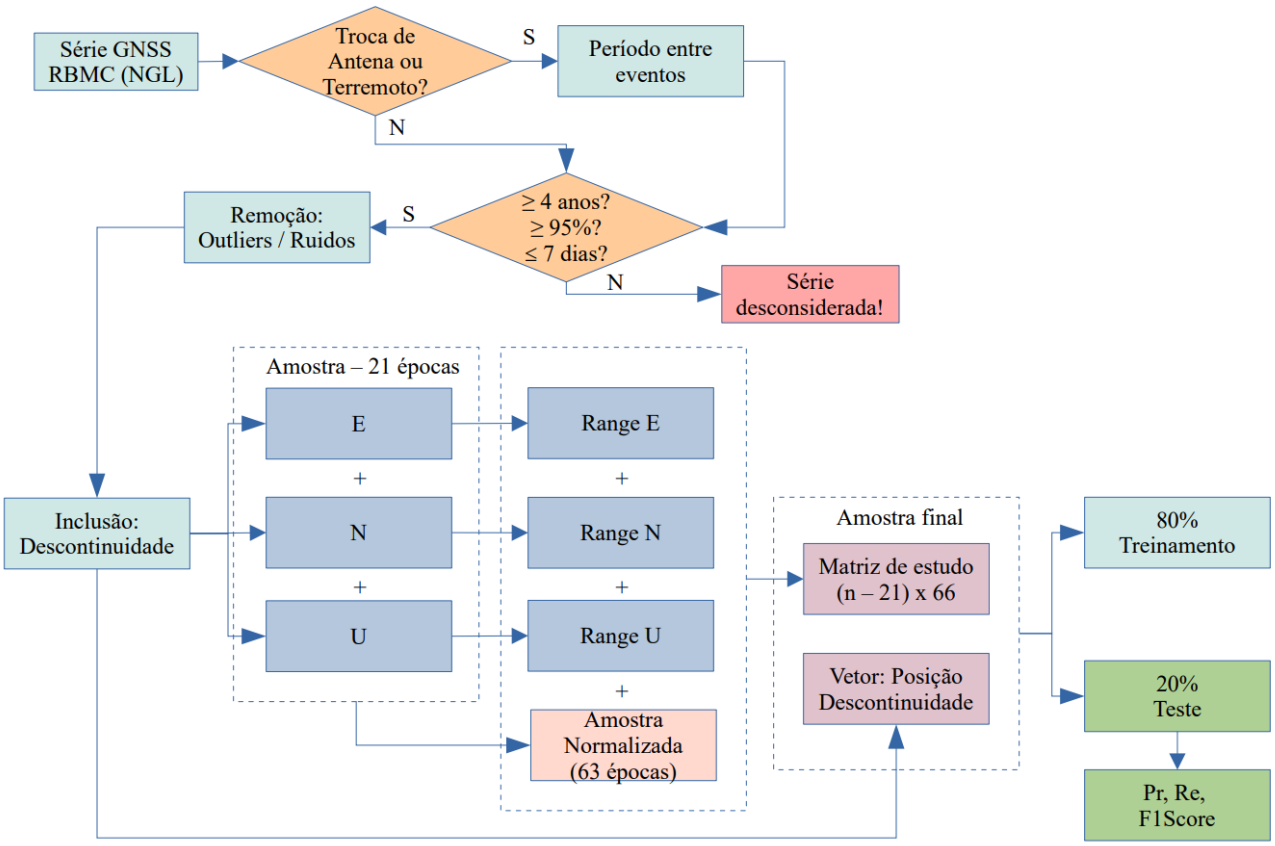


Figura 6 – Fluxograma ilustrando a metodologia aplicada na seleção dos dados, detecção de outliers e ruídos, montagem das amostras e análise da qualidade da classificação. Fonte: Os autores (2024).

4. Resultados

O desempenho de uma classificação por *machine learning* está associado a diversos fatores, como a lógica do algoritmo, a qualidade e o comportamento dos dados, a metodologia adotada na construção das amostras, a estratégia de filtragem no pré-processamento e a representatividade das classes dentro do conjunto de treinamento. No caso da detecção de descontinuidades em séries temporais de coordenadas oriundas de uma rede continental como a RBMC, somam-se ainda fatores adicionais, como o comportamento sazonal acentuado da componente altimétrica em determinadas estações, as diferentes precisões entre as componentes das coordenadas, a ampla distribuição geográfica das estações e a ausência de correlação espacial entre os eventos. Além disso, a falta de observações nos dias imediatamente anteriores aos saltos compromete a caracterização do padrão da descontinuidade, tornando a classificação ainda mais complexa.

4.1. Algoritmos de classificação

Analisamos os desempenhos de cinco algoritmos de *machine learning* aplicados à classificação multiclasse: *Random Forest* – RF (ALI et al., 2012), *Suporte Vector Machine* – SVM (LORENA e CARVALHO, 2007), *K-nearest neighbors* – KNN (SUN et al., 2018), Árvore de decisão – AD (ROKACH e MAIMON, 2005) e *Naive Bayes* – NB (WEBB, 2011). Conforme apresentado na Figura 7, e em concordância com Crocetti et al. (2021), o algoritmo *Random Forest* foi o que obteve o melhor desempenho, com um F1-Score de 0,78, sendo, portanto, o método adotado neste trabalho. Destaca-se que 77,5% das descontinuidades existentes foram corretamente identificadas (VP + VP*) pelo *Random Forest*, sendo que 39,6% foram classificadas exatamente na época de sua ocorrência (VP). Observa-se também que não houve ocorrência de falsos positivos (FP), ou seja, nenhuma descontinuidade foi classificada sem que houvesse registro correspondente, resultando em uma precisão (Pr) de 1 para o *Random Forest*. Por outro lado, como 22,5% das descontinuidades não foram identificadas (FN), a revocação (Re) foi de 0,64. É importante destacar que, para a identificação de descontinuidades, o valor de *Re* é mais relevante que o de *Pr*, uma vez que reflete a capacidade do algoritmo de detectar todos os eventos de interesse.

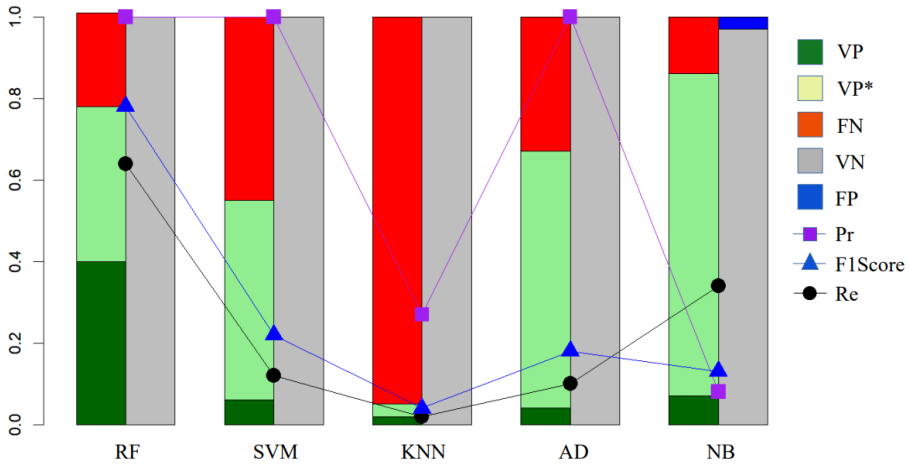


Figura 7 – Indicadores de qualidade dos algoritmos testados na classificação de descontinuidades nas séries temporais da RBMC. Fonte: Os autores (2024).

4.2. Capacidade de detecção quanto ao tamanho dos saltos

É natural que, quanto maior for o salto em uma série temporal, maior seja a eficácia de um algoritmo de classificação, já que a descontinuidade torna-se mais evidente em relação à variabilidade natural das amostras. De fato, ao aplicar o algoritmo *Random Forest* previamente treinado com descontinuidades de 1 cm, em amostras contendo saltos de 1 cm, 2 cm e 3 cm, observou-se uma melhora nos indicadores de desempenho à medida que se aumentou o tamanho do salto. O F1-Score passou de 0,78 para 0,86, com destaque para o aumento expressivo do número de Verdadeiros Positivos (VP), que subiu de 39,6% para 67,3%, conforme apresentado na Figura 8. É importante destacar que a quantidade de Falsos Negativos (FN) praticamente não se alterou, permanecendo em torno de 20%. Esse comportamento sugere que, mesmo com o aumento da magnitude das descontinuidades, o algoritmo treinado não foi capaz de definir os padrões associados aos saltos, o que indica uma possível limitação na representatividade dos eventos utilizados no treinamento.

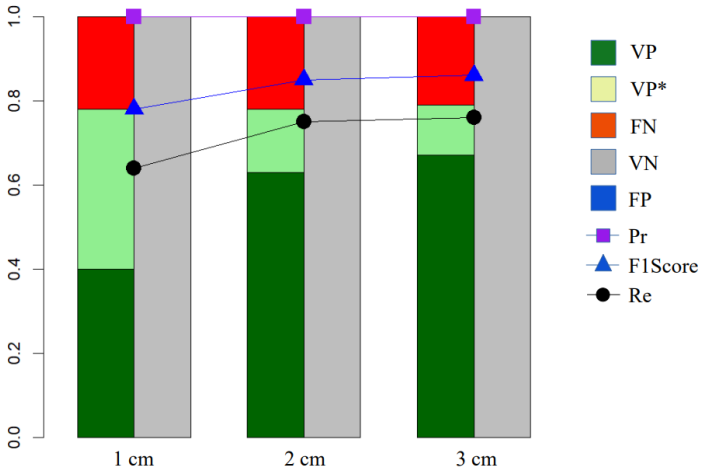
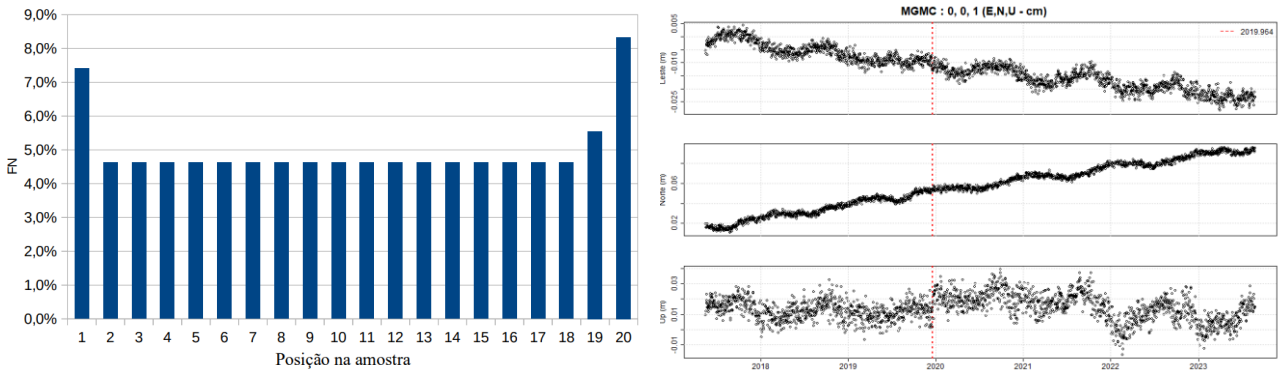


Figura 8 – Indicadores de qualidade do algoritmo Random Forest testados na classificação para diferentes tamanhos de descontinuidades. Fonte: Os autores (2024).

4.3. Descontinuidades não identificadas nas amostras

Com o objetivo de investigar as razões pelas quais aproximadamente 20% das descontinuidades presentes nas amostras não foram identificadas pelo algoritmo *Random Forest*, realizou-se uma análise focada na posição dos saltos dentro das janelas temporais. Verificou-se que descontinuidades localizadas nas extremidades das amostras apresentaram maior dificuldade de detecção, uma vez que há insuficiência de valores anteriores ou posteriores à ocorrência do salto, o que compromete a identificação do padrão treinado (Figura 9a). No entanto, também foram observados amostras com saltos posicionados na região central da janela que, apesar disso, não foram corretamente classificadas pelo algoritmo. Nessas situações, o número de amostras com salto permanece constante, sugerindo que as descontinuidades não estão suficientemente nítidas na série. Um exemplo ilustrativo pode ser observado na Figura 9b, para a estação MGMC (Montes Claros-MG) onde foi inserido um salto de 1 cm na componente altimétrica em 18/12/2019. Neste caso específico, nota-se uma maior dispersão dos dados na componente Up, o que dificulta a detecção da descontinuidade.



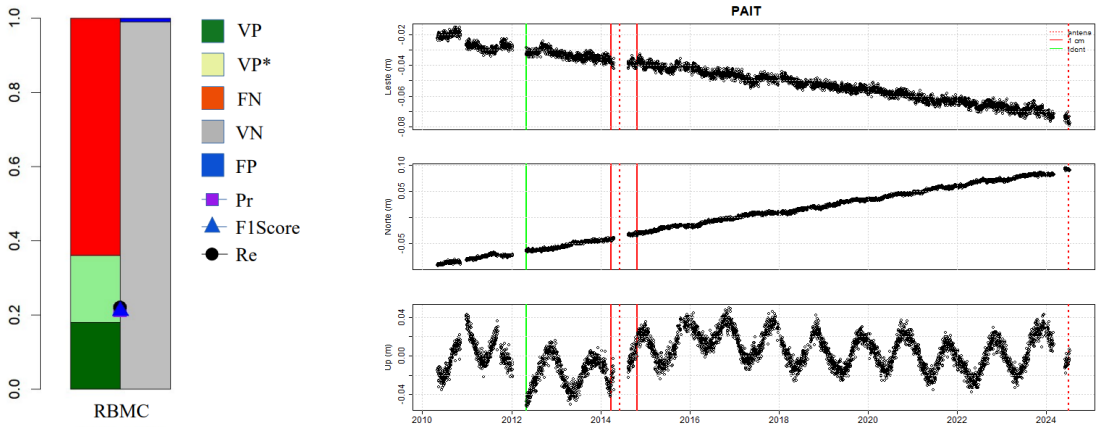
(a) (b)

Figura 9 – (a): Distribuição da posição dos FN dentro das amostras, indicando maior concentração nas extremidades. (b): descontinuidade de 1 cm na componente altimétrica da estação MGMC (Montes Claros-MG) não identificada pelo algoritmo de classificação, devido provavelmente à dispersão dos dados das amostras. Fonte: Os autores (2024).

4.4. Capacidade de classificação nas séries de coordenadas da RBMC com descontinuidades reais

Para avaliar a capacidade do modelo treinado em identificar descontinuidades reais nas séries temporais da RBMC, foram selecionadas 37 estações que apresentaram saltos superiores a 1 cm decorrentes de trocas de antenas, totalizando 47

eventos. Como esperado, os índices de qualidade dessa classificação foram inferiores aos apresentados na Seção 4.1, com F1-Score de 0,21 e aproximadamente 36% das descontinuidades corretamente identificadas nas amostras (VP + VP*). Por outro lado, observou-se a presença de descontinuidades detectadas pelo modelo, mas não registradas previamente como eventos superiores a 1 cm. Embora classificadas como Falsos Positivos (FP), elas não devem ser interpretadas automaticamente como falha do algoritmo, pois podem indicar eventos reais não registrados nas séries. Destaca-se ainda que quando presentes nos dados de treinamento, esses FP podem comprometer a eficácia do aprendizado do modelo. A Figura 10a apresenta os indicadores de desempenho da classificação frente às descontinuidades reais. Já a Figura 10b ilustra a série temporal da estação PAIT (Itaituba-PA) destacando descontinuidades conhecidas superiores a 1 cm (linhas vermelhas contínuas), inferiores a 1 cm (linha vermelha tracejada), e uma descontinuidade adicional identificada pela classificação, mas cuja origem é desconhecida (linha verde contínua).

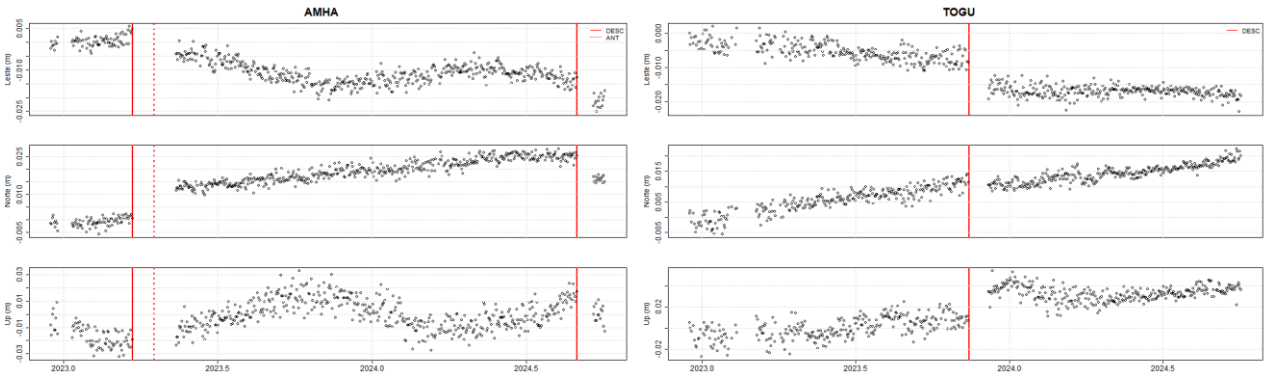


(a) (b)

Figura 10 – (a): Indicadores de qualidade da classificação por machine learning considerando as descontinuidades reais nas séries de dados da RBMC; (b): descontinuidade igual ou superior a 1 cm (linha vermelha contínua) identificada na estação PAIT (Itaituba-PA) entre 19/04/2012 e 01/05/2012, inferior a 1 cm (linha vermelha tracejada) e sem registro de troca de antena (linha verde contínua). Fonte: Os autores (2024).

4.5. Detecção de descontinuidades em séries temporais GNSS produzidas pelo IBGE

O algoritmo treinado pode ser aplicado a qualquer solução que envolva a estimativa sistemática de coordenadas, mesmo que estejam em referenciais distintos. O importante é que as séries sejam homogêneas e apresentem qualidades semelhantes. Neste trabalho, utilizamos o modelo *Random Forest* treinado com dados do NGL, para a detecção de descontinuidades em soluções diárias estimadas pelo IBGE. Essas soluções, referidas ao sistema IGS20, estão disponíveis desde 27/11/2022 e são determinadas por processamento relativo em rede, utilizando o software Bernese V5.4 (DACH et al., 2015), como parte da rotina interna da Coordenação de Geodésia para o controle de qualidade dos dados. Assim como observado nas séries do NGL, o algoritmo também foi capaz de identificar descontinuidades associadas a trocas de antenas nas soluções do IBGE, como nos casos das estações AMHA (Humaitá-AM) e TOGU (Gurupi-TO), conforme ilustrado na Figura 11. No caso específico da estação AMHA, a primeira descontinuidade identificada não coincide exatamente com a data da troca da antena. Essa divergência deve-se à ausência de dados nos dias imediatamente anteriores ao evento, o que impediu o algoritmo de reconhecer o salto no momento exato da ocorrência.



(a) (b)

Figura 11 – Descontinuidades identificadas em séries estimadas pelo IBGE (a): duas descontinuidades por trocas de antenas em AMHA (Humaitá-AM) sendo que a primeira foi identificada com data não coincidente ao evento; (b): descontinuidade em TOGU (Gurupi-TO) identificada na mesma época da troca da antena. Fonte: Os autores (2024).

5. Conclusão e recomendação

As observações da RBMC permitem estimar, com precisão milimétrica, as coordenadas diárias de suas estações. Esse nível de detalhamento possibilita a construção de séries temporais altamente ajustadas, fornecendo informações valiosas, como a velocidade e a direção do deslocamento das placas litosféricas. Para garantir a consistência temporal dessas coordenadas, e, quando necessário, definir múltiplas soluções, é fundamental considerar eventos perturbadores que possam afetar a posição das estações, como trocas de antenas e terremotos. Neste trabalho, avaliamos a capacidade de algoritmos de *machine learning* em identificar descontinuidades em séries temporais da RBMC causadas por trocas de antenas. Para isso, utilizamos séries produzidas pelo *Nevada Geodetic Laboratory – NGL da University of Nevada, Reno*, e informações sobre as épocas de ocorrência de terremotos e de trocas de antenas em cada estação.

O uso das descontinuidades reais geradas pelas trocas de antenas no treinamento do algoritmo não foi possível devido à ausência de dados imediatamente anteriores às épocas dos eventos. Para contornar esse problema, estabeleceram-se critérios para a seleção de períodos que permitisse considerar o maior número possível de estações da RBMC, assegurando, ao mesmo tempo, a representatividade dos dados e a presença dos efeitos sazonais observados nas séries. Ao todo, selecionamos 108 estações, que receberam, aleatoriamente, ao menos uma descontinuidade de 1 cm. Esses dados foram organizados em amostras temporais, geradas a partir de uma janela móvel de 21 dias que percorreu toda a série para cada uma das componentes das coordenadas, formando a matriz de estudo e, simultaneamente, estabelecendo um vetor alvo, indicando a existência (ou não) de descontinuidade e sua posição dentro da amostra.

Dentre os algoritmos de *machine learning* avaliados [*Random Forest* (RF), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT) e *Naive Bayes* (NB)] o *Random Forest* apresentou o melhor desempenho, com F1-Score de 0,78. O modelo foi capaz de identificar corretamente 77,5% das descontinuidades nas amostras (VP + VP*), sendo que 39,6% foram detectadas exatamente no instante de sua ocorrência (VP). Quando o valor dos saltos foi aumentado de 1 cm para 3 cm, a porcentagem de acerto dos VP subiu para 67,3% e o F1-Score para 0,86. Entretanto, cerca de 22% das descontinuidades não foram classificadas (FN), evidenciando a necessidade de um conjunto de dados mais representativo para o treinamento do modelo. O algoritmo treinado também foi aplicado às séries da RBMC contendo descontinuidades reais, apresentando um F1-Score de 0,21 e identificando aproximadamente 36% dos saltos (VP + VP*). Embora a eficácia tenha sido inferior à obtida com descontinuidades simuladas, o resultado era esperado, dado que parte dos dados imediatamente anteriores às trocas de antenas está ausente ou apresenta baixa qualidade. A mesma avaliação foi realizada nas séries produzidas pela Coordenação de Geodésia do IBGE no âmbito do controle de qualidade da RBMC, confirmando o grande potencial do modelo para aplicações práticas, uma vez que também foi capaz de identificar descontinuidades nessas soluções.

Este estudo demonstrou a capacidade de um modelo de *machine learning* em identificar, nas séries temporais, descontinuidades associadas a trocas de antenas em estações da RBMC. No entanto, considerando que os algoritmos aprendem com os dados a reconhecer padrões e que as descontinuidades representam apenas uma pequena fração do conjunto total de observações, recomenda-se a continuidade das pesquisas com o treinamento do modelo em um conjunto

mais amplo e diversificado de estações. Para isso, sugere-se a incorporação de dados de outras redes GNSS, como SIRGAS-CON, IGS, NOAA CORS, EUREF, entre outras.

Agradecimentos

Os autores agradecem ao NGL e ao IBGE pela disponibilização das séries temporais das coordenadas das estações da RBMC.

Referências

ALI, Jehad; KHAN, Rehanullah; AHMAD, Nasir; MAQSOOD, Imran. Random forests and decision trees. *International Journal of Computer Science Issues*, v. 9, n. 5, p. 272–278, set. 2012. Disponível em: http://www.ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf.

ALTAMIMI, Z.; REBISCHUNG, P.; COLLILIEUX, X.; MÉTIVIER, L.; CHANARD, K. *ITRF2020* [conjunto de dados]. IERS ITRS Center hospedado por IGN e IPGP, 2022. Disponível em: https://doi.org/10.18715/IPGP.2023.LDVIOBNL.

ALTAMIMI, Z.; REBISCHUNG, P.; COLLILIEUX, X. et al. ITRF2020: an augmented reference frame refining the modeling of nonlinear station motions. *Journal of Geodesy*, v. 97, p. 47, 2023. Disponível em: https://doi.org/10.1007/s00190-023-01738-w.

BLEWITT, G.; HAMMOND, W. C.; KREEMER, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos*, v. 99, 2018. Disponível em: https://doi.org/10.1029/2018EO104623.

COSTA, S.; LIMA, M.; MOURA JR, N.; DE ABREU, M.; SILVA, A.; FORTES, L.; RAMOS, A. RBMC in real time via NTRIP and its benefits in RTK and DGPS surveys. 2012. Disponível em: https://doi.org/10.1007/978-3-642-20338-1_115.

CROCETTI, L.; SCHARTNER, M.; SOJA, B. Discontinuity detection in GNSS station coordinate time series using machine learning. *Remote Sensing*, v. 13, p. 3906, 2021. Disponível em: https://doi.org/10.3390/rs13193906.

DACH, R.; LUTZ, S.; WALSER, P.; FRIDEZ, P. *Bernese GNSS Software Version 5.2: user manual*. Bern: Bern Open Publishing, 2015. ISBN 978-3-906813-05-9. DOI: https://doi.org/10.7892/boris.72297.

DAWIDOWICZ, K.; KRZAN, G.; WIELGOSZ, P. Offsets in the EPN station position time series resulting from antenna/radome changes: PCC type-dependent model analyses. *GPS Solutions*, v. 27, p. 9, 2023. Disponível em: https://doi.org/10.1007/s10291-022-01339-8.

DE DIEGO, I. M.; REDONDO, A. R.; FERNÁNDEZ, R. R. et al. General performance score for classification problems. *Applied Intelligence*, v. 52, p. 12049–12063, 2022. Disponível em: https://doi.org/10.1007/s10489-021-03041-7.

GAO, W.; LI, Z.; CHEN, Q. et al. Modelling and prediction of GNSS time series using GBDT, LSTM and SVM machine learning approaches. *Journal of Geodesy*, v. 96, art. 71, 2022. Disponível em: https://doi.org/10.1007/s00190-022-01662-5.

GAO, W.; WANG, C.; FENG, Y. A machine-learning-based missing data interpolation method for GNSS time series. In: YANG, C.; XIE, J. (org.). *China Satellite Navigation Conference (CSNC 2024) Proceedings*. Singapore: Springer, 2024. (Lecture Notes in Electrical Engineering, v. 1092). Disponível em: https://doi.org/10.1007/978-981-99-6928-9_20.

HEYDARIAN, M.; DOYLE, T. E.; SAMAVI, R. MLCM: multi-label confusion matrix. *IEEE Access*, v. 10, p. 19083–19095, 2022. Disponível em: https://doi.org/10.1109/ACCESS.2022.3151048.

HUANG, J.; HE, X.; HU, S.; MING, F. Impact of offsets on GNSS time series stochastic noise properties and velocity estimation. *Advances in Space Research*, v. 75, n. 4, p. 3397–3413, 2025. Disponível em: https://doi.org/10.1016/j.asr.2024.12.016.

IBGE. Instituto Brasileiro de Geografia e Estatística, 2024. Disponível em: https://geoftp.ibge.gov.br/informacoes_sobre_posicionamento_geodesico/rbmc/relatorio/log_sirgas/. Acesso em: 24 out. 2024.

JOHNSTON, G.; RIDDELL, A.; HAUSLER, G. The International GNSS Service. In: TEUNISSEN, P. J.; MONTENBRUCK, O. (eds.). *Springer Handbook of Global Navigation Satellite Systems*. Cham: Springer, 2017. p. 739–776. Disponível em: https://doi.org/10.1007/978-3-319-42928-1_33.

LAHTINEN, S.; JIVALL, L.; HÄKLI, P. et al. Updated GNSS velocity solution in the Nordic and Baltic countries with a semi-automatic offset detection method. *GPS Solutions*, v. 26, p. 9, 2022. Disponível em: https://doi.org/10.1007/s10291-021-01194-z.

LE, N.; MÄNNEL, B.; BUI, L. K. et al. Identifying neotectonic motions in Germany using discontinuity-corrected GNSS data. *Pure and Applied Geophysics*, v. 181, p. 87–108, 2024. Disponível em: https://doi.org/10.1007/s00024-023-03390-z.

LORENA, A. C.; DE CARVALHO, A. C. P. L. F. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Disponível em: https://doi.org/10.22456/2175-2745.5690.

ÖZBEY, V.; ERGINTAV, S.; TARI, E. GNSS time series analysis with machine learning algorithms: a case study for Anatolia. *Remote Sensing*, v. 16, p. 3309, 2024. Disponível em: https://doi.org/10.3390/rs16173309.

RAJPUT, D.; WANG, W. J.; CHEN, C. C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, v. 24, p. 48, 2023. Disponível em: https://doi.org/10.1186/s12859-023-05156-9.

ROKACH, L.; MAIMON, O. Decision Trees. In: MAIMON, O.; ROKACH, L. (eds.) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer, 2005. Disponível em: https://doi.org/10.1007/0-387-25465-X_9.

SAVCHUK, S.; DOSKICH, S.; GOŁDA, P.; RURAK, A. The Seasonal Variations Analysis of Permanent GNSS Station Time Series in the Central-East of Europe. *Remote Sensing*, v. 15, p. 3858, 2023. Disponível em: https://doi.org/10.3390/rs15153858.

SIEMERS, F.; FELDMANN, C.; BAJORATH, J. Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity. *Cell Reports Physical Science*, 2022. Disponível em: https://doi.org/10.1016/j.xcrp.2022.101113.

SIEMURI, A.; SELVAN, K.; KUUSNIEMI, H.; VÄLISUO, P.; ELMUSRATI, M. A systematic review of machine learning techniques for GNSS use cases. *IEEE Transactions on Aerospace and Electronic Systems*, p. 1–42, 2022. Disponível em: https://doi.org/10.1109/TAES.2022.3219366.

SUN, J. W.; DU, W. X.; SHI, N. C. A Survey of kNN Algorithm. *Information Engineering and Applied Computing*, v. 1, Article ID: 770, 2018. Disponível em: https://doi.org/10.18063/ieac.v1i1.770.

TUNINI, L.; MAGRIN, A.; ROSSI, G.; ZULIANI, D. Global Navigation Satellite System (GNSS) time series and velocities about a slowly convergent margin processed on high-performance computing (HPC) clusters: products and robustness evaluation. *Earth System Science Data*, v. 16, p. 1083-1106, 2024. Disponível em: https://doi.org/10.5194/essd-16-1083-2024.

WANNINGER, L. Correction of apparent position shifts caused by GNSS antenna changes. *GPS Solutions*, v. 13, p. 133-139, 2009. Disponível em: https://doi.org/10.1007/s10291-008-0106-z.

WEBB, G. I. Naïve Bayes. In: SAMMUT, C.; WEBB, G. I. (eds.) *Encyclopedia of Machine Learning*. Boston, MA: Springer, 2011. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_576.