

## SELEÇÃO DE ESTAÇÕES METEOROLÓGICAS AUTOMÁTICAS BASEANDO-SE NA QUALIDADE DO BANCO DE DADOS

SANTOS, A. A.<sup>1</sup>; SOUZA, J. L. M.<sup>1</sup>; ROSA, S. L. K.<sup>1</sup>

<sup>1</sup>Universidade Federal do Paraná, Programa de Pós-graduação em Ciência do Solo do Departamento de Solos e Engenharia Agrícola - Rua dos Funcionários, 1540 – CEP 80035 050 Curitiba PR.

### RESUMO

Os dados meteorológicos são importantes para o planejamento e uso dos recursos hídricos e do solo. Bem como, quando se trabalha com séries longas de dados meteorológicos há grande dificuldade de detectar valores atípicos, que não correspondem com a realidade do clima estudado. Teve-se por objetivo no presente trabalho aplicar e testar critérios para seleção e identificação de estações meteorológicas, baseando-se na existência ou não de dados de boa qualidade e representatividade de seus respectivos tipos climáticos brasileiros (*Af, Am, As, Aw, BSh, Cfa, Cfb, Cwa* e *Cwb*), empregando programas de fácil acesso, utilização e replicabilidade. Foram analisadas 168 estações meteorológicas automáticas do Instituto Nacional de Meteorologia (INMET) e seus respectivos dados meteorológicos, como: Temperatura máxima e mínima do ar ( $T_{máx}$ ,  $T_{mín}$  e  $T_{méd}$ ; °C); Umidade Relativa máxima e mínima do ar ( $UR_{máx}$ ,  $UR_{mín}$  e  $T_{méd}$ ; %), velocidade do vento à dez metros de altura ( $u_{10}$ ;  $m\ s^{-1}$ ); e, Radiação solar incidente ( $R_s$ ;  $MJ\ m^{-2}\ dia^{-1}$ ). As séries de dados foram obtidas na periodicidade horária, entre 01/01/2017 a 31/12/2020. Os critérios para seleção das estações e dados climáticos foram: I) “Falhas inferiores a 25% do total de dados horários da série, para cada variável climática”; II) “Não apresentar falha integral de todos os dados medidos em uma das estações do ano; e, III) “Não apresentar *outlier*”. Os critérios permitiram selecionar consistentemente as estações meteorológicas com melhor qualidade de dados. A região Norte apresentou o maior número de falhas e teve maior porcentagem de exclusão de dados. Verificou-se baixa cobertura espacial de estações na região Centro-Oeste do país. Os Critérios I, II e III analisados permitiram selecionar consistentemente as estações com melhor qualidade de dados. Das 168 estações meteorológicas analisadas, 95 contemplaram os critérios adotados.

**Palavras-chave:** Temperatura do ar; Umidade Relativa; Velocidade do vento; Radiação solar incidente.

### INTRODUÇÃO

A seleção de estações meteorológicas contendo dados com qualidade é importante para a área de engenharia de água e solo. As principais variáveis meteorológicas utilizadas em modelos para finalidades agrícolas são a umidade relativa do ar (UR), temperatura do ar (T), velocidade do vento a dois metros de altura ( $u_2$ ) e radiação solar insidente ( $R_s$ ) (Santos et al., 2021; Rosa, 2023).

O Instituto Nacional de Meteorologia (INMET) disponibiliza dados medidos em estações meteorológicas. As estações encontram-se distribuídas por todo o território nacional, representando os principais tipos climáticos brasileiros. No entanto, para compor um banco de

dados com qualidade, critérios devem ser adotados para escolha das estações e dados meteorológicos a serem utilizadas em análises. Santos (2021) considera que a existência de falhas de dados (horário, diário, mensal ou anual) prejudicam a utilização e o desempenho de modelos matemáticos, principalmente nos procedimentos para a sua calibração.

Na utilização de longa série de dados, ressalta-se a dificuldade para identificar os dados atípicos, que fogem da representatividade de um clima. Valadares Neto et al. (2017) recomendam a utilização do gráfico *boxplot*, que permite realizar a análise exploratória dos dados e detectar valores atípicos (*outliers*). O *boxplot* foi utilizado pela primeira vez pelo matemático Tukey (1970), tornando-se um método amplamente divulgado após a sua publicação em 1977 (Tukey, 1977). O autor desenvolveu o *boxplot* para simplificar e resumir os resultados da análise exploratória de dados. O *boxplot* é um recurso visual que resume os dados de uma amostra ou série, sendo apresentados a mediana, quartis e valores máximos e mínimos. Os gráficos *boxplot* são regularmente utilizados na divulgação de pesquisas científicas, sendo possível incorporar modificações como a forma que as hastes são desenhadas, ou a inserção de parâmetros como média (Valladares Neto et al., 2017).

O *boxplot* permite a disposição visual dos dados conforme três quartis (Q1, Q2 e Q3; Figura 1). As extremidades são representadas por uma linha (bigodes), limites superior e inferior. Na Figura 1 tem-se a representação das faixas de valores correspondentes a 99% de probabilidade de ocorrência. A leitura dos quartis é realizada considerando que 25% dos valores do conjunto de dados encontram-se no intervalo Q1, 50% em Q2 e 75% em Q3. Os pontos plotados fora dos limites são considerados *outliers*, sendo valores estatisticamente atípicos ao conjunto, amostra ou série de dados.

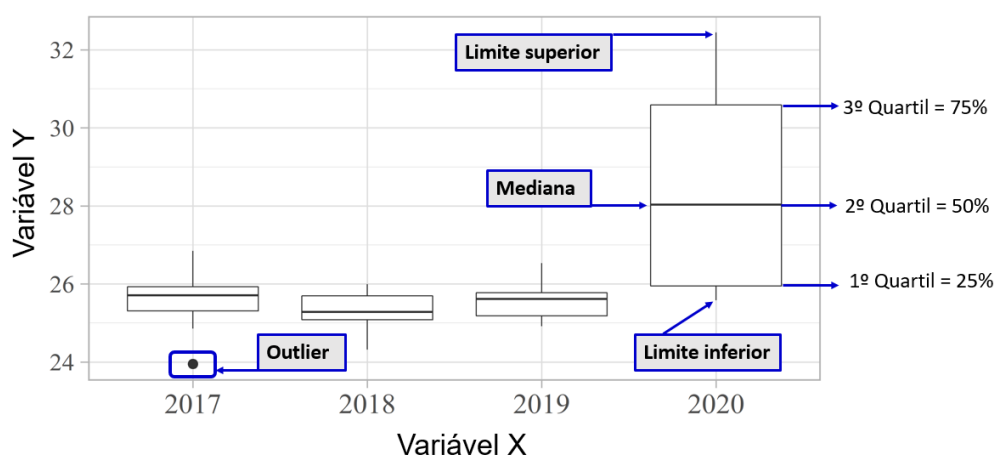


Figura 1 – Exemplificação da forma do gráfico *boxplot*, apresentando três quartis (com suas respectivas porcentagens), limites superior e inferior, mediana e outlier.

Diante do contexto apresentado, teve-se por objetivo no presente trabalho aplicar e testar critérios para seleção e identificação de estações meteorológicas, baseando-se na existência ou não de dados de boa qualidade e representatividade de seus respectivos tipos climáticos brasileiros (*Af, Am, As, Aw, BSh, Cfa, Cfb, Cwa e Cwb*), empregando programas de fácil acesso, utilização e replicabilidade.

## MATERIAL E MÉTODOS

Foram analisados dados de 168 estações meteorológicas automáticas do Instituto Nacional de Meteorologia (INMET), localizadas nas regiões e climas brasileiros (Figura 2). As variáveis analisadas foram: Temperaturas máxima, mínima e média do ar ( $T_{m\acute{a}x}$ ,  $T_{m\acute{i}n}$  e  $T_{m\acute{e}d}$ ; °C) Umidade Relativa máxima, mínima e média do ar ( $UR_{m\acute{a}x}$ ,  $UR_{m\acute{i}n}$ ,  $UR_{m\acute{e}d}$ ; %) Velocidade do vento à dez metros de altura ( $u_{10}$ ;  $m\ s^{-1}$ ) e Radiação solar incidente ( $R_s$ ;  $MJ\ m^{-2}$ ). Os dados foram obtidos na periodicidade horária, compreendendo o período entre 01/01/2017 e 31/12/2020, totalizando 47.093.760 dados meteorológicos.

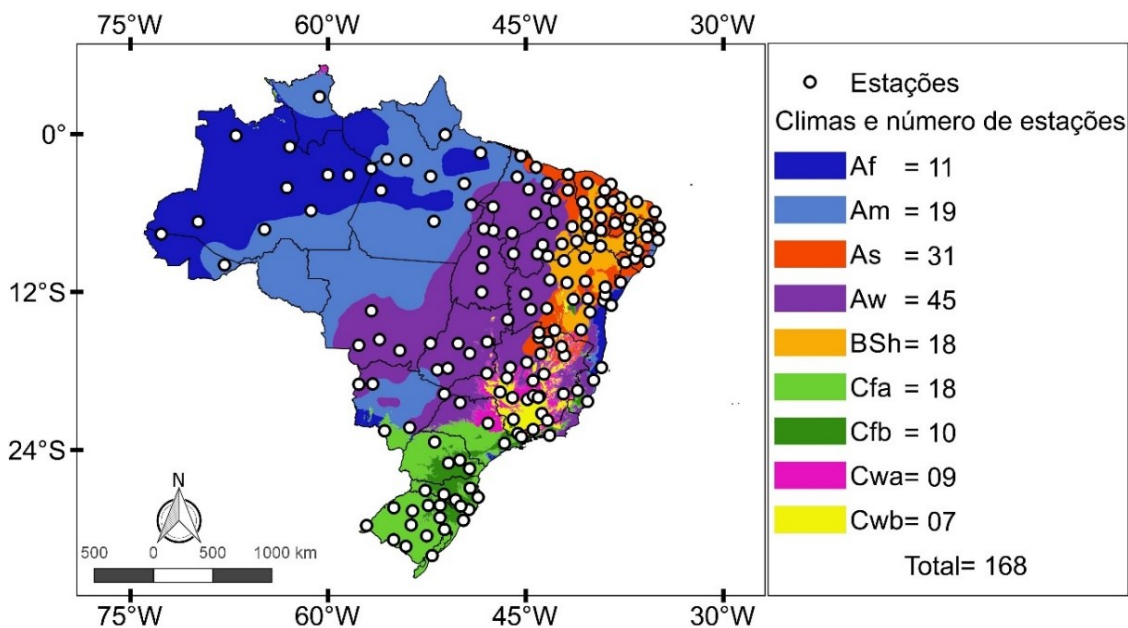


Figura 2 – Distribuição espacial das 168 estações meteorológicas automáticas do INMET analisadas, e quantificação das estações por tipo climático brasileiro.

Os Critérios para seleção das estações foram:

- I) “Falhas inferiores a 25% do total de dados horários da série, para cada variável climática”;
- II) “Não apresentar falha integral de todos os dados medidos em uma das estações do ano; e,
- III) “Não apresentar *outlier*”.

Os critérios foram estabelecidos baseando-se em: Santos et al. (2021), que considerou vários aspectos sobre a importância de critérios de escolha de estações meteorológicas, para que o banco de dados utilizado seja representativo do ambiente, para o estudo da evapotranspiração horária; e Valladares Neto (2017), que ressalta a qualidade da análise de dados em *boxplot* para detecção de *outliers*.

Os Critérios I e II foram aplicados aos dados por meio de funções e análise gráfica, realizada com ferramentas do Excel 2010. O Critério III foi aplicado por meio do software R, em interface RStudio (RStudio Team, 2021; versão 4.1.2) e pacote “ggplot2” (Wickham et al., 2022), com a função “boxplot”.

## RESULTADOS E DISCUSSÃO

Como cenário ilustrativo, nas Figuras 3 e 4 encontram-se os resultados das análises realizadas para o clima *Cwa*, quanto presença de *outliers* e falha de dados nas estações do ano. Constatou-se nas análises no Excel que não houve falhas superiores a 25%, e nem a presença de *outliers* (Figura 3). Contudo, na estação meteorológica de Caratinga (Figura 4) não houve dados para o período do outono. A estação foi excluída e o clima *Cwa* ficou com apenas oito estações representativas para serem analisadas (Tabela 1), das nove disponíveis. Desta forma, o mesmo procedimento foi aplicado para as 168 estações meteorológicas, e apenas 95 foram selecionadas (Tabela 1; Figura 5).

Na região Norte do Brasil, em que predominam os tipos climáticos *Am* e *Af* (Figuras 2 e 3), verificou-se as maiores falhas no banco de dados das estações meteorológicas. Houve 82% e 58% de exclusão das estações, respectivamente (Tabela 1), sendo a região com menor cobertura de estações meteorológicas, seguida da região Centro-Oeste (Figura 5b).

Santos et al. (2020), analisando bancos de dados meteorológicos brasileiros de diversas fontes, verificaram que o INMET apresenta boa cobertura de estações meteorológicas na região Norte. No entanto, muitas delas se encontram em local remoto, sendo comum ocorrência de falhas e perda de dados. Apesar da boa cobertura espacial das estações analisadas, verificou-se no presente estudo grandes áreas sem estações próximas (Figura 5a e 5b), dificultando a viabilidade na extrapolação de valores. Logo, a análise criteriosa para exclusão de estações é importante, pois as estações meteorológicas utilizadas devem apresentar dados robustos e poucas falhas.

Verificou-se também baixa cobertura espacial da região Centro-Oeste, em que predominam os tipos climáticos *Am* e *Aw* (Figuras 3 e 4). A região Norte, que apresenta a maior área territorial, foi a mais prejudicada quanto cobertura espacial. Apesar disso, a seleção com os critérios de escolha permitiu boa representatividade climática com as estações que foram selecionadas, devido a qualidade dos dados. Nos processos de calibração de um modelo, os dados utilizados devem apresentar qualidade, para não interferir no resultado e levar o pesquisador a interpretações equivocadas.

A necessidade de avaliar grandes séries de dados torna difícil a verificação de falhas ou a ocorrência de *outliers*. As ferramentas utilizadas no presente trabalho (Planilha eletrônica Excel e software R para *boxplot*), foram eficientes e simples para realizar a seleção do banco de dados.

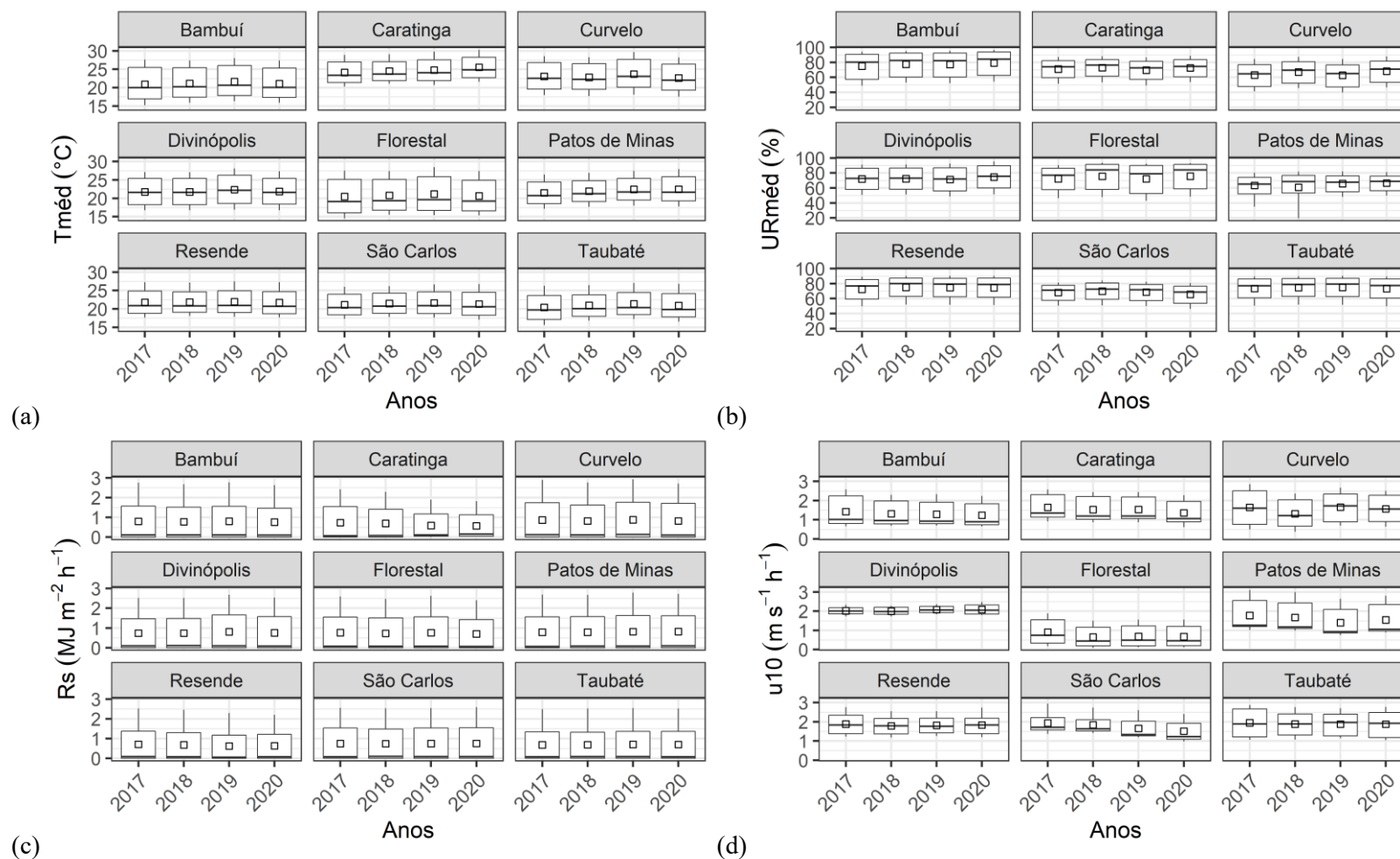


Figura 3 – *Outliers* ocorridos na periodicidade horária, nas estações meteorológicas com clima *Cwa*, para: a) temperatura média do ar ( $T_{méd}$ ); b) umidade relativa média do ar ( $UR_{méd}$ ); c) radiação solar incidente ( $R_s$ ); e, d) velocidade do vento a 10 m de altura ( $u_{10}$ ).

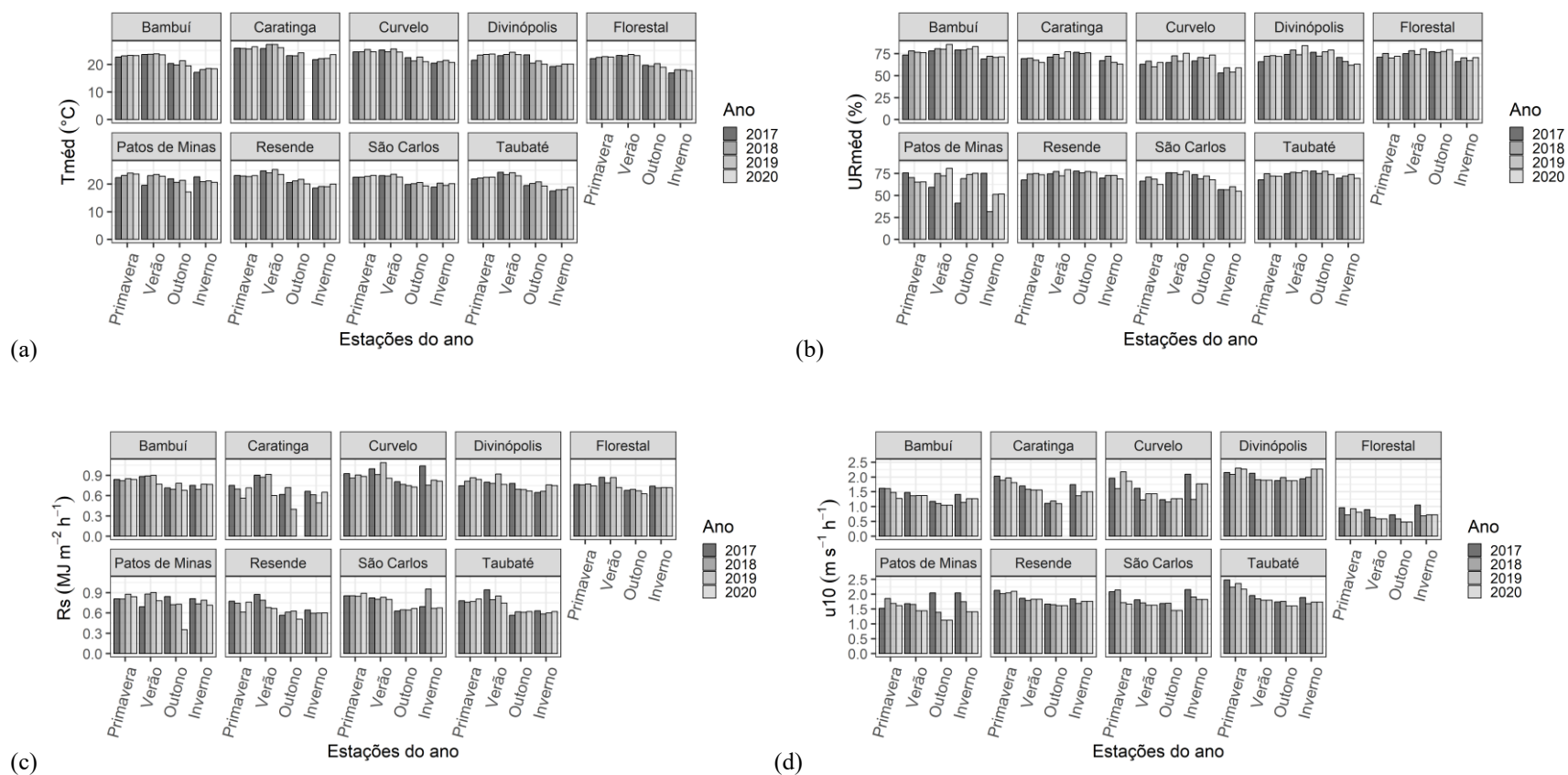


Figura 4 – Falhas dos dados verificadas nas estações do ano (falhas totais de dados na periodicidade horária, que impediram a realização de médias), entre 2017 e 2020, para as variáveis das estações meteorológicas em clima *Cwa*: a) temperatura média do ar ( $T_{méd}$ ); b) umidade relativa média do ar ( $UR_{méd}$ ); c) radiação solar incidente ( $R_s$ ); e d) velocidade do vento a 10 m de altura ( $u_{10}$ ).

Tabela 1 – Estações meteorológicas do instituto nacional de meteorologia (INMET) disponíveis, conforme o tipo climático, e percentual de estações excluídas após a aplicação dos critérios de seleção dos dados.

Climas	Estações disponíveis	Estações selecionadas	Excluídas
<i>Af</i>	12	5	58%
<i>Am</i>	17	3	82%
<i>As</i>	31	21	32%
<i>Aw</i>	46	29	37%
<i>BSh</i>	18	9	50%
<i>Cfa</i>	18	9	50%
<i>Cfb</i>	10	7	30%
<i>Cwa</i>	9	8	11%
<i>Cwb</i>	7	4	43%
Total	168	95	43%

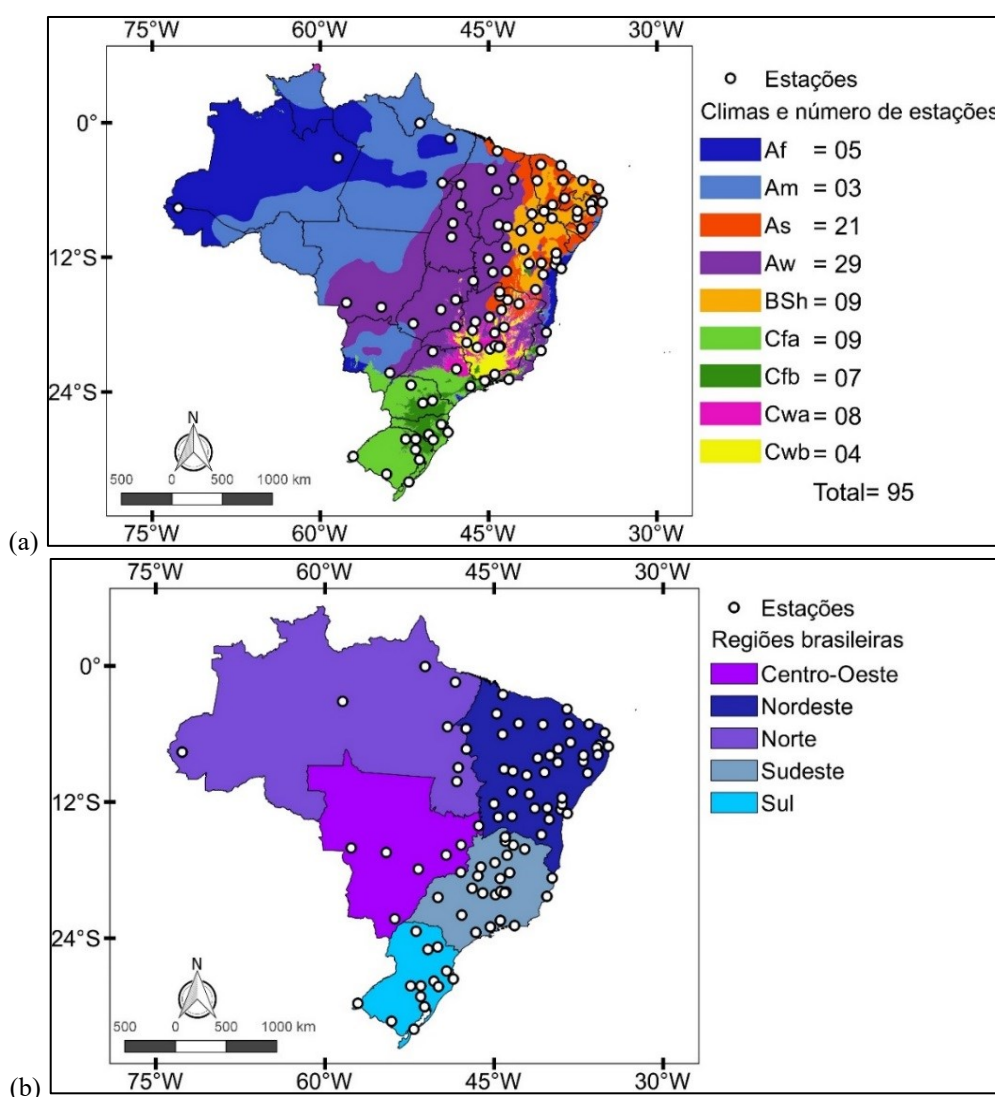


Figura 5 – Distribuição espacial das 95 estações meteorológicas automáticas do INMET selecionadas conforme os Critérios I, II e III, considerando: a) Os principais tipos climáticos brasileiros; e, b) Regiões brasileiras.

## CONCLUSÕES

Os Critérios I, II e III analisados permitiram selecionar consistentemente as estações com melhor qualidade de dados. Das 168 estações meteorológicas analisadas, 95 contemplaram os critérios adotados.

As estações meteorológicas localidades na região Norte apresentaram maior número de falhas, e conseqüentemente maior percentual de exclusão de dados e estações. Com a aplicação dos critérios, a região Centro-Oeste apresentou baixa cobertura espacial do território contento dados e estações confiáveis.

Os programas utilizados (Excel e *software* R para *boxplot*) foram eficientes para realização das análises e aplicação dos critérios para a seleção de banco de dados e estação meteorológica com qualidade.

## REFERÊNCIAS

ROSA, S. L. K. Alternativas para utilização e otimização do método ASCE Penman-Monteith na ausência de dados em quantidade e qualidade. Tese (Doutorado em Ciência do Solo). Universidade Federal do Paraná, Curitiba, 2023.

RStudio Team. **RStudio: Integrated Development for R**. RStudio, PBC, Boston, MA. URL: <http://www.rstudio.com/.2020>

SANTOS, A. A. **Evapotranspiração de referência horária e diária para o Estado do Paraná**. Dissertação (Mestrado em Agronomia, área de concentração Ciência do Solo) – Setor de Ciências Agrárias, Universidade Federal do Paraná. Curitiba, 2020. 214p.

SANTOS, A. A.; SOUZA, J. L. M.; ROSA, S. L. K. Hourly and Daily Reference Evapotranspiration with ASCE-PM Model for Paraná State, Brazil. **Revista Brasileira de Meteorologia**. 2021. ISSN 1982-4351. <https://doi.org/10.1590/0102-77863610009>

SANTOS, H. F. L.; COSTA, P. V. M.; ARAÚJO, M. L. L.; PIRES, F. O. T. S. B.; REGNIER, L.O. Que fonte de dados meteorológicos utilizar no Brasil? Que incerteza esperar? Uma comparação entre diferentes abordagens e variadas fontes de dados. VIII Congresso Brasileiro de Energia Solar – Fortaleza, 01 a 05 de junho de 2020. Disponível em: <https://anaiscbens.emnuvens.com.br/cbens/article/view/754/754>. Acesso em: 04 de abril de 2023.

TUKEY, J. W. **Exploratory Data Analysis**. Addison-Wesley, preliminary edition, 1970.

TUKEY, J. W. Box-and-Whisker Plots. In: **Exploratory Data Analysis**. Reading, MA: Addison-Wesley; 1977. p. 39-43

VALLADARES NETO, J. SANTOS, C. B.; TORRES, E. M.; ESTRELA, C. Boxplot: um recurso gráfico para a análise e interpretação de dados quantitativos. **Revista Odontologia Brasileira**. Central. ISSN 1981-3708. 2017; 26(76): p. 1-6

WICKHAM, H.; CHANG, W.; HENRY, L.; PEDERSEN, T. L; TAKAHASHI, K.; WILKE, C.; WOO, K.; YUTANI, H.; DUNNINGTON, D. **Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics**. R em interface RStudio, PBC, Boston, MA. URL: <https://ggplot2.tidyverse.org/>