UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
Acesso aberto

Ci. e Nat., Santa Maria, v. 43, e8, 2021 • https://doi.org/10.5902/2179460X43661
Submissão: 20/04/2020 • Aprovação: 13/08/2020 • Publicação: 30/01/2021

Meteorologia

Uso de parâmetros estatísticos para a classificação de regiões homogêneas de temperatura do ar

Statistical classification of homogenous surface temperature regions

Carlos Eduardo Salles de Araujo[I]

[I]Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina, Florianópolis, SC, Brasil

RESUMO

Séries temporais de temperatura horária de 146 estações meteorológicas do Estado de Santa Catarina foram utilizadas para avaliar se a representação compacta dos dados por meio de parâmetros estatísticos consegue subsidiar uma classificação de áreas homogêneas coerentes sob o ponto de vista físico. Nos resultados, a distribuição normal se ajustou muito bem as séries temporais de temperatura do ar, com um valor mediano de 0,9721 para o coeficiente de correlação de Pearson nas 146 estações meteorológicas empregadas na avaliação. As médias e desvios padrão obtidos pelo ajuste das funções foram utilizados como parâmetros de entrada para dois classificadores distintos: hierárquico e k-médias. Ambos classificadores separaram em quatro grupos distintos as estações meteorológicas de Santa Catarina. Estes grupos apresentaram relação direta com diferentes faixas de altitude e também com a influência da maritimidade. A classificação das estações meteorológicas em diferentes grupos homogêneos foi útil para identificar diferentes comportamentos climáticos de temperatura horária. Além da caraterização do clima em si, esta classificação pode ser utilizada como suporte na validação de modelos numéricos de previsão de tempo e também na identificação de séries temporais anômalas num contexto espacial regional.

Palavras-chave: Temperatura horária; Clima; Classificação estatística

ABSTRACT

Time series of hourly temperature from 146 weather stations located in Santa Catarina State – South Brazil were used to show that a compact data representation, using probability density functions (pdf) parameters, could be useful to classify homogeneous air temperature areas. The normal distribution fitted well the 146 weather stations temperature time series, presenting a median value of 0.9721 for the Pearson correlation coefficient. The means and standard deviations obtained by adjusting the Gaussian functions for the 146 stations were used as input parameters for two different classifiers: hierarchical and k-means. Both classifiers separated Santa Catarina's weather stations into four distinct groups. These groups had direct relationship with altitude ranges and with the influence of sea. The classification

of weather stations in different homogeneous groups was useful to identify climatic behaviors of hourly temperatures. In addition to the characterization of the climate itself, this classification can be useful as a support for the validation of numerical weather forecast models, and for the identification of abnormal temperature time series in a regional spatial context.

Keywords: Hourly temperature; Climate; Statistical classification

1 INTRODUÇÃO

O clima pode ser definido como o padrão dos diversos elementos atmosféricos numa região para um determinado período de tempo relativamente longo (VIANELLO; ALVES, 2013). Fatores como a latitude, altitude, exposição geográfica e a continentalidade influenciam nestes padrões em diferentes escalas espaciais e temporais (ROHLI; VEGA, 2012). A temperatura e a precipitação (KÖPPEN, 1936; Trewartha; Horn, 1980) e de forma menos frequente a umidade e a evapotranspiração (TORNTHWAITE, 1948) são os principais elementos utilizados na definição do clima.

As análises estatísticas constituem uma forma eficiente de se quantificar os elementos do clima incorporando as incertezas inerentes as limitações dos diferentes sistemas de monitoramento. O uso de valores médios de temperatura e acumulados de precipitação para cada localidade ou ponto de grade são frequentemente utilizados para definição e espacialização dos diferentes tipos de clima. Alvares *et al*. (2013) e Sá Junior *et al*. (2012) são exemplos de como estes dados são empregados para a classificação climática pelo sistema de Köppen (KÖPPEN, 1936). Além das médias, as amplitudes térmicas (UNAL *et al*., 2003) e os maiores e menores valores mensais de precipitação (PEEL *et al*., 2007) são utilizados para a definição de zonas climáticas.

As técnicas de classificação envolvem tradicionalmente o uso combinado de faixas de valores limite para as variáveis ou elementos que definem o clima (Trewartha; Horn, 1980). Com o aperfeiçoamento e a disseminação dos métodos computacionais e o aumento da disponibilidade de dados ambientais, técnicas estatísticas mais elaboradas passaram a ser utilizadas para a classificação climática, conforme relatado por Cannon (2012). Entre estas técnicas, os métodos não

supervisionados, como a análise de agrupamentos de amostras similares (clusters) tem sido amplamente empregada por diversos autores como: Fovell e Fovell (1993), Yao (1997), Gerstengarbe *et al*. (1999), Unal *et al*. (2003), Zscheischler *et al*. (2012), Zhang e Yan (2014).

A qualidade da caracterização climática depende em grande parte da qualidade dos dados de entrada. Dentre todos os tipos de sistemas de monitoramento, os dados das estações meteorológicas de superfície (EMS) são os mais utilizados para a caracterização climática. Existem testes de controle de qualidade que avaliam dados de um único sítio (MEEK; HATFIELD, 1994; ZAHUMENSKY, 2004) e testes desenvolvidos para comparar dados de uma estação contra estações vizinhas (HUBBARD, 2001).

Na comparação com estações vizinhas, diferenças significativas entre conjuntos de dados podem estar relacionadas a variações climáticas locais e não necessariamente a dados espúrios. Sobre estas variações, Daly (2006) ressalta a influência dos fatores climáticos em escalas de 1 km até escalas superiores a 100km. Nesse contexto, a principal questão da utilização dos dados de uma EMS se resume em determinar se a referida estação representa o comportamento climático da região na qual está inserida para a escala espacial da análise desejada. Nem sempre é uma questão simples de ser respondida pois está relacionada tanto com a variabilidade da fisiografia quanto com a variabilidade da dinâmica atmosférica.

1.1 Objetivos

Este trabalho se propõe a realizar um teste metodológico para avaliar se a representação compacta dos dados das estações meteorológicas de superfície (EMS), por meio de parâmetros estatísticos, consegue subsidiar uma classificação de áreas homogêneas coerentes sob o ponto de vista físico. Os objetivos específicos estão divididos em duas etapas distintas: (a) ajustar funções densidade probabilidade as distribuições de frequências simples de temperatura horária para as estações meteorológicas de Santa Catarina, avaliando-se o grau de representatividade destas

funções em relação as séries temporais; (b) classificar as estações meteorológicas por dois métodos estatísticos distintos: árvore hierárquica binária e k-médias, utilizando como variáveis de entrada dos classificadores os parâmetros das funções densidade probabilidade de temperatura horária de cada estação.

2 METODOLOGIA

Um conjunto de dados horários de temperatura média do ar de 156 estações meteorológicas automáticas e telemétricas do estado de Santa Catarina foi selecionado do banco de dados ambiental da EPAGRI para o período de 02/03/2018 a 11/06/2019. Este período foi escolhido de forma coincidente e para permitir futuras comparações com a previsão por conjuntos realizada pelo modelo WRF na EPAGRI. Os dados do banco são consistidos por um sistema de controle de qualidade automático que elimina valores espúrios, persistência de valores, e variações abruptas não realistas (MASSIGNAM *et al*., 2016). Adicionalmente foram eliminadas as estações com séries temporais menores que 7472 horas (311.33 dias), correspondente a dois terços da maior série temporal disponível, mantendo 146 estações para análise dos dados.

Para cada estação foi ajustada uma função densidade probabilidade gaussiana à série temporal de temperaturas horárias, avaliando-se o coeficiente de correlação de Pearson (ρ). A distribuição normal (ou gaussiana) é uma família de curvas de dois parâmetros: media e desvio padrão (Evans *et al*., 1993). Desta forma, as séries temporais de cada estação que apresentaram valores altos ($0.70 < \rho <= 0.90$) ou muito altos ($\rho > 0.90$) de correlação com a função gaussiana passaram a ser representadas de forma compacta por estes dois parâmetros.

A partir das médias e desvios padrão de temperatura de todas as estações construiu-se uma árvore hierárquica binária aglomerativa, utilizando-se como métrica o método da mínima variância (Ward, 1963). Esta árvore invertida

representa um método de classificação de agrupamentos (clusters) naturais dos dados onde cada estação é inicialmente seu próprio cluster. Pares de clusters são então agrupados em cada nível hierárquico da árvore, formando novos clusters à medida que a variância aumenta e os ramos da árvore se deslocam para um nível hierárquico superior.

Numa classificação paralela e para se verificar a robustez da classificação hierárquica inicial empregou-se o método de agrupamento por k-médias, utilizando-se a distância euclidiana ao quadrado como métrica. Este método funciona através de um processo de minimização de uma função, calculando-se um valor que resulta da soma das distâncias entre cada observação e o centro do grupo a que estão atualmente alocadas (Arthur; Vassilvitskii, 2007).

Com base neste valor o método iterativamente desloca as observações de um grupo para o outro de modo a minimizar o valor desta função. Para um conjunto grande de dados o agrupamento por k-médias é frequentemente mais apropriado do que a árvore hierárquica pois opera diretamente nas observações (ao invés de medidas de dissimilaridade) e particiona os dados em k agrupamentos mutuamente exclusivos, criando um único nível de agrupamentos (Arthur; Vassilvitskii, 2007).

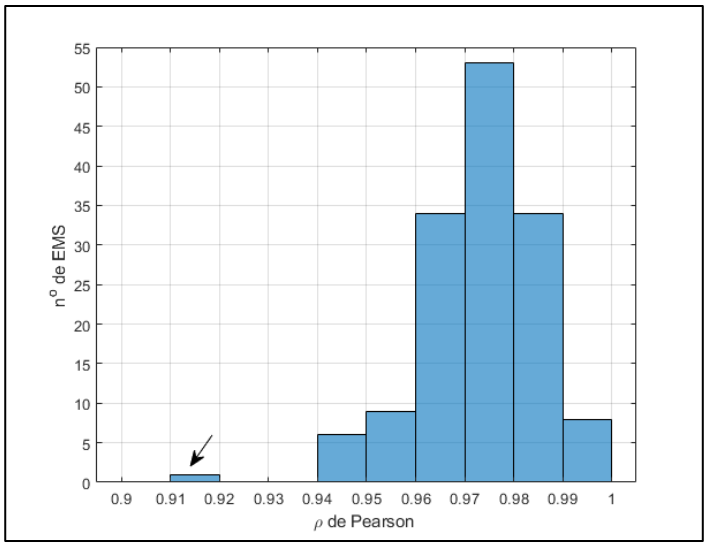
Um dos problemas do método k-médias é descobrir o valor ótimo para o número de grupos (clusters) que devem ser formados. Neste trabalho utilizou-se o cálculo do critério gap (lacuna) proposto por Tibshirani *et al*. (2001) realizando-se testes para o número de agrupamentos variando de um a oito. O valor ótimo é obtido pela maximização do critério gap.

Finalmente, utilizando-se o resultado da classificação por k-médias para o número ótimo de clusters, verificou-se a distribuição espacial dos grupos de estações identificados. Esta distribuição foi associada a diferentes faixas de altitude no Estado de Santa Catarina.

3 RESULTADOS

De acordo com a interpretação de Hinkle *et al*. (2003) a função gaussiana (distribuição normal) mostrou excelente ajuste para as séries de temperatura horária nas 146 estações meteorológicas, apresentando valores do coeficiente de correlação de Pearson (ρ) variando entre 0,9136 e 0,9924 e valor mediano de 0,9721. A Figura 1 apresenta o histograma de distribuição de frequências simples dos coeficientes de correlação de Pearson por número de EMS.

Figura 1 – Histograma de distribuição de frequências simples dos coeficientes de correlação de Pearson por número de SEM



A Figura 1 revela os valores de ρ concentrados no entorno do valor mediano. O valor mínimo observado (marcado com uma seta na Figura 1) é uma exceção a esse padrão e corresponde a EMS da Comunidade XV Dias, situada a 1220 metros de altitude, no município de Bom Jardim da Serra - SC. A Figura 2 mostra que a distribuição de frequências das temperaturas horárias da EMS situada na Comunidade XV Dias possui forma leptocúrtica (curtose maior que a distribuição normal) e assimetria ligeiramente negativa.

Figura 2 – Histograma de distribuição de frequências das temperaturas horárias da EMS da Comunidade XV Dias. A linha vermelha representa a função gaussiana ajustada a esta distribuição

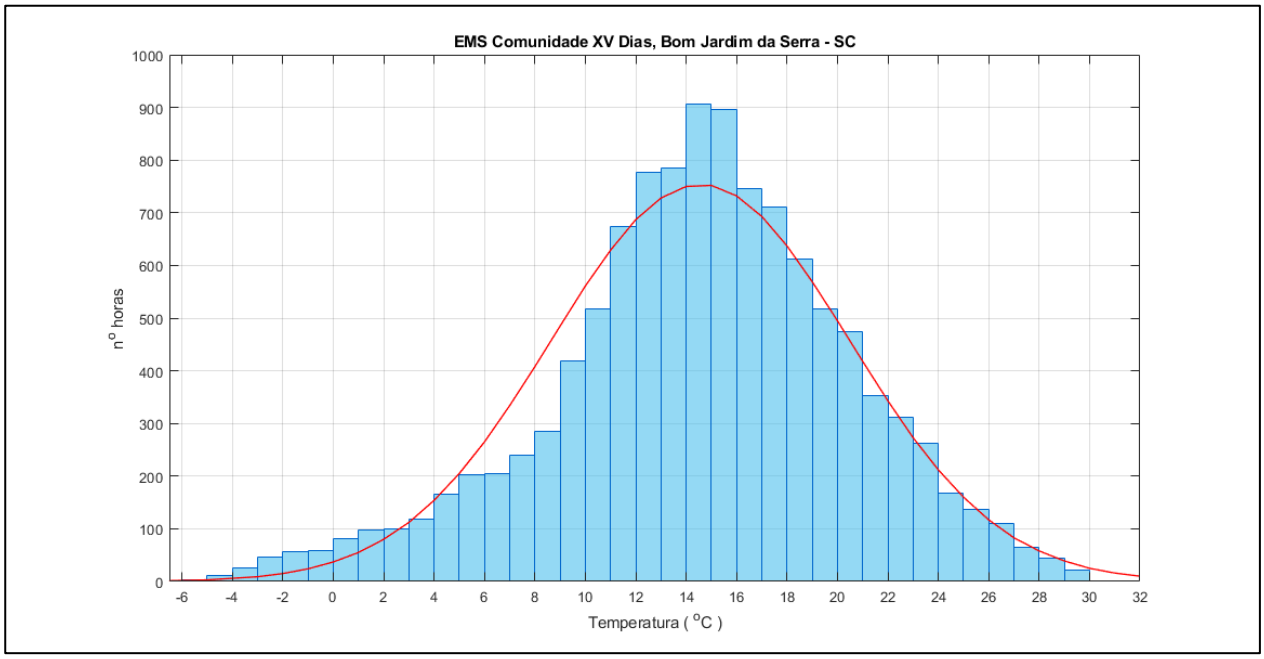
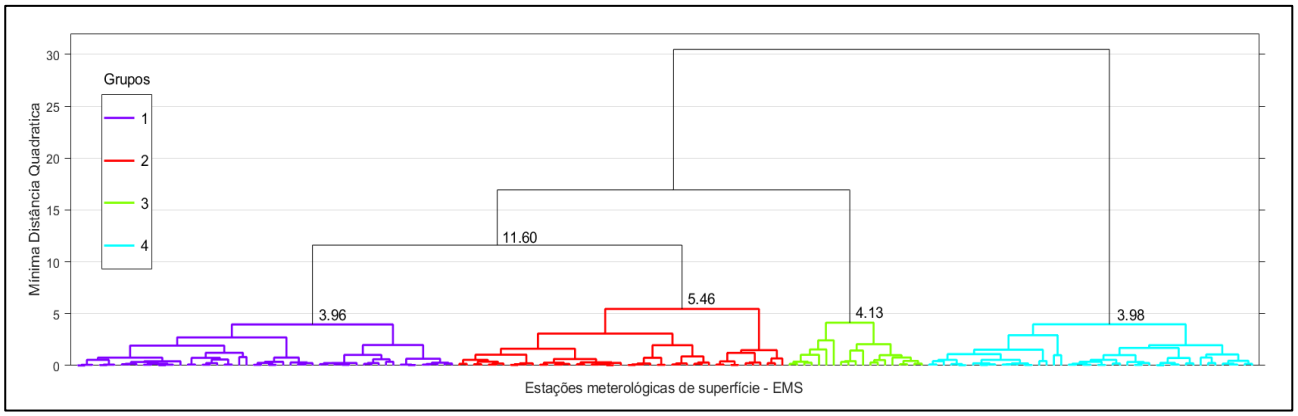


Figura 3 – Dendrograma das variáveis média e desvio padrão das séries temporais de temperatura média horária



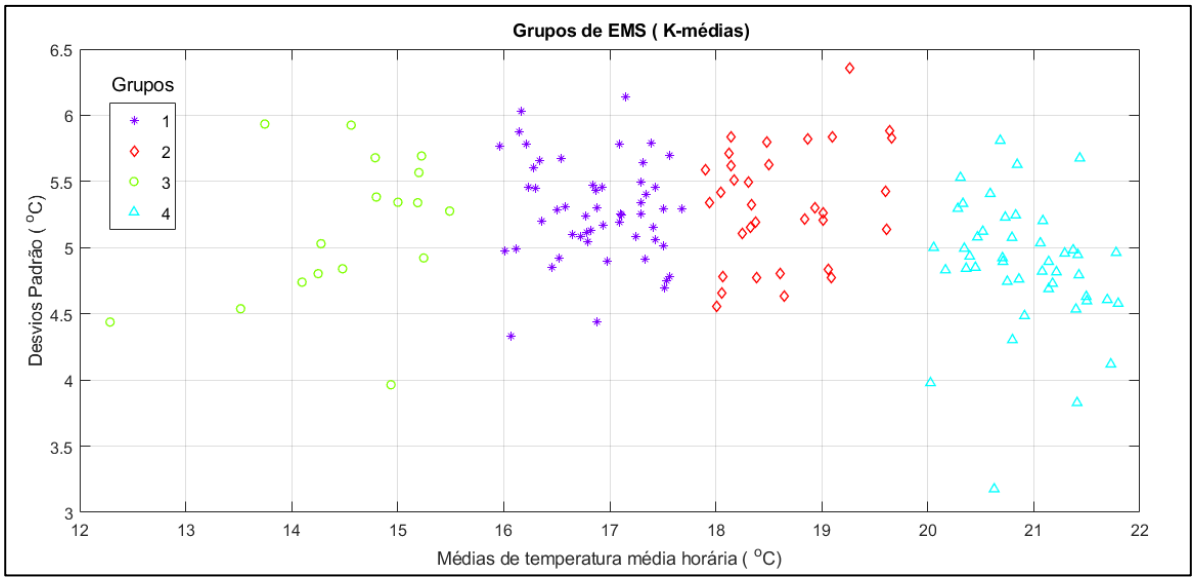
O dendrograma (Figura 3) apresenta a classificação por árvore hierárquica binária das EMS de Santa Catarina. As médias e desvios padrão de temperatura horária das estações meteorológicas foram utilizadas como parâmetros de entrada para a

classificação. A altura de cada segmento em forma de "colchete deitado" representa a distância entre os centroides dos grupos são conectados dois a dois.

Com base nas distâncias verticais entre os diferentes nós da Figura 3 é possível identificar quatro grupos distintos. Para ressaltar essa observação os grupos foram realçados com diferentes cores. Observe que o valor mínimo de variância entre os quatro grupos ocorre na junção dos grupos 1 e 2 e apresenta um valor elevado (11,60), correspondendo a mais que o dobro do valor máximo (5,46) da variância interna do grupo 2.

A Figura 4 apresenta a dispersão das estações meteorológicas no plano cartesiano formado pelos eixos média e desvio padrão das temperaturas. O critério "gap" revelou como quatro o número ótimo de grupos na classificação por k-médias, coincidindo com a classificação hierárquica. Para a visualização destes grupos utilizou-se na Figura 4 o mesmo padrão de cores adotado na Figura 3.

Figura 4 – Dispersão das estações meteorológicas no plano cartesiano formado pelos eixos média e desvio padrão das temperaturas. As cores indicam a qual cluster da classificação por k-médias cada EMS pertence

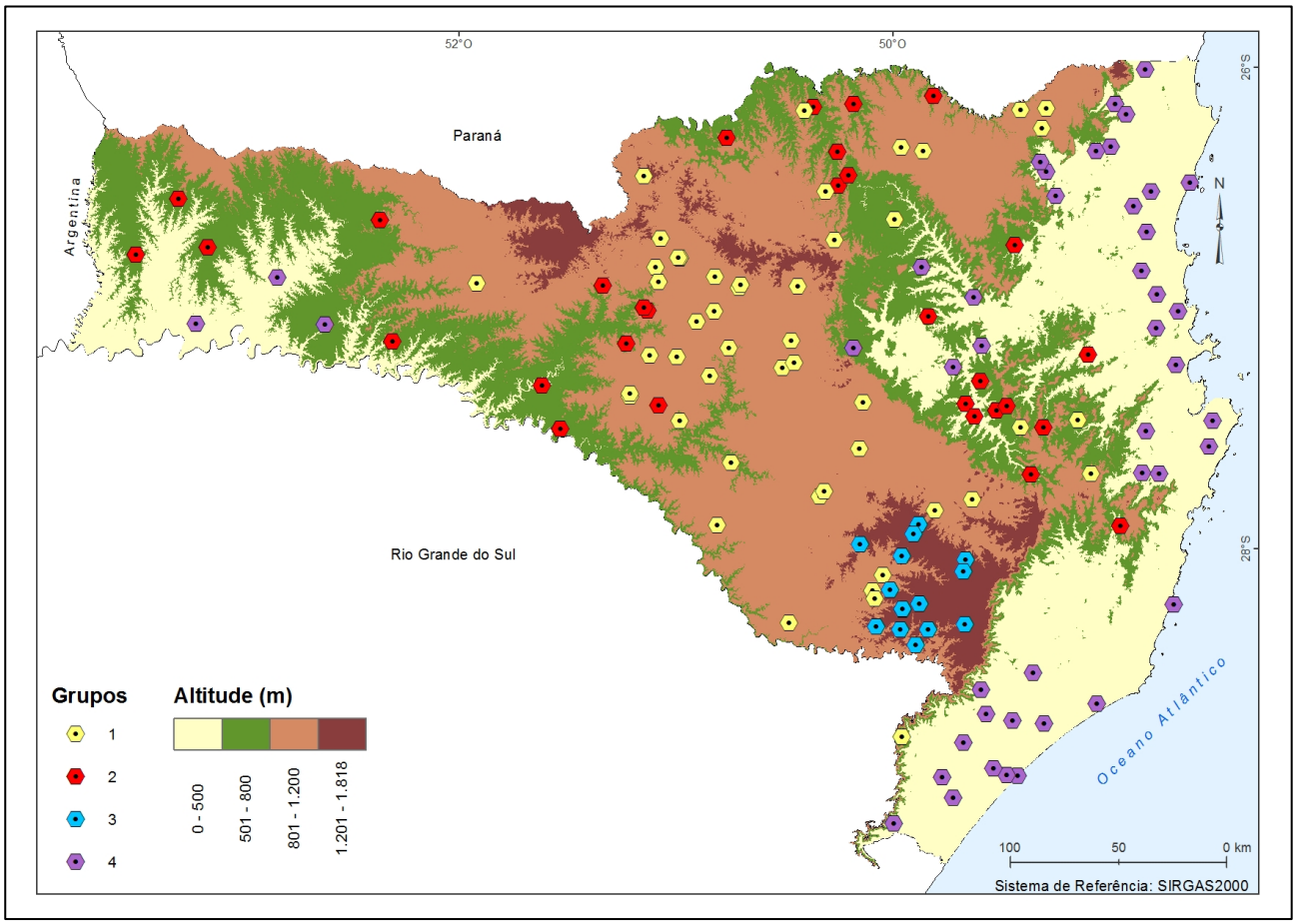


A Tabela 1 mostra os valores de média e desvio padrão da classificação por k-médias para o centroide dos quatro clusters da Figura 4 e também as respectivas faixas de altitude em que se encontram as EMS de cada grupo.

Tabela 1 – Média e desvio padrão dos centroides dos clusters e faixas de altitude

Cluster	**Centroide do cluster**		**Faixa de altitude (m)**
	Média (°C)	Desvio (°C)	
1	16.89	5.26	801-1200
2	18.49	5.35	501-800
3	14.62	5.21	1201-1818
4	20.91	4.92	0-500

Figura 5 – Localização geográfica das estações meteorológicas por grupos (classes) e faixas de altitude para Santa Catarina



A Figura 5 apresenta um mapa da distribuição espacial das estações para os quatro grupos classificados pelo algoritmo k-médias. Na mesma Figura 5 é possível notar que os grupos apresentam uma forte identificação com diferentes faixas de altitude do Estado de Santa Catarina revelando não apenas uma homogeneidade estatística, mas também uma coesão espacial.

4 DISCUSSÕES

A distribuição normal se ajustou muito bem as séries temporais horárias de temperatura das estações meteorológicas de Santa Catarina, permitindo utilizar os parâmetros da curva normal (média e desvio padrão) para classificar as EMS em diferentes grupos.

De uma forma geral a altitude foi o fator dominante na separação entre as quatro classes. Analisando de forma conjunta o dendrograma (Figura 3) e o mapa (Figura 5) é possível notar que as estações situadas na faixa de altitude entre 801m e 1200m (grupo 1) são mais similares as estações situadas na faixa entre 501m e 800m (grupo 2). Do ponto de vista fisiográfico, as estações do grupo 1 estão associadas predominantemente as regiões de planalto enquanto as estações do grupo 2 estão associadas predominantemente as áreas de encostas e vales encaixados.

As estações da serra catarinense (grupo 3) possuem um comportamento de temperatura horária diferente em comparação aos dois primeiros grupos. As estações do litoral, vales do rio Itajaí e do rio Uruguai (grupo 4) são as mais diferentes dentre todos os grupos.

A análise da Tabela 1 e da Figura 4 ajudam a explicar estas diferenças. No grupo 4 a média de temperatura das estações é mais elevada e o desvio padrão mais baixo, por causa de fatores como a própria altitude baixa e a maritimidade (a maior parte das estações deste grupo está próxima ao litoral) que reduz as amplitudes diárias. Na serra (grupo 3), por causa das elevadas altitudes, as temperaturas são mais baixas, embora os desvios sejam próximos aos verificados nos grupos 1 e 2.

A grande variedade e complexidade das formas do relevo (fisiografia) de Santa Catarina se reflete na distribuição espacial da classificação das estações. Considerando o contexto das faixas de altitude é interessante notar na Figura 5 que algumas EMS parecem ter sido classificadas de forma errônea. De fato, isso não ocorre, sendo apenas um artifício da escala de visualização. O mesmo comportamento pode ser observado nas faixas de transição entre as classes de altitude, onde estações distantes menos de 10 km entre si pertencem a diferentes classes.

Essa observação corrobora a discussão da importância da escala de análise espacial para considerar os dados de uma EMS como representativos de um comportamento regional. Em regiões que possuem uma alta densidade de EMS é possível selecionar apenas aquelas que atendam a algum critério de homogeneidade. Além da própria caracterização da região essa seleção pode ser extremamente útil para a realização de comparações com outras bases de dados.

As informações das EMS são de natureza pontual no espaço, enquanto as informações obtidas por meio das imagens de satélites e dos modelos numéricos de previsão de tempo são representativas da área de uma célula, correspondendo a um elemento de grade no modelo ou a um pixel na imagem. Os dados medidos pelas EMS são usualmente assimilados pelos modelos numéricos e posteriormente empregados como referência nas avaliações de desempenho dos mesmos (HAIDEN *et al*., 2018). São utilizados também na validação dos produtos de satélite (POVEY; GRAINGER, 2015). Em ambos os casos é importante garantir a qualidade e homogeneidade destes dados de referência, minimizando as fontes de erro dessas avaliações.

5 CONCLUSÕES

A representação dos dados por meio dos parâmetros de uma função, média e desvio padrão no caso da gaussiana, é uma forma compacta e computacionalmente eficiente para se trabalhar com a classificação de um número grande de EMS. A virtude

deste método é também sua principal deficiência porque não é possível considerar na classificação as influências temporais, sejam elas sazonais ou interanuais.

Esta representação compacta é útil para a classificação estatística das estações em grupos homogêneos. As classificações realizadas para a temperatura horária, tanto pelo método hierárquico como por k-médias, separaram em quatro grupos distintos as estações meteorológicas de Santa Catarina. Estes grupos têm relação direta com diferentes faixas de altitude e também com a influência da maritimidade.

Embora a temperatura seja apenas um dos elementos definidores do clima, A classificação das estações meteorológicas em diferentes grupos homogêneos é útil para identificar diferentes comportamentos climáticos. Além da caraterização do clima em si, esta classificação pode ser utilizada como ferramenta de a suporte na validação de modelos numéricos de previsão de tempo e produtos de satélite.

Em trabalhos futuros pretende-se avaliar a aplicação desta metodologia a variáveis que se ajustam a outras distribuições paramétricas como a chuva (Gama ou Gumbel, por exemplo) e ao vento (Weibull). Deve-se também avaliar diferentes escalas temporais, como a horária, diária e decendial. Finalmente, após concluída estas avaliações pretende-se fazer uma classificação considerando os parâmetros estatísticos de duas ou mais variáveis meteorológicas concomitantemente, como chuva e temperatura, por exemplo.

REFERÊNCIAS

Alvares CA, Stape JL, Sentelhas PC, de Moraes G, Leonardo J, Sparovek G. Köppen's climate classification map for Brazil**. Meteorologische Zeitschrift**. 2013;22(6):711-28.

ARTHUR D, VASSILVITSKII S. K-means: The Advantages of Careful Seeding. *In:* Proceedings of the Eighteenth Annual ACM-SIAM **Symposium on Discrete Algorithms**. 2007, p. 1027–35.

Cannon, AJ. Köppen versus the computer: Comparing Köppen-Geiger and multivariate regression tree climate classifications in terms of climate homogeneity. Hydrol. **Earth Syst. Sci**. 2012;(16):217–29.

Daly C. Guidelines for assessing the suitability of spatial climate data sets, Int. **J**. **Climatol**. 2006; (26):707–721.

EVANS M, HASTINGS N, PEACOCK B. Statistical Distributions. 2.ed. **Hoboken**, NJ: John Wiley & Sons, Inc., 1993.

Haiden T, Dahoui M, Ingleby B, Rosnay P de, Prates C, Kuscu E, *et al*. Use of in situ surface observations at ECMWF. ECMWF Tech Memo [Internet]. 2018;(November). Available from: https://www.ecmwf.int/node/18748.

Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: **Houghton Mifflin**; 2003.

Fovell RG, Fovell MYC. Climate zones of the conterminous United States defined using cluster analysis. **J**. **Climate**. 1993;(6):2103–35.

Gerstengarbe F, Werner P, Fraedrich K. Applying Non-Hierarchical Cluster Analysis Algorithms to Climate Classification: Some Problems and their Solution. **Theor. Appl. Climatol**. 1999;(64):143–150.

Hubbard KG. 2001 Multiple station quality control procedures. *In:* **Automated Weather Stations for Applications in Agriculture and Water Resources Management** - AGM-3 WMO-TD1074. 2001, p.133-136.

MASSIGNAM AM, ANTUNES EN, MARASCHIN F. Banco de Dados Agrometeorológicos. *In:* Silva, E. **Boletim Ambiental Síntese Trimestral** - Inverno 2015. Florianópolis: Epagri, 2016. 51 p. (Documento, 253).

MEDEIROS M, MUNARI DB, BEZERRA ALQ, ALVES MA. Pesquisa qualitativa em saúde: implicações éticas. *In:* Ghilhem D, Zicker F, editors. **Ética na pesquisa em saúde: avanços e desafios**. Brasília: Letras Livres UnB; 2007. p. 99-118.

Meek DW, Hatfield JL. Data quality checking for single station meteorological databases**. Agric. For. Meteor.** 1994; (69):85-109.

KÖPPEN W. Das geographisca System der Klimate. *In:* Köppen W, Geiger R, editors. **Handbuch der Klimatologie**. Berlin*:* Gebrüder Borntraeger; 1936. 1C, p. 1–44.

PEEL MC, FINLAYSON BL, McMAHON, TA. Updated world map of the KÖPPEN-Geiger climate classification. Hydrol. **Earth Syst. Sci.** 2007; (11):1633–44.

Povey AC, Grainger RG. Known and unknown unknowns: Uncertainty estimation in satellite remote sensing. **Atmos. Meas. Tech**. 2015;8(11):4699–718.

ROHLI RV, VEGA AJ **Climatology**. 2nd ed. Jones & Bartlett Learning, 2012.

SÁ JUNIOR A., CARVALHO LG, SILVA FF, ALVES, MC Application of the KÖPPEN classification for climatic zoning in the state of Minas Gerais, Brazil. **Theor. Appl.Climatol.** 2012;(108):1–7.

THORNTHWAITE CW. An approach toward a rational classification of climate. **Geogr. Rev**. 1948;(38):55–94.

TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the number of clusters in a data set via the gap statistic. **J. R. Stat. Soc. Ser. B Methodol**. 2001, 63(2): 411–423.

Trewartha GT, Horn LH **Introduction to climate**, 5th ed. McGraw Hill, New York, NY, 1980.

Unal Y, Kindap T, Karaca M. Redefining the climate zones of Turkey using cluster analysis. Int **J Climatol**. 2003 Jul;23(9):1045–55.

VIANELLO R. ALVES. AR **Meteorologia básica e aplicações. Viçosa**, MG: UFV. 1991.

WARD JH. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 1963;(58):236–244.

Yao CS. A new method of cluster analysis for numerical classification of climate. **Theor. Appl. Climatol**. 1997;(57):111–118.

Zahumenský I. Guidelines on Quality Control Procedures for Data from Automatic Weather Stations. **World Meteorol**. **Organ**. 2004. Available from: https://www.wmo.int/pages/prog/www/IMOP/meetings/Surface/ET-STMT1_Geneva2004/Doc6.1(2).pdf .

Zhang X, Yan X. Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. **Climate Dyn**. 2014;(43):595–605.

Zscheischler, J., Mahecha MD, Harmeling S. Climate classifications: The value of unsupervised clustering. **Procedia Comput.** Sci. 2012;(9):897–906.

Contribuições de autoria

1 – Carlos Eduardo Salles de Araujo
Doutorado em Engenharia, Pesquisador EPAGRI/CIRAM
https://orcid.org/0000-0002-6377-8536 - kadu_araujo@epagri.sc.gov.br
Contribuição: Concepção, processamento e análise dos dados, escrita, confecção de figuras e tabelas, submissão e revisão.

Como citar este artigo

ARAUJO, C. E. S. Uso de parâmetros estatísticos para a classificação de regiões homogêneas de temperatura do ar. **Ciência e Natura**, Santa Maria, v. 43, e8, p. 1-14, 2021. Available from: https://doi.org/10.5902/2179460X43661. Accessed: Month Abbreviated. day, year.