ROSANE SOARES MOREIRA VIANA

O USO DA GEOESTATÍSTICA ESPAÇO-TEMPORAL E APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DA TEMPERATURA MÁXIMA DO AR

VIÇOSA
MINAS GERAIS – BRASIL
2019

ROSANE SOARES MOREIRA VIANA

O USO DA GEOESTATÍSTICA ESPAÇO-TEMPORAL E APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DA TEMPERATURA MÁXIMA DO AR

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
V614u
2019

Viana, Rosane Soares Moreira, 1969-
O uso da geoestatística espaço-temporal e aprendizagem de máquina na predição da temperatura máxima do ar / Rosane Soares Moreira Viana. – Viçosa, MG, 2019.
xiv, 106 f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Gérson Rodrigues dos Santos.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Geologia - Estatísticas. 2. Dados geoespaciais. 3. Covariância. 4. Gráficos aelatórios. 5. Análise de regressão. 6. Máquina de vetores de suporte. 7. Algorítmos. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 551.021

ROSANE SOARES MOREIRA VIANA

O USO DA GEOESTATÍSTICA ESPAÇO-TEMPORAL E APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DA TEMPERATURA MÁXIMA DO AR

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 21 de fevereiro de 2019.

Lidiane Maria Ferraz Rosa

Adriana Maria R. Trancoso Santos

Leandro Roberto de Macêdo

Danilo Pereira Barbosa

Gerson Rodrigues dos Santos
(Orientador)

Aos meus pais, Antônio (in memorian) e Maria;
ao meu esposo Dênio;
aos meus filhos Taiane e Luis Felipe;
aos meus irmãos;
A Deus.

dedico.

AGRADECIMENTOS

À Universidade Federal de Viçosa (UFV) e ao departamento de Matemática da UFV pela oportunidade oferecida em fazer o doutorado com afastamento parcial;

Ao Instituto Nacional de Meteorologia (INMET) pela disponibilização das bases de dados meteorológicos históricos inseridos no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP);

Ao professor Dr. Gérson Rodrigues dos Santos, pela orientação e pelo constante estímulo transmitido durante todo o trabalho;

Aos meus pais por me ensinarem valores como ética, honestidade, humildade, perseverança e solidariedade.

Ao meu esposo Dênio e aos meus filhos Taiane e Luis Felipe, com meus pedidos de desculpas pelos inúmeros momentos de ausência, ainda que fisicamente presente, pela compreensão, apoio, palavras de incentivo e principalmente pela paciência e amor;

Ao professor Dr. Demerval Soares Moreira, por me socorrer nas horas de desespero, pelas constantes discussões durante a elaboração deste trabalho e pelas inúmeras ajudas com a programação do software R;

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria da UFV por contribuírem para minha formação;

Aos meus coorientadores, professores Dr. João Marcos Louzada e Dr. Paulo César Emiliano pelos ensinamentos e pelas contribuições neste trabalho;

Aos membros da banca avaliadora, pelas considerações que contribuíram para melhorar o texto final;

A todos meus amigos que, embora não estivessem presentes diariamente, tenho certeza que torceram e oraram por mim;

Aos meus irmãos, cunhados e sobrinhos, pela amizade, carinho e pelo convívio que, por várias vezes, trouxeram momentos de descontração e contribuíram com palavras e pensamentos positivos para a conclusão desse trabalho;

A Deus por tudo!

"O sucesso nasce do querer, da determinação e persistência
em se chegar a um objetivo. Mesmo não atingindo o alvo,
quem busca e vence obstáculos, no mínimo fará
coisas admiráveis."

(José de Alencar)

SUMÁRIO

LISTA DE FIGURAS

INTRODUÇÃO GERAL

CAPÍTULO 1

CAPÍTULO 2

APÊNDICE

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

RESUMO

VIANA, Rosane Soares Moreira, D.Sc., Universidade Federal de Viçosa, fevereiro de 2019. **O uso da geoestatística espaço-temporal e aprendizagem de máquina na predição da temperatura máxima do ar.** Orientador: Gérson Rodrigues dos Santos. Coorientadores: João Marcos Louzada e Paulo César Emiliano.

Dados espaço-temporais são caracterizados pela descrição da variabilidade no tempo e no espaço. Atualmente, os estudos desses tipos de dados têm proporcionado grandes avanços em áreas como ciências ambientais, geofísicas, biologia, epidemiologia e outras. Os procedimentos comuns de estatística, frequentemente, não são suficientes para descrever os processos espaço-temporais, pois não conseguem captar a variabilidade nas dimensões espaço e tempo conjuntamente. Para estes processos existem três tipos de abordagem: análise puramente espacial, que considera cada tempo separadamente, ou seja, desconsidera a dependência temporal e analisa os dados do processo utilizando técnicas usuais de estatística espacial para cada tempo; análise puramente temporal, onde cada localização desconsidera-se a dependência espacial e analisa os dados do processo utilizando técnicas usuais de séries temporais; e análise espacial e temporal, que é capaz de analisar conjuntamente tanto as dependências espaciais quanto as temporais existentes no conjunto de dados. Ainda não existe um consenso sobre quais são as técnicas mais adequadas de modelagem que atendem às necessidades de aplicações que envolvam simultaneamente tempo e espaço. O desenvolvimento destas técnicas e a construção de representações computacionais apropriadas é um dos grandes desafios da geoinformação. Desta forma, este trabalho tem como objetivo fazer uma exposição teórica de algumas metodologias disponíveis na geoestatística espaço-temporal e/ou aprendizagem de máquina, bem como utilizar um conjunto de dados reais para fazer predição via estrutura de funções de covariâncias espaço-temporais e via modelos de regressão baseados em aprendizagem de máquina, em especial, os algoritmos de random Forest e support vector machine.

ABSTRACT

VIANA, Rosane Soares Moreira, D.Sc., Universidade Federal de Viçosa, February, 2019. **The use of space-time geostatistics and machine learning in the prediction of maximum air temperature.** Advisor: Gérson Rodrigues dos Santos. Co-advisors: João Marcos Louzada and Paulo César Emiliano.

Spatial-temporal data are characterized by the description of variability in time and space. Currently, studies of these types of data has provided great advances in areas such as environmental sciences, geophysics, biology, epidemiology and others. Common statistical procedures are often not sufficient to describe spatio-temporal processes because they fail to capture the variability in space and time dimensions together. For these processes there are three types of approaches: purely spatial analysis, which considers each time separately, ie, disregards the temporal dependence and analyzes the process data using usual techniques of spatial statistics for each time; purely temporal analysis, where each location is disregarded the spatial dependence and analyzes the data of the process using usual techniques of time series; and spatial and temporal analysis, which is able to analyze both spatial and temporal dependencies in the dataset together. There is still no consensus on which are the most appropriate modeling techniques that meet the needs of applications that involve both time and space. The development of these techniques and the construction of appropriate computational representations is one of the great challenges of geoinformation. In this way, this work has as objective to make a theoretical exposition of the available methodologies in the space-time geostatistics and machine learning, as well as to use a real data set to make prediction via space-time covariance function structures and via regression models based on machine learning, especially the algorithms of random forest and support vector machine.

INTRODUÇÃO GERAL

Um breve histórico da Geoestatística foi apresentado por Landim (2006). Segundo esse autor, a teoria das variáveis regionalizadas, que é o princípio da Geoestatística, surgiu em uma série de publicações realizadas por Matheron (1962, 1963, 1965 e 1971), inspirado nos trabalhos de Krige (1951) e De Wijs (1951, 1953). Krige (1951) trabalhou com dados de mineração e concluiu que para descrever a variabilidade dos dados foi preciso considerar a distância entre os locais que foram observadas as amostras.

Dados espaciais podem ser tratados pela ciência denominada Geoestatística e eles se encontram em diversas áreas, tais como agricultura de precisão, meteorologia e geologia.

Duas ferramentas fundamentais da geoestatística são o variograma e a krigagem. O variograma é empregado para descrever a dependência espacial dos dados, isto é, determinar a distância dentro da qual as amostras apresentam-se correlacionadas espacialmente. Já a Krigagem, segundo Landim (2006), é um método de estimativa de valores de um atributo disseminado no espaço e/ou tempo, com base em valores vizinhos correlacionados espacialmente pela análise variográfica.

Em diversos processos físicos seus dados espaciais também variam ao longo do tempo, estando assim associados ao momento em que foram observados. Esses são denominados dados espaço-temporais.

Os modelos espaço-temporais estão associados a vários campos da ciência, tais como: Hidrologia (ROUHANI, WACKERNAGEL, 1990; GOOVAERTS, SONNET, 1993; VAROUCHAKIS, 2018); Meteorologia (CRESSIE, HUANG, 1999; DE IACO, MYERS, POSA, 2002; STEIN, 2005; GNEITING, GENTON, GUTTORP, 2006; RAJA et al., 2017; MOREIRA et al., 2017); Meio Ambiente (DE CESARE, MYERS, POSA, 1997; CHRISTAKOS, VYAS, 1998; PAEZ, GAMERMAN, 2005; GRÄLER, PEBESMA, HEUVELINK, 2016), Ciência do Solo (SNEPVANGERS, HEUVELINK, HUISMAN, 2003) e Agricultura de Precisão (COELHO, 2005; MANTOVANI et al., 2007).

Segundo Cressie e Wikle (2011), a questão é se o aspecto temporal tem importância para a análise do conjunto de dados espaciais ou não. Em algumas aplicações, a componente temporal simplesmente é descartada e o mesmo pode acontecer com a componente espacial.

Segundo Schabenberger e Gotway (2005), as análises separadas no tempo (espaço) permitem previsões apenas no espaço (tempo) e os procedimentos dessa geoestatística clássica frequentemente não são suficientes para descrever os processos espaço-temporais, pois não conseguem captar a variabilidade nas dimensões espaço e tempo conjuntamente.

A busca constante do desenvolvimento de modelos espaço-temporais capazes de descrever de forma adequada os processos do mundo real é um dos grandes desafios da geoinformação. Outro grande desafio é encontrar um equilíbrio entre o desenvolvimento e o meio ambiente a fim de reduzir os impactos ambientais causados pelo homem tal como: ilha de calor, poluição do ar, desmatamentos, efeito estufa e aquecimento global.

Com os acentuados avanços tecnológicos e programas computacionais que permitem trabalhar com numerosos dados disponibilizados por órgãos governamentais e institutos de pesquisa, técnicas de análise conjunta espaço-temporal estão sendo desenvolvidas no intuito de levar em consideração as interações entre os componentes espaciais e temporais.

Segundo Gräler, Pebesma e Heuvelink (2016), o processo de levar em consideração as observações feitas em outros momentos pode oferecer melhorias para a estimativa e previsão de parâmetros, bem como mostrar a evolução do fenômeno ao longo do tempo para poder prever o futuro.

Entre os procedimentos para análise de dados espaço-temporais destaca-se a geoestatística espaço-temporal via modelagem da estrutura de covariância. A Figura 1 descreve tipos de abordagens da Estatística Espacial até convergir para a modelagem por estrutura de covariância.

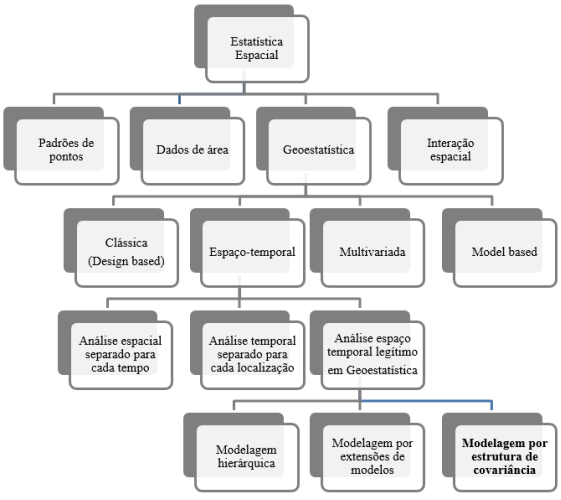


Figura 1 – Abordagens da Estatística Espacial até chegar a modelagem por estrutura de covariância.
Fonte: Adaptada de Paula Alves (2016).

A geoestatística espaço-temporal via modelagem da estrutura de covariância é uma generalização da geoestatística puramente espacial, que propõe levar em consideração as interações existentes entre os componentes espaciais e os temporais de modo a permitir predições no tempo e no espaço. Essa metodologia utiliza-se exclusivamente de funções de covariâncias espaço-temporais que asseguram a condição de positiva definida.

Segundo Montero, Fernández-Avilés e Mateu (2015), construir modelos válidos de funções de covariâncias espaço-temporais é um desafio bastante difundido na modelagem geoestatística espaço-temporal.

Gneiting (2002) propôs classes bem gerais de funções de covariância espaço-temporais que envolve combinações de funções completamente monótonas e funções positivas com derivadas completamente monótonas.

Além das funções propostas por Gneiting (2002), na literatura podem ser encontradas várias abordagens de funções válidas de covariâncias separáveis e não separáveis como: Rouhani e Hall (1989), Dimitrakopoulos e Luo (1994), De Cesare, Myers e Posa (1997), Cressie e Majure (1997), Cressie e Huang (1999), Christakos (2000), Stein (2005), Porcu, Mateu e Bevilacqua (2007), Rodrigues e Diggle (2010), Fonseca e Steel (2011), De Iaco, Myers, Posa (2001; 2002; 2013), entre outros.

Nos trabalhos de Christakos (2000) e Cressie e Wikle (2011) podem ser encontradas contribuições ordenadas sobre análise espaço-temporal legítima em Geoestatística, bem como várias classes de funções de covariância espaço-temporal válidas.

Recentemente, com o avanço da área da inteligência artificial, técnicas computacionais capazes de adquirir conhecimento de forma automática, denominada aprendizagem de máquina, surgiram como alternativas competitivas frente aos métodos clássicos de estatística.

Entre os algoritmos de aprendizagem de máquina mais frequentes, pode-se citar Random Forest (BREIMAN, 2001; LIAW, WEINER, 2002) e Support Vector Machine (SMOLA, SCHÖLKOPF, 2004; VAPNIK, 2013), além do método tradicional de Regressão Linear Múltipla (GUJARATI, PORTER, 2011).

O aprendizado de máquina se beneficiou do crescente número de informações digitalizadas disponíveis via internet. Segundo James et al. (2013), ele explora o estudo e a construção de algoritmos para extrair regras e padrões, a partir de entradas amostrais, a fim de fazer previsões sobre os dados.

Para avaliar o desempenho dos modelos, normalmente utilizam-se funções matemáticas, referidas como métricas. O erro médio (Mean Error – ME), o erro quadrático médio (Mean Squared Error – MSE), a raiz do erro quadrático médio (Root Mean Square Error – RMSE), o erro médio absoluto (Mean absolute error – MAE) e coeficiente de determinação (R^2) são métricas frequentemente utilizadas em problemas com variável de resposta contínua (problemas de regressão).

Isaaks e Srivastava (1989) e Chai e Draxler (2014) apresentam uma explicação detalhada das métricas de erro descritas pelas Equações (1) a (4):

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4}$$

em que y_i é o valor observado na localização i, $\hat{y}_i$ o valor predito e n o tamanho da amostra.

As métricas de erro medem o desempenho em termos do desvio de suas previsões e dos valores de referência. Assim, resultados com erros mais baixos normalmente indicam que ele é melhor do que os outros que obtiveram erros maiores. Já o R^2, conforme abordado por Gujarati e Porter (2011), mede o quão bem o modelo se ajusta aos dados, ou seja, mede a proporção da variação total da variável dependente Y que é explicada, em termos lineares, pelo modelo de regressão. Pode ser obtido de acordo com a Equação (5):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} = 1 - \frac{MSE}{\hat{\sigma}_y^2} \tag{5}$$

em que $\bar{Y}$ é o valor médio de Y, $\hat{\sigma}_y^2$ a estimativa da variância total de Y, calculada com n como divisor (em vez de $n-1$) e y_i, $\hat{y}_i$, n conforme definidos acima. Ele varia de 0 a 1. Um $R^2 = 1$ significa um ajuste perfeito, ou seja, quando $\hat{y}_i = y_i$, para cada i. Por outro lado, um $R^2 = 0$, indica que as variações de Y são aleatórias de forma que o modelo não explica as variações Y.

Dessa forma, o objetivo deste trabalho é aplicar a metodologia geoestatística espaço-temporal de funções de covariâncias e/ou aprendizagem de máquina com a finalidade de inferir acerca da temperatura máxima do ar do estado de Minas Gerais, visando contribuir com desafios tais como, aquecimento global, urbanização descontrolada, escassez de recursos naturais, epidemias e catástrofes naturais.

Além desta seção introdutória, o texto deste trabalho está organizado em outras três seções: Capítulo 1, Capítulo 2 e Conclusões Gerais. Os capítulos seguem uma mesma organização e apresentam resultados provenientes de dois artigos científicos, resultantes do projeto de doutorado desenvolvido. Por fim, a seção denominada Considerações Finais, apresenta um breve resumo sobre o que foi abordado.

Os Capítulos podem ser lidos de forma independente e têm, em comum, registros de 61 estações meteorológicas convencionais de observações de superfícies que se encontram distribuídas no estado de Minas Gerais e regiões/estados circunvizinhos.

O Capítulo 1 trata-se do uso da geoestatística espaço-temporal na predição da temperatura máxima diária do ar do estado de MG, no período de 01 de janeiro 1996 a 31 de dezembro de 2016, utilizando a abordagem via modelagem da estrutura de covariância e a krigagem ordinária como método de interpolação. Já o Capítulo 2, trata-se de viabilizar o uso dos algoritmos de aprendizagem de máquina supervisionada (regressão linear múltipla, random forest, support vector machine) na modelagem da temperatura máxima do ar do estado de MG.

REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, 2001, p. 5-32.

DE CESARE, L.; MYERS, D. E.; POSA, D. Spatial-temporal modeling of SO_2 in Milan district. **Geostatistics wollongong'96**, Kluwer Academic, v. 2, p. 1031–1042, 1997.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. **Geoscientific model development.** v. 7, n. 3, p. 1247-1250, 2014.

CHRISTAKOS, G.; VYAS, V. M. A composite space/time approach to studying ozone distribution over eastern United States. **Atmospheric Environment**, v. 32; n:16, p. 2845-2857, 1998.

CHRISTAKOS, G. **Modern spatiotemporal Geostatistics**. New York: Oxford University Press, 2000. 288 p.

COELHO, A. M. Agricultura de precisão: manejo da variabilidade espacial e temporal dos solos e culturas. **Embrapa Milho e Sorgo-Documentos** (INFOTECA-E), 2005.

CRESSIE, N.; HUANG, H. Classes of nonseparable, spatio-temporal stationary covariance functions**. Journal of the American Statistical Association**, v. 94, n. 448, p. 1330–1339, 1999.

CRESSIE, N.; MAJURE, J. J. Spatio-temporal statistical modeling of livestock waste in streams. **Journal of Agricultural, Biological, and Environmental Statistics**, p. 24-47, 1997.

CRESSIE, N.; WIKLE, C. K. **Statistics for spatio-temporal data**. New Jersey: John Wiley & Sons, 2011. 531 p.

DE IACO, S.; MYERS, D. E.; POSA, D. Space-time analysis using a general product-sum model. **Statistics & Probability Letters**, v. 52, n. 1, p. 21-28, 2001.

DE IACO, S.; MYERS, D. E.; POSA, D. Nonseparable space-time covariance models: some parametric families. **Mathematical Geology**, vol. 34, n. 1, p. 23-42, 2002.

DE IACO, S.; POSA, D.; MYERS, D. E. Characteristics of some classes of space–time covariance functions. **Journal of Statistical Planning and Inference**, v. 143, n. 11, p. 2002-2015, 2013.

DIMITRAKOPOULOS, R.; LUO, X. Spatiotemporal modelling: covariances and ordinary kriging systems. In: **Geostatistics for the next century**. Springer, Dordrecht, p. 88-93, 1994.

FONSECA, T. C. O.; STEEL, M. F. J. A general class of nonseparable space-time covariance models. **Environmetric**s, v. 22, n. 2, p. 224–242, 2011.

GRÄLER, B.; PEBESMA, E.; HEUVELINK, G. Spatio-temporal interpolation using gstat. **RFID Journal**, v. 8, n. 1, p. 204-218, 2016.

GNEITING, T. Nonseparable, stationary covariance functions for space-time data. **Journal of the American Statistical Association**, v. 97, n. 458, p. 590–600, 2002.

GNEITING, T.; GENTON, M. G.; GUTTORP, P. Geostatistical space-time models, stationarity, separability, and full symmetry. **Monographs On Statistics and Applied Probability**, v. 107, p. 151, 2006.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica,** 5ª ed., Porto Alegre: AMGH Editora, 2011.

GOOVAERTS, P.; SONNET, Ph. Study of spatial and temporal variations of hydrogeochemical variables using factorial kriging analysis. In: **Geostatistics Tróia'92**. Springer, Dordrecht, p. 745-756, 1993.

ISAAKS, E. H.; SRIVASTAVA, R. M. **Applied Geostatistics**. New York: Oxford University Press, 1989. 561 p.

JAMES, G. et al. **An introduction to statistical learning:** with applications in R (Springer texts in Statistics), 2013. 426p.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, v. 52, n. 6, p. 119-139, 1951.

LANDIM, P. M. B. Sobre geoestatística e mapas. **Terrae Didatica**, v. 2, n. 1, p. 19-33, 2006.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.

MANTOVANI, E. C. et al. Management crop production system using precision farming concept for decision making./gerenciamento do sistema de produção utilizando conceitos de agricultura de precisão para tomada de decisão. **Revista Brasileira de Engenharia de Biossistemas**, v. 1, n. 2, p. 127-136, 2007.

MATHERON, G. Traité de géostatistique appliquée, Tome I: Mémoires du bureau de recherches géologiques et minières. **Editions Technip**, Paris, v. 14, 1962.

MATHERON, G. Principles of goestatistics. **Economic Geology**, Lancaster, v. 58, n.8, p. 1246-1266, 1963.

MATHERON, G. **Les variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature**. Masson et CIE, 1965.

MATHERON, G. **The theory of regionalized variables and its applications**. Les Cahiers du Centre de Morphologie Mathemathique, Fas. 5, C. G. Fontainebleau. 1971.

MONTERO, J. M.; FERNÁNDEZ-AVILÉS, G.; MATEU, J. **Spatial and spatio-temporal geostatistical modeling and kriging**. Chennai, India: John Wiley e Sons, 2015.

MOREIRA, D. S. et al. Modeling the radiative effects of biomass burning aerosols on carbon fluxes in the Amazon region. **Atmospheric Chemistry and Physics**, v. 17, n. 23, p.14785-14810, 2017. Disponível em: <https://doi.org/10.5194/acp-17-14785-2017>. Acesso em: 12 de dez 2017.

PAEZ, M. S.; GAMERMAN, D. Modelagem de processos espaço-temporais. **Escola de Séries Temporais e Econometria**, v. 11, 2005.

PORCU, E.; MATEU, J.; BEVILACQUA, M. Covariance functions that are stationary or nonstationary in space and stationary in time. **Statistica Neerlandica**, v. 61, n. 3, p. 358–382, 2007.

RAJA, N. B. et al. Space-time kriging of precipitation variability in Turkey for the period 1976–2010. **Theoretical and Applied Climatology**, v. 129, n. 1-2, p. 293-304, 2017.
RODRIGUES, A.; DIGGLE, P. J. A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. **Scandinavian Journal of Statistics**, v. 37, n. 4, p. 553–567, 2010.

ROUHANI, S.; HALL, T. J. Space-time kriging of groundwater data. In: Armstrong M. (eds) **Geostatistics**. Springer, Dordrecht, p. 639–650, 1989.

ROUHANI, S.; WACKERNAGEL, H. Multivariate geostatistical approach to space-time data analysis. **Water Resources Research**, v. 26, n. 4, p. 585-591, 1990.

SCHABENBERGER, O.; GOTWAY, C. A. **Statistical methods for spatial data analysis**. Chapman e Hall/CRC, 2005.

PAULA ALVES, H. J. P. **Modelo geoestatístico espaço-temporal com funções de covariância estacionárias não-separáveis aplicado ao albedo de superfície**. 2016. 53 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG, 2016. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/10966>. Acesso em: 02 jul. 2017.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199-222, 2004.

SNEPVANGERS, J. J. J. C.; HEUVELINK, G. B. M.; HUISMAN, J. A. Soil water content interpolation using spatio-temporal kriging with external drift. **Geoderma**, v. 112, n. 3-4, p. 253-271, 2003.

STEIN, M. L. Space time-covariance functions. **Journal of the american statistical association**, v. 100, n. 469, p. 310–321, 2005.

VAPNIK, V. N. **The nature of statistical learning theory**. Springer science & business media, New York, 2nd ed, 2013.

VAROUCHAKIS, E. A. Spatiotemporal geostatistical modelling of groundwater level variations at basin scale: a case study at Crete's Mires Basin. **Hydrology Research**, v. 49, n.4, p. 1131-1142, 2018.

CAPÍTULO 1 – O USO DA GEOESTATÍSTICA ESPAÇO-TEMPORAL NA PREDIÇÃO DA TEMPERATURA MÁXIMA DO AR

RESUMO

Processos estocásticos de natureza espaço-temporais consistem de fenômenos que são caracterizados por meio da variabilidade espacial e temporal. Atualmente, é uma das áreas de maior crescimento com diversas aplicações em ciências ambientais, geográficas, biológicas, epidemiológicas, entre outras. Certamente, os métodos da estatística convencional não são adequados para modelar estruturas autocorrelacionadas no espaço e no tempo. De fato, ainda há grandes desafios no que tange à implementação computacional da metodologia geoestatística para análise de processos espaço-temporais, com destaque para o pacote spacetime do programa R, utilizado neste estudo. Assim, este trabalho tem como objetivo aplicar a metodologia geoestatística espaço-temporal de funções de covariância a fim de inferir acerca da temperatura máxima do ar do Estado de Minas Gerais de 1996 a 2016, visando contribuir com desafios, tais como aquecimento global, urbanização descontrolada, escassez de recursos naturais, epidemias e catástrofes naturais. Utilizando dados de 61 estações meteorológicas foi realizada a análise geoestatística espaço-temporal, no qual o modelo de covariância soma-métrico foi o mais adequado, considerando-se o critério do erro quadrático médio. Dessa forma, foi possível elaborar mapas de predições das temperturas máximas do ar no estado de Minas Gerais por meio da krigagem ordinária, assumindo-se estacionariedade de primeira ordem do processo estocástico avaliado. Pode-se observar que os modelos da geoestatística espaço-temporal mostraram ser eficientes nos estudos espaço-temporais das temperaturas máximas do ar.

Palavras-chave: Modelagem de Dados Espaço-Temporal, Covariância, Variograma, Krigagem Ordinária.

ABSTRACT

Stochastic processes of spatio-temporal nature consist of phenomenons that are characterized by spatial and temporal variability. Currently, it is one of the great growing areas with diverse applications in environmental, geographic, biological, epidemiological sciences, among others. Certainly, conventional statistical methods are not adequate to modeling self-correlated structures in space and time. In fact, there are still major challenges regarding the computational implementation of the geostatistical methodology for the analysis of space-time processes, with emphasis on the spacetime package of the R program used in this study. Thus, this work aims to apply the geostatistical methodology of covariance functions in order to infer about the maximum air temperature of the State of Minas Gerais from 1996 to 2016, aiming to contribute with challenges such as heating uncontrolled urbanization, scarcity of natural resources, epidemics and natural disasters. Using the data from 61 meteorological stations, the geostatistical space-time analysis was performed, in which the sum-metric covariance model was the most adequate, considering the criterion of the Mean Squared Error. Thus, it was possible to prepare maps of predictions of maximum air temperatures in the state of Minas Gerais through of ordinary kriging, assuming first order stationarity of the evaluated stochastic process. It can be observed that the models of space-time geostatistics have shown to be efficient in the space-time studies of maximum air temperatures.

Keywords: Spatial-temporal Data Modeling. Covariance. Variogram. Ordinary Kriging.

1 INTRODUÇÃO

Dados espaço-temporais são observações tomadas em diferentes localizações no espaço e, para cada localização, em diferentes tempos. Existem três formas distintas de análise de processos espaço-temporais: análise espacial para cada tempo, análise temporal para cada ponto no espaço e análise espacial e temporal conjunta. As duas primeiras possibilidades isolam a parte espacial ou a parte temporal e aplicam-se as técnicas padrões para o tipo de processo resultante. A terceira possibilidade considera as variações espaciais e temporais conjuntamente.

Nos últimos anos essa área teve um grande crescimento científico, acompanhado de uma diversificação das temáticas abordadas, com diversas aplicações práticas em uma ampla variedade de ramos da ciência, tais como, climatologia, meteorologia, agricultura ou qualquer outra área preocupada com o estudo de fenômenos que ocorrem tanto no espaço quanto no tempo. Na literatura pode-se citar trabalhos recentes para o estudo e análise de dados espaço-temporais tais como, Reboita et al. (2015), Raja et al. (2017), Moreira et al. (2017) e Varouchakis (2018).

Em Climatologia, segundo o quinto relatório do Intergovernmental Panel on Climate Change (IPCC, 2014) uma elevação na temperatura média do planeta poderá causar consequências devastadoras para o nosso planeta Terra e consequentemente, para os seres vivos, incluindo animais e vegetais. Entre essas consequências, pode-se citar: derretimento de grande parte das calotas polares; elevação do nível médio do mar; submersão de ilhas oceânicas e ameaça de desaparecimento de cidades litorâneas que atualmente estão em níveis próximos ao nível médio do mar; aumento de eventos extremos, tais como tornados, furacões, tempestades severas e secas prolongadas, incluindo desertificação de algumas regiões. É de consenso da grande maioria dos pesquisadores que o efeito estufa é o principal fator que tem contribuído para o aquecimento global que já vem sendo observado nas últimas décadas.

O estudo da variável temperatura do ar é de suma importância para a comunidade científica poder quantificar as emissões dos gases de efeito estufa causado principalmente por atividades antrôpicas.

A temperatura do ar atmosférico é uma variável meteorológica que possui grande dependência temporal devido principalmente aos movimentos de rotação e translação da terra,

fazendo com que a energia recebida pela Terra varie durante o dia e o ano, respectivamente. A temperatura também possui variações espaciais, dependendo, por exemplo, dos movimentos das massas de ar e de variações na superfície, como cobertura do solo, albedo, altitude e umidade.

Devido ao grande custo de instalação e manutenção de estações meteorológicas, elas são bem escassas, principalmente em regiões de difícil acesso como oceanos e florestas. Normalmente, são instaladas em locais mais acessíveis, mas não equidistantes, principalmente as convencionais que necessitam de observadores para coletar os dados regularmente. Portanto, na maioria das vezes é necessário utilizar métodos matemáticos para estimar os valores de temperatura nos pontos de uma grade regular. Também é comum observar falhas na série de dados, devido à falta da observação ou a problemas no instrumento ou na transmissão e armazenamento dos dados. Desta forma, em muitas situações é comum o emprego de métodos para deixar a série sem falhas tanto espacial quanto temporal.

Na literatura são encontrados diversos estudos com abordagens distintas para estimar as temperaturas do ar, em diferentes estados e regiões brasileiras (GOMES et al., 2014; GARCIA, ANDRE, 2015; MEDEIROS et al., 2015, entre outros). Esses estudos, embora trabalhem com dados espaço-temporais, não exploram a interação entre os componentes espaciais e temporais conjuntamente.

A geoestatística espaço-temporal, via modelagem da estrutura de covariância, é um dos procedimentos para análise de dados espaço-temporais que leva em consideração as interações existentes entre os componentes espaciais e temporais e permite interpolações no tempo e no espaço.

Assim, o objetivo deste trabalho é aplicar a metodologia geoestatística espaço-temporal de funções de covariâncias com a finalidade de inferir acerca da temperatura máxima do ar do estado de Minas Gerais, visando contribuir com desafios tais como, aquecimento global, urbanização descontrolada, escassez de recursos naturais, epidemias e catástrofes naturais.

2 MATERIAL E MÉTODOS

Nas seções 2.1 e 2.2 caracteriza a área de estudo e descrição dos dados, enquanto que na seção 2.3 serão apresentados conceitos espaço-temporais relevantes para a compreensão dos resultados.

2.1 Área de Estudo

A região de estudo é o Estado de Minas Gerais (MG), localizado na região sudeste do Brasil, com área de 586.520,732 km^2. A região de trabalho é delimitada pelas latitudes 14º13´58´´ S e 22º54´00´´ S e longitude 39º51´32´´ W e 51º02´35´´ W.

O estado de MG é subdividido em 853 municípios com distância linear entre os pontos extremos de 986 km no sentido norte-sul e, de 1248 km, no leste-oeste (MG.GOV.BR, 2017). Abrange os fusos 22, 23 e 24. Adotou-se o fuso 23 como principal e as coordenadas do fuso 22 e 24 foram projetadas para o fuso 23, permitindo juntar todas as estações utilizadas em um plano único. Assim, as coordenadas geográficas, em graus decimais, foram convertidas para metros usando a projeção de coordenadas UTM (Universal Transversa de Mercator), Zona 23 Sul, Datum WGS84.

2.2 Descrição dos Dados

O conjunto de dados foi obtido do Banco de Dados Meteorológicos para Ensino e Pesquisa do Instituto Nacional de Meteorologia (INMET, 2018). Foram registros de 61 estações meteorológicas convencionais de observações de superfícies que se encontram distribuídas nas regiões ou estados circunvizinhos a MG. Nota-se, na Figura 1, que a distribuição espacial dessas estações é bastante irregular.

A Figura 1, foi gerada pelo aplicativo Grid Analysis and Display System (GrADS: DOTY, KINTER, 1993) e a topografia, com resolução de 5 km, foi proveniente do United

States Geological Surveys (USGS) vinculado ao Earth Resources Observation Systems (EROS) (GESCH, VERDIN, GREENLEE, 1999).

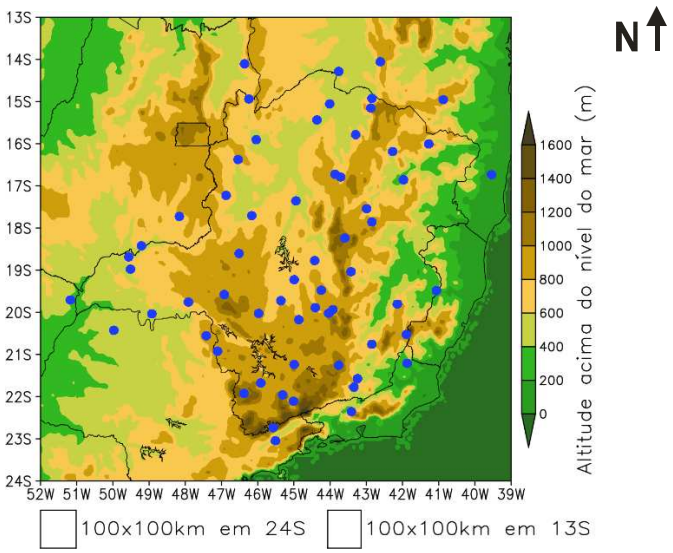


Figura 1 – Localizações das 61 estações meteorológicas convencionais utilizadas neste estudo e que estão instaladas em Minas Gerais e estados circunvizinhos (pontos azuis). As cores representam a topografia da região.

A variável em estudo é a temperatura máxima diária do ar e cada estação está associada a uma série temporal de 7671 tempos distintos representando os dias entre janeiro de 1996 a dezembro de 2016. No entanto, existem algumas falhas ao longo da série.

Para execução das análises descritivas e geoestatística foi utilizado o software livre R versão 3.5.0 (TEAM, 2018) com os pacotes gstat (PEBESMA, GRÄLER, 2018), sp (PEBESMA, BIVAND, 2005), spacetime (PEBESMA, 2018), geoR (RIBEIRO JR, DIGGLE, 2016), lattice (SARKAR, 2017), xts (RYAN, ULRICH, 2018), maptools (BIVAND, LEWIN-KOH, 2017) e rgdal (BIVAND, KEITT, ROWLINGSON, 2018) que são essenciais para se trabalhar com a geoestatística espaço-temporal.

A

Tabela 1 apresenta o formato padrão de uma base de dados espaço-temporais. São devidamente identificadas as coordenadas e municípios das estações meteorológicas, como também a data das medições das temperaturas máxima do ar. Os valores figurados com NA são valores perdidos por falhas de medição e foram tratados, segundo a função na.locf do pacote spacetime do R.

Tabela 1 – Formato Padrão de uma Base de dados Espaço-Temporal.

Data	**Mun.**[1]	**ID**	**Lat**	**Lon**	**T_M**
01-01-1996	Belo Horizonte	83587	–19,9	–43,9	24,4
01-01-1996	Viçosa	83642	–20,7	–42,8	24,6
…	…	…	…	…	…
01-01-1996	Juiz de Fora	83692	–21,7	–43,3	23,8
…	…	…	…	…	…
31-12-2016	Juiz de Fora	83692	–21,7	–43,3	NA

[1] Mun.: Nome do município de localização da instalação, ID: Código da estação meteorológica convencional, Lat: Latitude (°), Lon: Longitude (°) e T_M: Temperatura máxima do ar (°C). Fonte: INMET (2018).

Para plotagem dos mapas do estado de MG foi utilizado o arquivo vetorial obtido no portal de mapas do Instituto Brasileiro de Geografia e Estatística (IBGE, 2016).

A modelagem geoestatística espaço-temporal via funções de covariâncias espaço-temporais foi utilizada para obter as predições da temperatura máxima diária do ar. O método de interpolação escolhido neste estudo foi a krigagem ordinária por ser considerado, segundo Perin et al. (2015), o mais frequente dos métodos geoestatísticos aplicado ao estudo da temperatura máxima do ar.

As equações para krigagem no domínio espaço-temporal são exatamente as mesmas que as equações de krigagem padrão. Detalhes das equações de krigagem ordinária podem ser encontradas em Varouchakis (2018).

Segundo Rios (2018), a estratégia de coletar pontos amostrais fora da área de interesse é eficaz na minimização de efeito de borda. Assim, das 61 estações utilizadas, 48 estão localizadas nos municípios do estado de MG e 13 nos estados circunvizinhos a MG para que os mapas de predição da temperatura máxima do ar do estado de MG tenham valores interpolados também nos limites do estado.

Os mapas dos valores observados da temperatura máxima diária do ar medida nas estações meteorológicas do estado de MG, apresentados na Figura 2, têm como objetivo fornecer uma ideia inicial da distribuição dos valores da temperatura. Opta-se, por simples critério, a apresentação desses mapas apenas para o dia 15 de janeiro, para os anos de estudo.

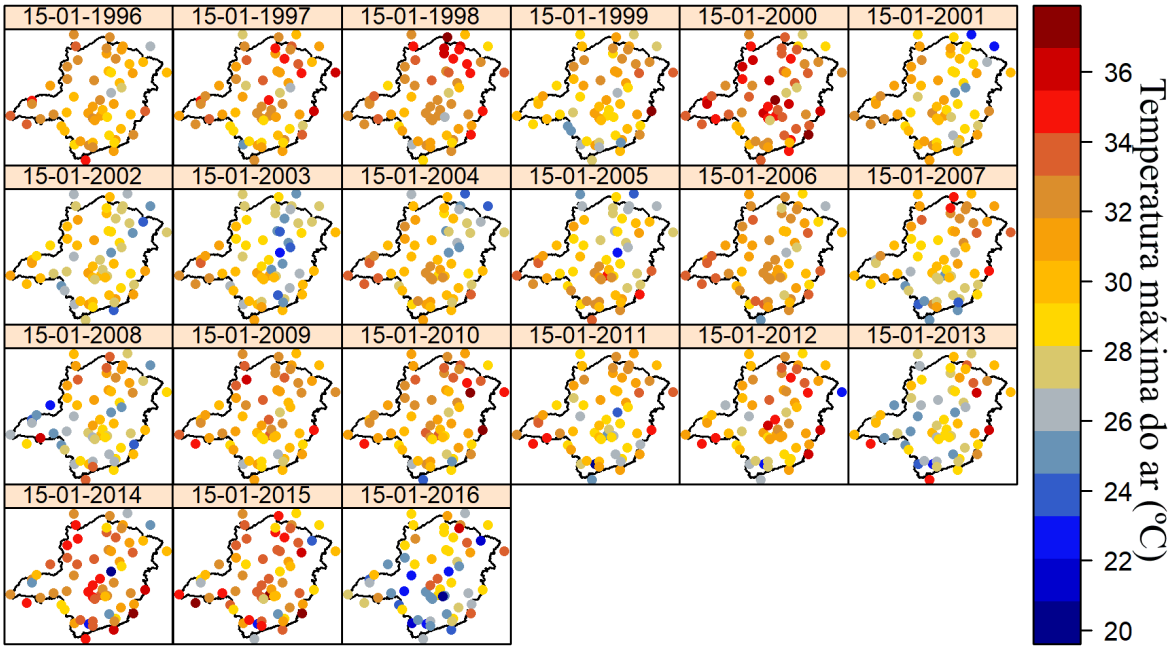


Figura 2 – Localização das estações meteorológicas estudadas no presente trabalho e valores observados da temperatura máxima do ar registrada no dia 15 de janeiro (1996 a 2016).

Para traçar isolinhas que possibilitem visualizar regiões de máximos, mínimos e gradientes, é necessário que os dados estejam em grade regular. Dessa forma, para obter um campo regular, a partir dos dados observados, é necessário utilizar métodos matemáticos que estimem os valores nos pontos de grade. Esses métodos de interpolação também têm como objetivo preencher falhas nos dados, tanto espacial quanto temporal, que normalmente existem devido à problemas no instrumento de medida, na sua coleta, na transmissão ou no armazenamento.

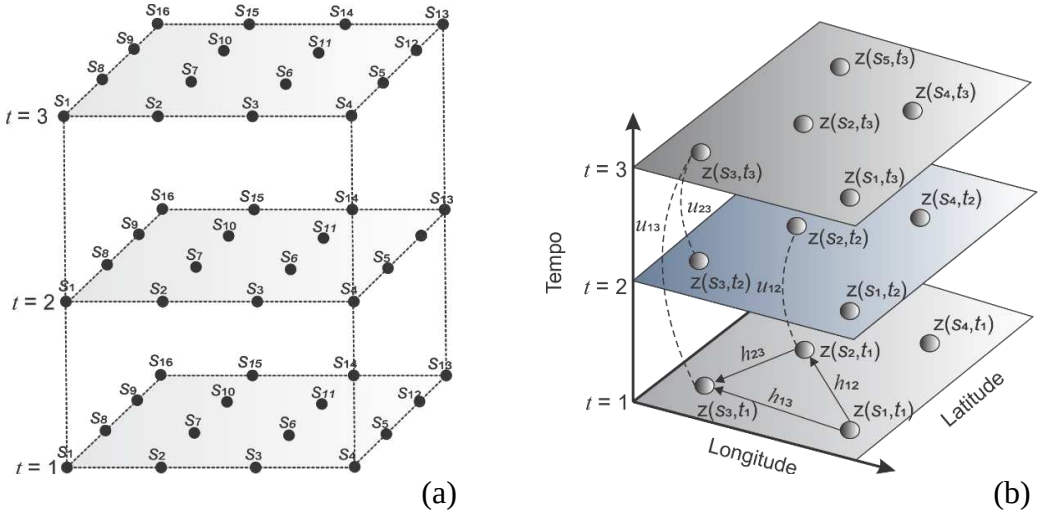
Entre as classes de funções de covariância espaço-temporais disponíveis na literatura, destaca-se, segundo De Iaco, Posa e Myers (2013), as classes de funções de Gneiting (2002) e segundo Gräler, Pebesma e Heuvelink (2016), as classes de funções de covariâncias separáveis (modelo soma e modelo produto) e não-separáveis (modelos soma-produto, métrico, soma-métrico e soma-métrico simplificado).

2.3 Conceitos espaço-temporais relevantes

Denotam-se $\mathbb{R}^d$ o espaço euclidiano d-dimensional, $D \subset \mathbb{R}^d$ o domínio espacial e $T \subset \mathbb{R}$ o domínio temporal. Para esse trabalho utilizou-se o domínio espacial $D \subset \mathbb{R}^2$.

Para estender a geoestatística para o caso espaço-temporal, denomina-se coordenadas espaço-temporal em $D \times T$ um par de elementos $(\boldsymbol{s}_i, t_i)$, com $\boldsymbol{s}_i \in D$ e $t_i \in T$ e campo aleatório (ou função aleatória) espaço-temporal em $D \times T$, o conjunto $\{ Z(\boldsymbol{s}, t) : s \in D, t \in T \}$, em que a variável aleatória $Z(\boldsymbol{s}, t)$ é o atributo de interesse. Já a relação entre duas observações $z(\boldsymbol{s}_i, t_i)$ e $z(\boldsymbol{s}_j, t_j)$ normalmente dependerá de $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$, denominado vetor distância de separação espacial (lag espacial) e de $u = t_i - t_j$, denominado distância temporal (lag temporal). Ao par $(\boldsymbol{h}, u)$, refere-se um lag espaço-temporal (CRESSIE e HUANG, 1999).

Para compreensão dos conceitos, a Figura 3a ilustra uma grade espacialmente regular em $\mathbb{R}^2 \times \mathbb{R}$ com 16 localizações espaciais em 3 instantes de tempo sucessivos, ou seja, 48 coordenadas espaço-temporais. Já a Figura 3b ilustra as realizações $z(\boldsymbol{s}_i, t_j)$, com $i = 1, \ldots, 5$, $j = 1, 2, 3$, os lags espaciais $\boldsymbol{h}_{12} = \boldsymbol{s}_1 - \boldsymbol{s}_2$, $\boldsymbol{h}_{13} = \boldsymbol{s}_1 - \boldsymbol{s}_3$, $\boldsymbol{h}_{23} = \boldsymbol{s}_2 - \boldsymbol{s}_3$ e os lags temporais $u_{12} = t_1 - t_2$, $u_{13} = t_1 - t_3$ e $u_{23} = t_2 - t_3$.



(a) (b)

Figura 3 – (a) Grade espacialmente regular em $\mathbb{R}^2 \times \mathbb{R}$ composta por 48 coordenadas espaço-temporais. (b) Realizações espaço-temporais com representação das distâncias espaciais e temporais.
Fonte: Adaptada de Montero, Fernández-Avilés e Mateu (2015).

De acordo com Sherman (2011), dado um campo espaço-temporal $\{ Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D, t \in T \}$ e quaisquer coordenadas espaço-temporais $(\boldsymbol{s}_i, t_i)$, $(\boldsymbol{s}_j, t_j)$ em $D \times T$, o valor esperado, a variância (a priori), a covariância e o variograma são descritos, respectivamente, pelas Equações (1) a (4):

$$E[Z(\boldsymbol{s}_i, t_i)] = \mu(\boldsymbol{s}_i, t_i) \tag{1}$$

$$Var[Z(\boldsymbol{s}_i, t_i)] = E\left[\left(Z(\boldsymbol{s}_i, t_i) - \mu(\boldsymbol{s}_i, t_i)\right)^2\right] \tag{2}$$

$$Cov\left[Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)\right] = E\left[\left(Z(\boldsymbol{s}_i, t_i) - \mu(\boldsymbol{s}_i, t_i)\right)\left(Z(\boldsymbol{s}_j, t_j) - \mu(\boldsymbol{s}_j, t_j)\right)\right] \tag{3}$$

$$2\gamma_{st}((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)) = Var[Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)]\,. \tag{4}$$

Utilizando propriedades da variância, tem-se que se as funções de covariância e variograma existem e estão relacionadas de acordo coma a Equação (5):

$$2\gamma_{st}((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)) = Var\,[Z(\boldsymbol{s}_i, t_i)] + Var\,[Z(\boldsymbol{s}_j, t_j)] - 2Cov\left[Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)\right] \tag{5}$$

O uso de variogramas tem sido particularmente utilizado em problemas puramente espaciais. Segundo Gneiting e Raftery (2007), ao discutirem os modelos geoestatísticos espaço-temporais os autores, em sua grande maioria, trabalham com estruturas de funções de covariâncias e correlações. No entanto, sempre que é possível definir ambas as funções, é possível mudar de uma para outra utilizando a Equação (5).

Na prática, um modelo parte de um conjunto de n variáveis aleatórias,

$$\{Z(\boldsymbol{s}_1, t_1),\ Z(\boldsymbol{s}_2, t_2),\ \ldots,\ Z(\boldsymbol{s}_n, t_n)\}, \tag{6}$$

correlacionadas entre si, da qual só se conhece uma realização da função aleatória espaço-temporal $Z(\boldsymbol{s}, t)$. Teoricamente, com uma única realização, é impossível estimar qualquer parâmetro estatístico. Segundo Kyriakidis e Journel (1999), para permitir a inferência de algumas estatísticas, deve-se assumir graus de estacionariedade sobre o campo aleatório espaço-temporal $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D,\ t \in T\}$.

Segundo Montero, Fernández-Avilés e Mateu (2015), um campo $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D,\ t \in T\}$ é estacionário de primeira ordem quando a média é constante ao longo de $D \times T$, isto é,

$$E\,[Z(\boldsymbol{s}_i, t_i)] = E\left[Z(\boldsymbol{s}_j, t_j)\right] = \mu,\ \ \forall(\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j) \in D \times T\,. \tag{7}$$

Nesse caso, a Equação (4) torna-se equivalente a

$$\begin{aligned}\gamma_{st}((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)) &= \frac{1}{2}Var\left(Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)\right) \\ &= \frac{1}{2}E\left[\left(Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)\right)^2\right] = \gamma_{st}(\boldsymbol{h}, u)\end{aligned} \tag{8}$$

em que $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ é o lag espacial e $u = t_i - t_j$ o lag temporal.

Segundo Kyriakidis e Journel (1999), um campo $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D,\ t \in T\}$ é estacionário de segunda ordem quando: (i) a média $\mu = E[Z(\boldsymbol{s}, t)]$ é constante em toda área, ou seja, o campo é estacionário de primeira ordem; (ii) para quaisquer par de coordenadas espaço-temporais $(\boldsymbol{s}_i, t_i)$, $(\boldsymbol{s}_j, t_j)$ em $D \times T$, a função de covariância $Cov\left[Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)\right]$

existe e depende apenas do lag espaço-temporal $(\boldsymbol{h}, u)$, em que $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ é o lag espacial e $u = t_i - t_j$ o lag temporal, ou seja,

$$\begin{aligned} Cov\big[Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)\big] &= E\left[Z(\boldsymbol{s}_i, t_i) Z(\boldsymbol{s}_j, t_j)\right] - \left[E[Z(\boldsymbol{s}_i, t_i)]\right]^2 \\ &= E\left[Z(\boldsymbol{s}_i, t_i) Z(\boldsymbol{s}_j, t_j)\right] - \mu^2 = C_{st}(\boldsymbol{h}, u) \end{aligned} \tag{9}$$

Observe que a estacionariedade da função de covariância implica a estacionariedade da variância e do variograma, já que

$$\begin{aligned} Var\,[Z(\boldsymbol{s}_i, t_i)] &= \mathrm{E}\left[\left(Z(\boldsymbol{s}_i, t_i)\right)^2\right] - \left[E[Z(\boldsymbol{s}_i, t_i)]\right]^2 = \mathrm{E}\left[\left(Z(\boldsymbol{s}_i, t_i)\right)^2\right] - \mu^2 \\ &= Cov(\mathbf{0}, 0) = C_{st}(\mathbf{0}, 0) > 0 \end{aligned} \tag{10}$$

e, pela Equação (8),

$$\begin{aligned} 2\gamma_{st}(\boldsymbol{h}, u) &= 2\gamma_{st}((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)) = Var[Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)] \\ &= Var\,[Z(\boldsymbol{s}_i, t_i)] + Var\left[Z(\boldsymbol{s}_j, t_j)\right] - 2Cov\left[Z(\boldsymbol{s}_i, t_i), Z(\boldsymbol{s}_j, t_j)\right] \\ &= 2C_{st}(\mathbf{0}, 0) - 2C_{st}\big(\boldsymbol{s}_i - \boldsymbol{s}_j, t_i - t_j\big) = 2C_{st}(\mathbf{0}, 0) - 2C_{st}(\boldsymbol{h}, u) \end{aligned} \tag{11}$$

A Equação (10) diz que, em um campo aleatório espaço-temporal estacionário, a variância a priori é finita e vale $C_{st}(\mathbf{0}, 0)$. Já a Equação (11), justifica o fato que a existência da covariância é uma hipótese mais restritiva do que a existência do variograma e diz também que, sob hipótese de estacionariedade, o variograma é a diferença entre a variância a priori e a função de covariância estacionária.

A pressuposição de estacionariedade da hipótese intrínseca (ou hipótese de estacionariedade fraca) é suficiente para a inferência de algumas estatísticas obtidas de processos de medição de fenômenos físicos que possui um variograma definido, porém, de elevada capacidade de dispersão (não têm uma variância a priori finita).

De acordo com Cressie e Huang (1999), um campo $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D,\ t \in T\}$ é intrinsecamente estacionário quando, para quaisquer duas coordenadas espaço-temporais $(\boldsymbol{s}_i, t_i)$, $(\boldsymbol{s}_j, t_j)$ em $D \times T$, com lag espacial $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ e lag temporal $u = t_i - t_j$, a variável aleatória $Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)$ é estacionária de segunda ordem, ou seja,

$$E\big[Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)\big] = 0 \tag{12}$$

e

$$Var\big[Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)\big] = E\left\{\left[Z(\boldsymbol{s}_i, t_i) - Z(\boldsymbol{s}_j, t_j)\right]^2\right\} = 2\gamma_{st}(\boldsymbol{h}, u)\,. \tag{13}$$

A Equação (12) expressa a similaridade dos valores e a Equação (13) a existência do variograma espaço-temporal em função apenas do lag espaço-temporal $(\boldsymbol{h}, u)$. Isso significa que, para quaisquer $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ e $u = t_i - t_j$ fixado, as variáveis regionalizadas: $Z(\boldsymbol{s}_1, t_1)$,

$Z(\boldsymbol{s}_2, t_2), \ldots, Z(\boldsymbol{s}_n, t_n)$ e $Z(\boldsymbol{s}_1 + \boldsymbol{h}, t_1 + u),\ Z(\boldsymbol{s}_2 + \boldsymbol{h}, t_2 + u), \ldots, Z(\boldsymbol{s}_n + \boldsymbol{h}, t_n + u)$, que são realizações diferentes do mesmo processo aleatório $Z(\boldsymbol{s}, t)$, têm a mesma função de distribuição multivariada. Então, ao admitir estacionariedade de primeira ordem, o parâmetro $\mu = E[Z(\boldsymbol{s}_1, t_1)] = E[Z(\boldsymbol{s}_2, t_2)] = \cdots = E[Z(\boldsymbol{s}_n, t_n)]$ pode ser estimado pela média aritmética dos valores das realizações das variáveis aleatórias.

Assim, substituindo a repetição nas realizações da função aleatória (tendo apenas uma realização) por repetição no espaço, presume-se que os valores encontrados nas diferentes regiões do campo têm as mesmas características e podem ser consideradas como realizações diferentes do mesmo processo aleatório.

De forma análoga a metodologia espacial, diz-se que o conjunto $\{Z(\boldsymbol{s}, t): \boldsymbol{s} \in D,\ t \in T\}$ é um campo aleatório espaço-temporal Gaussiano se para qualquer conjunto de localização espaço-temporal $\{(\boldsymbol{s}_i, t_i) : i = 1, \ldots, n\}$, o vetor $\big(Z(\boldsymbol{s}_1, t_1), Z(\boldsymbol{s}_2, t_2), \ldots, Z(\boldsymbol{s}_n, t_n)\big)$ têm distribuição Gaussiana multivariada. Nesse caso, cada variável $Z(\boldsymbol{s_i}, t_i)$, $i = 1, \ldots, n$, têm distribuição normal e assim, o campo aleatório espaço-temporal é totalmente caracterizado pelo vetor de médias e matriz de covariância.

A distribuição espaço-temporal das observações é modelada como uma distribuição gaussiana. Esta distribuição será perfeitamente caracterizada pelos momentos de primeira e segunda ordem, ou seja, sua função de esperança e covariância espaço-temporal. O sucesso no processo de krigagem espaço-temporal depende dessa função de covariância ou da função variograma que fornecem informações sobre a estrutura da dependência espaço-temporal presente na realização observada.

Segundo Gneiting e Raftery (2007), a complexidade da análise espaço-temporal legítimo em Geoestatística está relacionada com a presença ou ausência de três características importantes: separabilidade, completamente simétrica e estacionariedade.

A hipótese de estacionariedade, sob o ponto de vista matemático, afirma que a covariância, a correlação e o variograma são funções invariantes por translações, já que

$$Cov[Z(\boldsymbol{s} + \boldsymbol{h}, t + u), Z(\boldsymbol{s}, t)] = C_{st}(\boldsymbol{h}, u) \tag{14}$$

$$Corr[Z(\boldsymbol{s} + \boldsymbol{h}, t + u), Z(\boldsymbol{s}, t)] = \frac{C_{st}(\boldsymbol{h}, u)}{C_{st}(\boldsymbol{0}, 0)} = 1 - \frac{\gamma_{st}(\boldsymbol{h}, u)}{C_{st}(\boldsymbol{0}, 0)} = \rho_{st}(\boldsymbol{h}, u) \tag{15}$$

$$\begin{aligned} \gamma_{st}((\boldsymbol{s} + \boldsymbol{h}, t + u)) &= \frac{1}{2} Var[Z(\boldsymbol{s} + \boldsymbol{h}, t + u) - Z(\boldsymbol{s}, t)] \\ &= \frac{1}{2} E\left[\big(Z(\boldsymbol{s} + \boldsymbol{h}, t + u) - Z(\boldsymbol{s}, t)\big)^2\right] = \gamma_{st}(\boldsymbol{h}, u) \end{aligned} \tag{16}$$

e, nesse caso, elas são referidas como funções espaço-temporais estacionárias.

De acordo com Montero, Fernández-Avilés e Mateu (2015), um campo estacionário espaço-temporal $\{Z(\boldsymbol{s},t):\ \boldsymbol{s}\in D,\ t\in T\}$ é dito ter função de covariância espaço-temporal

(i) separável, se existem funções de covariância puramente espacial $C_s(\boldsymbol{h})$ e puramente temporal $C_t(u)$ tal que

$$C_{st}(\boldsymbol{h},\mathrm{u}) = C_s(\boldsymbol{h})\ C_t(\mathrm{u}),\ \ \forall(\boldsymbol{h},u) \tag{17}$$

ou seja, a função de covariância espaço-temporal representa-se como produto de funções de covariâncias puramente espacial e puramente temporal. Caso contrário, é dita ser não-separável;

(ii) isotrópica quando

$$C_{st}(\boldsymbol{h},\mathrm{u}) = C_{st}(\|\boldsymbol{h}\|,|u|),\ \ \forall(\boldsymbol{h},u) \tag{18}$$

(iii) par quando

$$C_{st}(\boldsymbol{h},u) = C_{st}(-\boldsymbol{h},-u),\ \ \forall(\boldsymbol{h},u) \tag{19}$$

(iv) completamente simétrica quando

$$C_{st}(\boldsymbol{h},u) = C_{st}(\boldsymbol{h},-u) = C_{st}(-\boldsymbol{h},u) = C_{st}(-\boldsymbol{h},-u),\ \ \forall(\boldsymbol{h},u). \tag{20}$$

Note que, uma função de covariância espaço-temporal estacionária e isotrópica é também completamente simétrica.

Para visualizar a condição de simetria, considere o domínio espacial $D = \{s_1, s_2, s_3, s_4\} \subset \mathbb{R}$, o domínio temporal $T = \{1, 2, 3, 4\}$ e as cinco coordenadas espaço-temporais dadas pela Equação (21).

$$(s_2, t_3),\ (\overbrace{s_2 - h}^{s_1}, \overbrace{t_3 - u}^{t_2}),\ \ (\overbrace{s_2 - h}^{s_1}, \overbrace{t_3 + u}^{t_4}),\ (\overbrace{s_2 + h}^{s_3}, \overbrace{t_3 + u}^{t_4})\ \ e\ \ (\overbrace{s_2 + h}^{s_3}, \overbrace{t_3 - u}^{t_2}). \tag{21}$$

Assume-se também, que o campo aleatório $\{Z(s,t):\ s\in\{s_1,s_2,s_3,s_4\},\ t\in\{1,2,3,4\}\}$ é estacionário e completamente simétrico. Logo, a covariância entre os pares de variáveis, $\big(Z(s_2,t_3), Z(s_1,t_2)\big)$, $\big(Z(s_2,t_3),(s_1,t_4)\big)$, $\big(Z(s_2,t_3), Z(s_3,t_4)\big)$, $\big(Z(s_2,t_3), Z(s_3,t_2)\big)$ é a mesma (Figura 4).

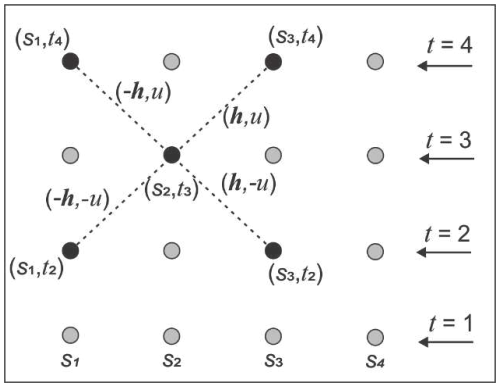


Figura 4 – Análise da covariância sob as hipóteses de estacionariedade e completamente simétrica.
Fonte: Adaptada de Montero, Fernández-Avilés e Mateu (2015).

Ainda de acordo com Montero, Fernández-Avilés e Mateu (2015), uma função C, definida em $D \times T \subset \mathbb{R}^d \times \mathbb{R}$, é uma função de covariância espaço-temporal válida se, e somente se, é simétrica e positiva-definida, ou seja, dado um conjunto qualquer de coordenadas espaço-temporais, $\{(s_1, t_1), (s_2, t_2), \dots, (s_m, t_m)\} \subset D \times T$ e $a_i, a_j \in \mathbb{R}$, $i = 1, \dots, m$ e $j = 1, \dots, m$, tem-se satisfeitas as Equações (20) e (23)

$$C(Z(s_i, t_i), Z(\boldsymbol{s}_j, t_j)) = C(Z(s_i, t_j), Z(\boldsymbol{s}_j, t_i)), \ \forall i, j = 1, \dots, m \tag{22}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \, C(Z(s_i, t_i), Z(\boldsymbol{s}_j, t_j)) \geq 0 \, . \tag{23}$$

E, sob hipótese de estacionariedade, tem-se que a Equação (24), juntamente com a Equação (20), tornam-se as condições necessárias e suficientes para uma função de covariância espaço-temporal ser válida.

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i \, a_j \, C(\boldsymbol{h}, u) \geq 0 \tag{24}$$

No entanto, é um desafio bastante difundido na modelagem geoestatística espaço-temporal construir funções de covariâncias espaço-temporais válidas. Alguns autores como Cressie e Huang (1999), Gneiting (2002), Stein (2005), De Iaco, Posa e Myers (2013) propõem famílias paramétricas de funções de covariância espaço-temporais válidas (separáveis e não-separáveis), que nos seus fatores aparecem, por exemplo, soma e/ou produto de funções de covariâncias puramente espacial e puramente temporal, já conhecidas anteriormente.

De acordo com Montero, Fernández-Avilés e Mateu (2015), para um campo espaço-temporal $\{Z(\boldsymbol{s},t) : \boldsymbol{s} \in D,\ t \in T\}$ intrinsecamente estacionário e n coordenadas espaço-temporais, $\{(\boldsymbol{s}_1,t_1),(\boldsymbol{s}_2,t_2),\dots,(\boldsymbol{s}_n,t_n)\} \subset D \times T$, um estimador de $2\gamma_{st}(\boldsymbol{h},u)$, generalização do estimador clássico proposto por Matheron (1989), é dado pela Equação (25)

$$2\hat{\gamma}_{st}(\boldsymbol{h},u) = \frac{1}{N(\boldsymbol{h},u)} \sum_{i=1}^{N(\boldsymbol{h},u)} \left[Z(\boldsymbol{s}_i,t_i) - Z(\boldsymbol{s}_j,t_j)\right]^2 \tag{25}$$

em que $Z(\boldsymbol{s}_i,t_i)$ e $Z(\boldsymbol{s}_j,t_j)$ são pares de observações da função aleatória espaço-temporal $Z(\boldsymbol{s},t)$ separadas por um lag espacial $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ e um lag temporal $u = t_i - t_j$; e $N(\boldsymbol{h},u)$ é o número de pares de observações separadas pelo lag espaço-temporal $(\boldsymbol{h},u)$.

Se o processo é também estacionário de segunda ordem, um estimador da covariância é dado pela Equação (26),

$$\hat{C}_{st}(\boldsymbol{h},u) = \frac{1}{N(\boldsymbol{h},u)} \sum_{i=1}^{N(\boldsymbol{h},u)} (Z(\boldsymbol{s}_i,t_i) - \bar{Z})(Z(\boldsymbol{s}_j,t_j) - \bar{Z}) \tag{26}$$

em que $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(\boldsymbol{s}_i,t_i)$ é um estimador da média do campo aleatório espaço-temporal e $N(\boldsymbol{h},u)$ definido como anteriormente.

A representação gráfica de $\hat{\gamma}(\boldsymbol{h},u)$, obtida da Equação (25) para cada lag $(\boldsymbol{h},u)$, permite gerar uma superfície suavizada denominada variograma experimental espaço-temporal. Já a representação gráfica $\hat{C}_{st}(\boldsymbol{h},u)$, obtida da Equação (25) para cada lag $(\boldsymbol{h},u)$, é referido como covariograma experimental. O variograma e o covariograma experimentais são instrumentos utilizados para quantificar a continuidade e estimar a variabilidade espacial de um determinado fenômeno em estudo.

Gneiting (2002) propõe uma família paramétrica de funções de covariância espaço-temporais válidas usando funções monótonas e funções em que a primeira derivada é completamente monótona. Ele demonstrou que, se $\varphi(x)$, $x \geq 0$, é uma função completamente monótona (significa que, $\forall n \in \mathbb{N}$, existem as derivadas $\varphi^{(n)}$ de ordem n e $(-1)^n\varphi^{(n)}(x) \geq 0$, $\forall x \geq 0$) e $\psi(x)$, $x \geq 0$, é uma função positiva com derivada completamente monótona, então

$$C(\boldsymbol{h},u) = \frac{\sigma^2}{\psi(|u|^2)^{d/2}}\ \varphi\left(\frac{\|\boldsymbol{h}\|^2}{\psi(|u|^2)}\right),\ \ \forall(\boldsymbol{h},u) \in D \times T \tag{27}$$

é uma família de funções de covariância espaço-temporais estacionária e completamente simétrica em $D \times T$, em que d é dimensão do espaço euclidiano d-dimensional.

Possíveis escolhas para as funções φ e ψ podem ser obtidas da Tabela 2 e da Tabela 3.

Tabela 2 – Funções completamente monótonas $\varphi(x)$, $x \geq 0$.

Forma funcional	Parâmetros
$\varphi(x) = \exp(-cx^{\gamma})$	$c > 0,\ 0 < \gamma \leq 1$
$\varphi(x) = (1 + cx^{\gamma})^{-\upsilon}$	$c > 0,\ 0 < \gamma \leq 1,\ \upsilon > 0$
$\varphi(x) = \frac{1}{2^{\upsilon-1}\Gamma(\upsilon)}\left(cx^{\frac{1}{2}}\right)^{\upsilon} \mathrm{K}_{\upsilon}(cx^{\frac{1}{2}})$	$c > 0,\ \upsilon > 0$
$\varphi(x) = 2^{\upsilon}\left[\exp\left(cx^{\frac{1}{2}}\right) + \exp\left(-cx^{\frac{1}{2}}\right)\right]^{-\upsilon}$	$c > 0,\ \upsilon > 0$

K_{υ}: função de Bessel modificada de segunda espécies de ordem υ (OLIVEIRA, TYGEL, 2005)
Fonte: Gneiting (2002).

Tabela 3 – Funções com derivada completamente monótona $\psi(x)$, $x \geq 0$.

Forma funcional	Parâmetros
$\psi(x) = (ax^{\alpha} + 1)^{\beta}$	$a > 0,\ 0 < \alpha \leq 1,\ 0 \leq \beta \leq 1$
$\psi(x) = \frac{\ln(ax^{\alpha} + b)}{\ln(b)}$	$a > 0,\ b > 1,\ 0 < \alpha \leq 1$
$\psi(x) = \frac{(ax^{\alpha} + b)}{b(ax^{\alpha} + 1)}$	$a > 0,\ 0 < b \leq 1,\ 0 < \alpha \leq 1$

Fonte: Gneiting (2002).

Por exemplo, a escolha de $\varphi(x) = \exp(-cx^{\gamma})$ e $\psi(x) = (ax^{\alpha} + 1)^{\beta}$ na Tabela 2 e Tabela 3, respectivamente, resulta da Equação (27), com $d = 2$, a família de funções de covariância da Equação (28).

$$C(\boldsymbol{h}, u) = \frac{\sigma^2}{(a|u|^{2\alpha} + 1)^{\beta}}\ exp\left(-\frac{c\|\boldsymbol{h}\|^{2\gamma}}{(a|u|^{2\alpha} + 1)^{\beta\gamma}}\right),\ \ \forall (h, u) \in D \times T \tag{28}$$

Observe-se que na Equação (28), para $\beta = 0$ a família de funções de covariância não depende do lag temporal. Multiplicando-se a Equação (28) pela função puramente temporal $C_t(u) = (a|u|^{2\alpha} + 1)^{-\delta},\ \ \delta \geq 0,$ resulta uma função de covariância espaço-temporal válida (Equação(29)).

$$C(\boldsymbol{h}, u) = \frac{\sigma^2}{(a|u|^{2\alpha} + 1)^{\delta+\beta}}\ exp\left(-\frac{c\|\boldsymbol{h}\|^{2\gamma}}{(a|u|^{2\alpha} + 1)^{\beta\gamma}}\right),\ \ \forall (h, u) \in D \times T \tag{29}$$

em que $a > 0,\ c > 0,\ 0 < \alpha \leq 1,\ 0 \leq \beta \leq 1,\ 0 < \gamma \leq 1, \delta \geq 0$ são os parâmetros do modelo e σ^2 é a variância a priori do campo espaço-temporal.

De acordo com Schabenberger e Gotway (2005), quando $C(h,u)$ é uma função de covariância espaço-temporal estacionária em $\mathbb{R}^d \times \mathbb{R}$, as marginais $C(h,0)$ e $C(0,u)$ são funções de covariâncias puramente espaciais em $\mathbb{R}^d$ e puramente temporal em $\mathbb{R}$, respectivamente.

Os principais modelos de covariância espaço-temporal utilizados neste estudo, de acordo com Montero, Fernández-Avilés e Mateu (2015) e Gräler, Pebesma e Heuvelink (2016), foram:

2.3.1 Modelos separáveis (modelo soma e modelo produto)

Não considera a interação entre o espaço e o tempo. Assume-se que a função de covariância espaço-temporal (C_{sep}) pode ser representada como soma ou produto de uma componente puramente espacial e outra puramente temporal, da forma:

$$C_{sep}\left((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)\right) = C_s(\boldsymbol{s}_i, \boldsymbol{s}_j) + C_t(t_i, t_j) \tag{30}$$

ou

$$C_{sep}\left((s_i, t_i), (\boldsymbol{s}_j, t_j)\right) = C_s(\boldsymbol{s}_i, \boldsymbol{s}_j)\, C_t(t_i, t_j) \tag{31}$$

para todo $s_i, s_j \epsilon\, D$ e $t_i, t_j \epsilon\, T$ sendo C_s e C_t funções de covariâncias puramente espacial e puramente temporal, respectivamente.

Se C_s e C_t são estacionárias de segunda ordem, o modelo soma e o modelo produto tornam-se, respectivamente:

$$C_{sep}(\boldsymbol{h}, u) = C_s(\boldsymbol{h}) + C_t(u) \tag{32}$$

e

$$C_{sep}(\boldsymbol{h}, u) = C_s(\boldsymbol{h})\, C_t(u) \tag{33}$$

para todo $(\boldsymbol{h}, u) \epsilon\, D \times T$.

Quanto a função variograma espaço-temporal, no caso estacionário, o modelo soma e modelo produto tomam-se, respectivamente, a forma:

$$\begin{aligned} \gamma_{sep}(\boldsymbol{h}, u) &= C_{sep}(\boldsymbol{0}, 0) - C_{sep}(\boldsymbol{h}, u) = C_s(\boldsymbol{0}) + C_t(0) - C_s(\boldsymbol{h}) - C_t(u) \\ &= \gamma_{sep}(\boldsymbol{h}) + \gamma_{sep}(u) \end{aligned} \tag{34}$$

e

$$\begin{aligned}\gamma_{sep}(\boldsymbol{h},u) &= C_{sep}(\mathbf{0},0) - C_{sep}(\boldsymbol{h},u) = C_s(\mathbf{0})\, C_t(0) - C_s(\boldsymbol{h})\, C_t(u) \\ &= C_s(\mathbf{0})C_t(0) - (C_s(\mathbf{0}) - \gamma_s(\boldsymbol{h}))(C_t(0) - \gamma_t(u)) \\ &= C_s\,(\mathbf{0})\gamma_t\,(u) + C_t\,(0)\gamma_s\,(\boldsymbol{h}) - \gamma_s\,(\boldsymbol{h})\gamma_t\,(u)\end{aligned} \tag{35}$$

em que γ_s e γ_t são os variogramas puramente espacial e puramente temporal correspondentes a C_s e C_t, respectivamente.

2.3.2 Modelo soma-produto

No caso estacionário, as funções de covariância (C_{sp}) e variograma (γ_{sp}) tomam-se, respectivamente, as formas:

$$C_{sp}(\boldsymbol{h},u) = k_1 C_s(\boldsymbol{h}) C_t\,(u) + k_2 C_s(\boldsymbol{h}) + k_3 C_t(u) \tag{36}$$

e

$$\begin{aligned}\gamma_{sp}(\boldsymbol{h},u) &= C_{sp}(\mathbf{0},0) - C_{sp}(\boldsymbol{h},u) \\ &= \big(k_2 + k_1 C_t(0)\big)\,\gamma_s\,(\boldsymbol{h}) + \big(k_3 + k_1 C_s(\mathbf{0})\big)\,\gamma_t(u) - k_1\gamma_s(\boldsymbol{h})\,\gamma_t(u) \\ &= \big(k_2 + k_1\, sill_t\big)\,\gamma_s(\boldsymbol{h}) + \,(k_3 + k_1 sill_s)\,\gamma_t(u) - k_1\gamma_s(\boldsymbol{h})\,\gamma_t(u)\end{aligned} \tag{37}$$

em que C_s e C_t, são funções de covariância espacial e temporal, γ_s e γ_t os correspondentes variogramas espacial e temporal, $k_1 > 0$, $k_2 \geq 0$ e $k_3 \geq 0$ são constantes para assegurar a validade de C_{sp}.

Nota-se que, estimar e modelar as funções $\gamma_{sp}\,(\boldsymbol{h},0)$ (variograma marginal espacial) e $\gamma_{sp}\,(\mathbf{0},u)$ (variograma marginal temporal) é equivalente a estimar e modelar, respectivamente, as funções de variograma puramente espacial $\gamma_s\,(\boldsymbol{h})$ e puramente temporal $\gamma_t\,(u)$, já que

$$\gamma_{sp}\,(\boldsymbol{h},0) = \big(k_2 + k_1\, C_t(0)\big)\,\gamma_s\,(\boldsymbol{h}) = k_s\gamma_s\,(\boldsymbol{h}) \tag{38}$$

e

$$\gamma_{sp}\,(\mathbf{0},u) = \big(k_3 + k_1\, C_s(0)\big)\gamma_t\,(\boldsymbol{u}) = k_t\gamma_t\,(u). \tag{39}$$

Da Equação (36) resulta que o patamar geral $C_{sp}(\mathbf{0},0)$, denotado por $sill_{st}$, é dado por:

$$\begin{aligned}sill_{st} &= k_1 C_s(\mathbf{0}) C_t(0) + k_2\, C_s(\mathbf{0}) + k_3 C_t(0) \\ &= k_1\, sill_s\, sill_t + k_2\, sill_s + k_3\, sill_t\end{aligned} \tag{40}$$

em que os parâmetros k_1, k_2 e k_3, satisfazem as equações

$$\begin{cases} k_1 C_t\,(0) + k_2 = k_s \\ k_1 C_s\,(\mathbf{0}) + k_3 = k_t \\ k_1 C_s(\mathbf{0}) C_t(0) + k_2 C_s(\mathbf{0}) + k_3 C_t(0) = C_{ps}(\mathbf{0},0) \end{cases} \tag{41}$$

Os parâmetros k_1, k_2 e k_3, expressos em termos dos patamares $C_{sp}(\mathbf{0},0)$, $C_s(\mathbf{0})$, $C_t(0)$ e dos coeficientes k_s e k_t, tornam-se:

$$\begin{cases} k_1 = \dfrac{k_s C_s(\mathbf{0}) + k_t C_t(0) - C_{sp}(\mathbf{0},0)}{C_s(\mathbf{0}) C_t(0)} \\ k_2 = \dfrac{C_{sp}(\mathbf{0},0) - k_t C_t(0)}{C_s(\mathbf{0})} \\ k_3 = \dfrac{C_{sp}(\mathbf{0},0) - k_s C_s(\mathbf{0})}{C_t(0)} \end{cases} \tag{42}$$

Assim, utilizando as Equações (37), (38), (39) e (42), a função variograma espaço-temporal do modelo soma-produto, expressa com um único parâmetro, torna-se:

$$\begin{aligned} \gamma_{sp}(\boldsymbol{h},u) &= k_s\gamma_s(\boldsymbol{h}) + k_t\gamma_t(u) - k_1\gamma_s(\boldsymbol{h})\gamma_t(u) \\ &= \gamma_{sp}(\boldsymbol{h},0) + \gamma_{sp}(\mathbf{0},u) - k_1\frac{\gamma_{sp}(\boldsymbol{h},0)}{k_s}\frac{\gamma_{sp}(\mathbf{0},u)}{k_t} \\ &= \gamma_{sp}(\boldsymbol{h},0) + \gamma_{sp}(\mathbf{0},u) - k\gamma_{sp}(\boldsymbol{h},0)\gamma_{sp}(\mathbf{0},u) \end{aligned} \tag{43}$$

em que

$$k = \frac{k_1}{k_s k_t} = \frac{k_s C_s(\mathbf{0}) + k_t C_t(0) - C_{sp}(\mathbf{0},0)}{\big(k_s C_s(\mathbf{0})\big)\big(k_t C_t(0)\big)}. \tag{44}$$

2.3.3 Modelo métrico

Combina as distâncias espacial, temporal e espaço-temporal por uma correção de anisotropia espaço-temporal $\kappa > 0$ (número de unidades espacial equivalentes a uma unidade temporal), resultam em um único modelo de covariância conjunto C_{st} para as covariâncias espacial e temporal, com possíveis mudanças no alcance.

Segundo Montero, Fernández-Avilés e Mateu (2015), a função de covariância espaço-temporal para o modelo métrico (C_m), no caso estacionário, toma-se a forma

$$C_m(\boldsymbol{h},u) = C_{st}(\|\boldsymbol{h}\| + \kappa|u|), \tag{45}$$

em que $\|\boldsymbol{h}\| + \kappa|u|$ é uma métrica em $\mathbb{R}^d \times \mathbb{R}$.

Uma alternativa para o modelo métrico é utilizar a métrica equivalente $\|\boldsymbol{h}\|^2 + \kappa|u|^2$ ou a métrica $\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}$ (Figura 5). Nesse caso,

$$C_m(\boldsymbol{h},u) = C_{st}(\|\boldsymbol{h}\|^2 + \kappa^2|u|^2) \tag{46}$$

ou

$$C_m(\boldsymbol{h},u) = C_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right) \tag{47}$$

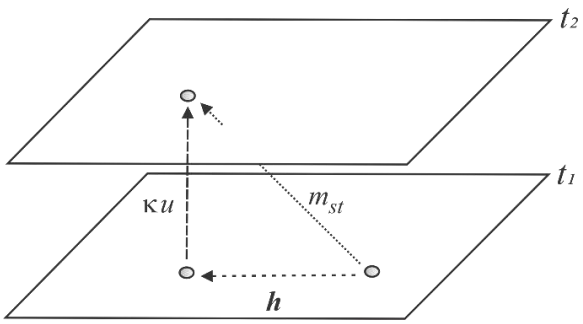


Figura 5 – Representação da distância espacial $\boldsymbol{h}$, da distância temporal u e da distância espaço-temporal $m_{st} = \sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}$ em $\mathbb{R}^2 \times \mathbb{R}$, sendo κ uma correção de anisotropia espaço-temporal.
Fonte: Adaptada de Montero, Fernández-Avilés e Mateu (2015).

Assim, independente das métricas, as funções de covariância marginal espacial e marginal temporal obedecem ao mesmo tipo de modelo e patamar, mesmo que possam vir a ter alcances diferentes.

Quanto à função de variograma espaço-temporal (γ_m), no caso estacionário, o modelo métrico toma-se a forma

$$\gamma_m(\boldsymbol{h}, u) = \gamma_{st}(\|\boldsymbol{h}\| + \kappa|u|) \tag{48}$$

ou

$$\gamma_m(\boldsymbol{h}, u) = \gamma_{st}(\|\boldsymbol{h}\|^2 + \kappa^2|u|^2) \tag{49}$$

ou ainda,

$$\gamma_m(\boldsymbol{h}, u) = \gamma_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right) \tag{50}$$

em que γ_{st} é o variograma conjunto, associado a C_{st}, que inclui o efeito pepita.

2.3.4 Modelo soma-métrico

O modelo de covariância soma-métrico (C_{sm}) é uma combinação do modelo soma e do modelo métrico, com todas as componentes configuradas independentemente. No caso estacionário, ao utilizar a métrica $\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}$, a função covariância é dada por:

$$C_{sm}(\boldsymbol{h}, u) = C_s(\boldsymbol{h}) + C_t(u) + C_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right) \tag{51}$$

em que $\kappa > 0$.

Quanto à função de variograma, ela é representada por:

$$\gamma_{sm}(\boldsymbol{h}, u) = \gamma_s(\boldsymbol{h}) + \gamma_t(u) + \gamma_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right) \tag{52}$$

em que γ_s, γ_t e γ_{st} são os correspondentes variogramas puramente espacial, puramente temporal e conjunto, respectivamente, com parâmetros configurados independentes.

Formas equivalentes das funções de covariância e variograma podem ser obtidas ao utilizar as métricas $\|\boldsymbol{h}\| + \kappa|u|$ ou $\|\boldsymbol{h}\|^2 + \kappa^2|u|^2$. Por exemplo, quando utiliza-se a Equação (51), a expressão da função de covariância soma-métrico exponencial é dada por

$$\begin{aligned} C_{sm}(h, u) = {} & C_s(\mathbf{0})exp\left(-\frac{\|\boldsymbol{h}\|}{a_s}\right) + C_t(0)exp\left(-\frac{|u|}{a_t}\right) \\ & + C_{st}(\mathbf{0}, 0)exp\left(-\frac{\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}}{a}\right) \end{aligned} \tag{53}$$

em que a_s, a_t e a são os alcances dos variogramas estacionários puramente espacial, puramente temporal e conjunto, respectivamente, mas todos exponenciais.

2.3.5 Modelo soma-métrica simplificado

É uma forma mais simples que o modelo soma-métrico porém, ao invés de flexibilidade total para cada componente, adicionalmente introduz-se um único efeito pepita, denotado por C_o. Nesse caso, ao utilizar a métrica $\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}$, a função variograma é dada por:

$$\gamma_{ssm}(\boldsymbol{h}, u) = C_o + \gamma_s(\boldsymbol{h}) + \gamma_t(u) + \gamma_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right). \tag{54}$$

A estimação do modelo teórico de variograma espaço-temporal baseia-se no variograma experimental que é derivado dos dados observados. De acordo com Gräler, Pebesma e Heuvelink (2016), os melhores ajustes dos modelos espaço-temporais podem sugerir diferentes famílias de variogramas e parâmetros para os modelos puramente espacial e puramente temporal.

Nesse estudo, a função aleatória $Z(s_i, t_j)$ representa a temperatura máxima diária do ar, em ^oC, na localização espacial s_i, $i = 1, 2, \cdots, 61$, e no tempo t_j, $j = 1, 2, \cdots, 7671$.

3 RESULTADOS E DISCUSSÃO

Inicialmente, foi realizada uma análise descritiva dos dados com o objetivo de obter padrões que auxiliam no processo de modelagem geoestatística espaço-temporal (Tabela 4).

Tabela 4 – Estatística Descritiva dos valores Observados da Temperatura Máxima Diária do Ar de 01 de janeiro 1996 a 31 de dezembro 2016 nas 61 estações de estudo.

Estatísticas	**Resultados**
Média ($\bar{Z}$)	29,3 ºC
Desvio Padrão (S)	3,8 ºC
Mínima	9,0 ºC
Máxima	42,6 ºC
Primeiro Quartil (Q_1)	27,0 ºC
Mediana (M_d)	29,5 ºC
Terceiro Quartil (Q_3)	31,9 ºC
Coeficiente de Variação (CV)	12,8%
Coeficiente de Assimetria (C_a)	–0,3
Coeficiente de Curtose (C_m)	3,1

Com base no critério de classificação proposto por Garcia (1989), o atributo temperatura máxima do ar apresentou média dispersão dos dados ($10\% < CV < 20\%$).

Os resultados da Tabela 4 e Figura 6 sugerem uma possível distribuição simétrica para os dados de temperatura máxima do ar em MG, já que a distribuição de frequência apresentou média e mediana bem próximas, além de um coeficiente de assimetria próximo de zero.

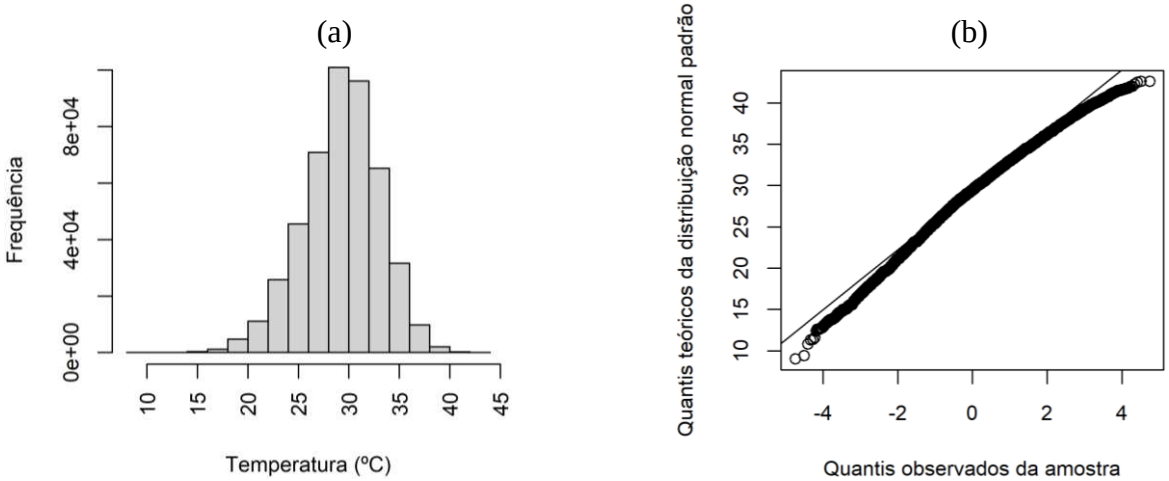


Figura 6 – (a) Histograma dos dados; (b) Gráfico qq-plot.

Segundo Yamamoto e Landim (2015), para dados que apresentem distribuição normal ou coeficiente de assimetria negativa, não há necessidade de nenhuma transformação dos dados e a krigagem ordinária pode ser aplicada diretamente sobre o conjunto de dados observados. Dessa forma, as estimativas geoestatísticas foram realizadas usando o preditor krigagem ordinária, sem nenhuma transformação nos dados.

Com base nas semivariâncias estimadas pela Equação (25) em que $Z(s_i, t_j)$ representa os valores da temperatura máxima diária do ar, em ^{o}C, na localização espacial s_i, $i = 1, 2, \cdots, 61$, e no tempo t_j, j$= 1, 2, \cdots, 7671$, foram geradas série dos variogramas experimentais para 12 lags temporais (Figura 7a) e a superfície suavizada tridimensional do variograma espaço-temporal experimental (Figura 8a).

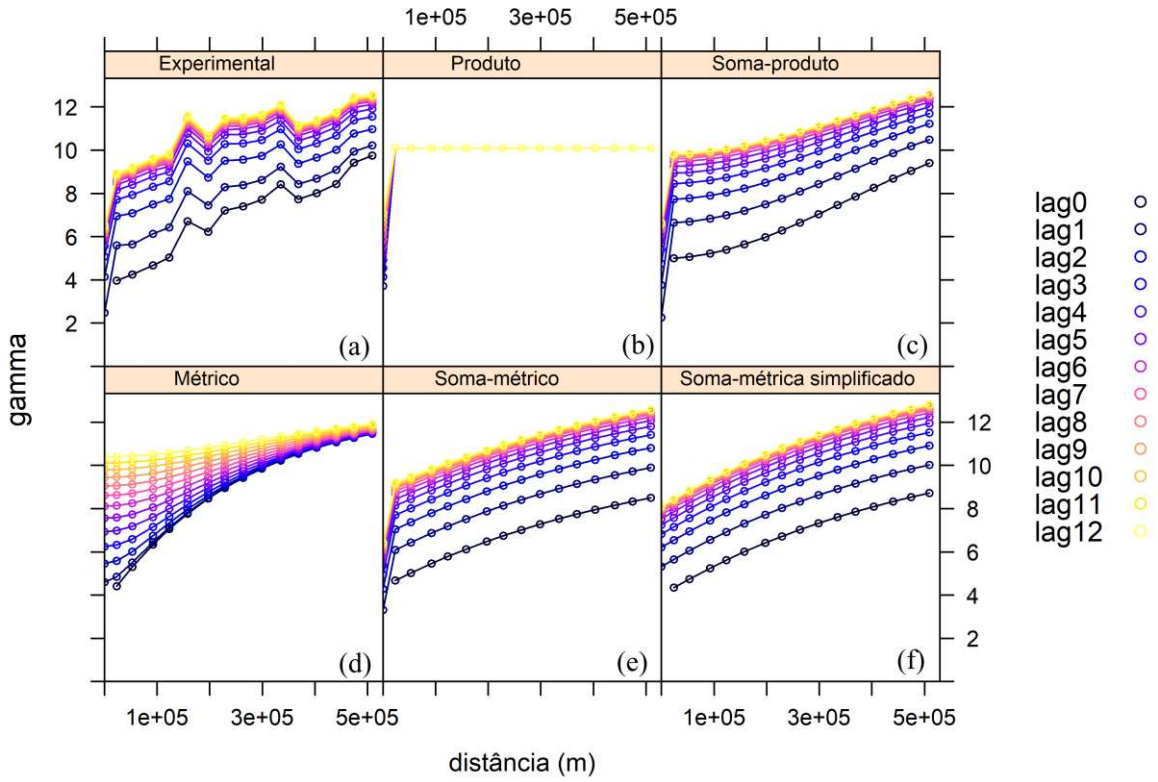


Figura 7 – Série dos variogramas para 12 lags temporais (a) Experimental, (b)-(f) Teóricos ajustados às semivariâncias estimadas.

Em seguida, foram ajustados às semivariâncias estimadas, modelos de variogramas teóricos espaço-temporais em que utilizou-se funções de covariâncias da família de Gneiting (2002) e os modelos separáveis e não separáveis apresentados neste estudo. A seleção dos

modelos e métodos de ajuste foi efetuada quanto ao critério do MSE, conforme sugerido por Pebesma e Gräler (2018).

Além da seleção da família de modelos de variogramas espaço-temporal, cada componente do modelo (puramente espacial, puramente temporal ou conjunto) foram ajustadas a partir de modelos de variogramas isotrópicos unidimensional.

Já as estimações dos parâmetros dessas componentes foram obtidas conjuntamente, via procedimentos numéricos iterativos, com valores iniciais retirados do variograma experimental. Para o ajuste da componente puramente espacial a série na defasagem de tempo lag0 foi utilizada e para a componente temporal o variograma marginal temporal $\hat{\gamma}(\mathbf{0}, u)$.

Para uma comparação visual do ajuste dos modelos foram geradas, para 12 lags temporais, as séries dos variogramas teóricos (Figura 7b -7f) e sua superfície tridimensional, para cada modelo (Figura 8b - 8f).

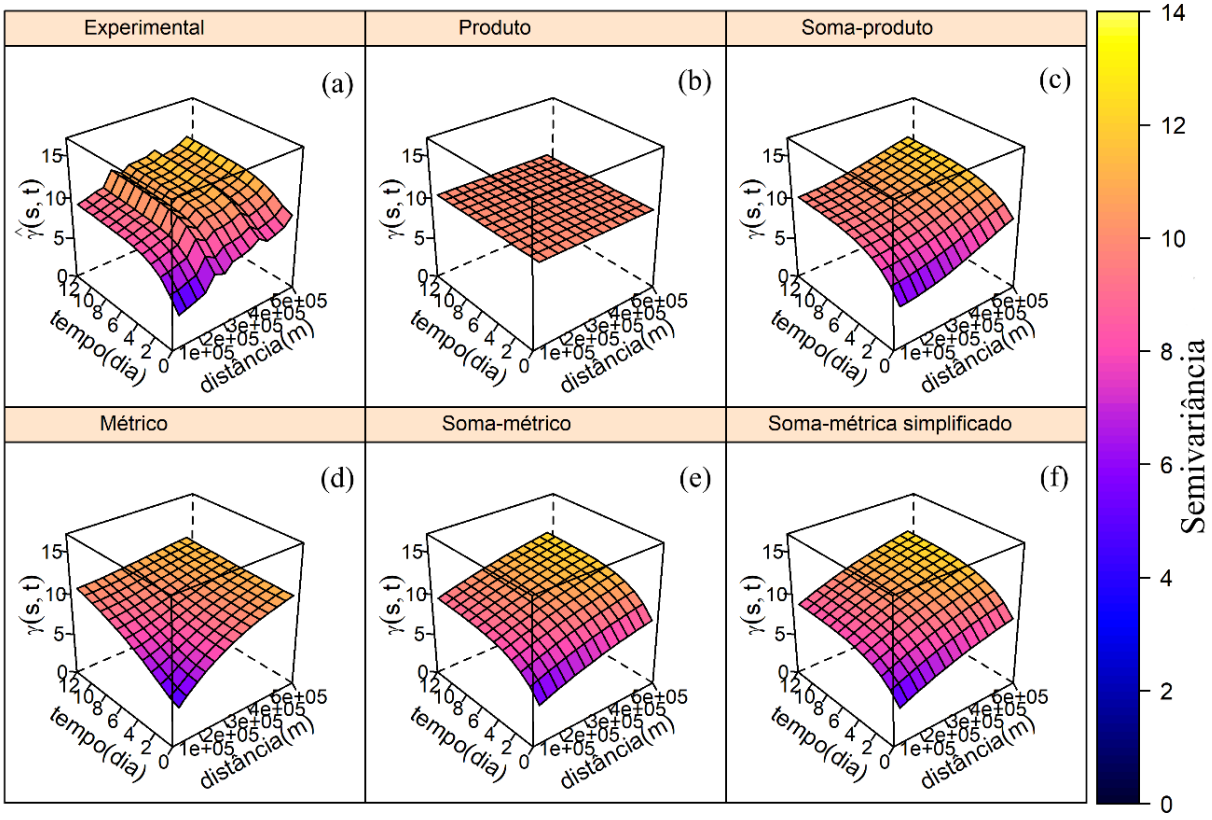


Figura 8 – Superfície dos variogramas espaço-temporais. (a) Experimental; (b)-(f) Modelos teóricos ajustados às semivariâncias estimadas.

Observa-se que o modelo de covariância espaço-temporal produto não foi adequado para ser utilizado na predição da temperatura máxima diária do ar no estado de MG, pois não ocorreu dependência espacial e temporal (modelo de pepita puro) (Figuras 7b e 8b).

A Tabela 5 apresenta os valores do MSE dos modelos produto, soma-produto, métrico, soma-métrico e soma-métrico simplificado de variogramas teóricos espaço-temporais ajustados. Pode-se observar que, os MSE´s de todos os modelos ajustados foram relativamente baixos, devido ao fato dos dados observados apresentarem variações pequenas e cada modelo tomar como base a média dos dados.

Tabela 5 – Indicador de qualidade dos modelos.

	Produto	Soma-produto	Métrico	Soma-métrico	Soma-métrico simplificado
MSE	2,86	0,37	1,34	0,24	0,49

Dentre os modelos teóricos espaço-temporal analisados, o mais apropriado para ser utilizado na predição da temperatura máxima do ar no estado de MG, no período de 1996 a 2016, foi o soma-métrico por apresentar-se o mais acurado, com menor erro quadrático médio ($\mathrm{MSE} = 0{,}24$), além de aspecto visual da série dos variogramas (Figura 7e) e superfície tridimensional (Figura 8e), que mais se aproxima do experimental.

A Tabela 6 apresenta as estimativas dos parâmetros dos variogramas estacionários puramente espacial, puramente temporal e conjunto do modelo ajustado soma-métrico, sendo todos exponenciais.

Tabela 6 – Estimativas dos parâmetros para o modelo ajustado soma-métrico exponencial.

	C_o	C_1	a
Espacial	2,5	5,1	500 km
Temporal	0	4	2,4 $dias$
Conjunto	1,9	1,3	500 km

C_o: efeito pepita, C_1: contribuição, a: alcance.

Dessa forma, ao utilizar a anisotropia espaço-temporal estimada antecipadamente segundo a função estiStAni do pacote gstat do R em $\kappa = 22{,}3\ \mathrm{k}m/dia$, os valores das estimativas dos parâmetros apresentados na Tabela 6 e, também, o modelo de variograma

teórico exponencial unidimensional, tem-se que a função do modelo ajustado do variograma espaço-temporal soma-métrico, segundo a Equação (52), é dada por

$$\begin{aligned}\gamma_{sm}(\boldsymbol{h},u) &= \gamma_s(\boldsymbol{h}) + \gamma_t(u) + \gamma_{st}\left(\sqrt{\|\boldsymbol{h}\|^2 + \kappa^2|u|^2}\right) \\ &= 2{,}5 + 5{,}1\left[1 - exp\left(-\frac{\|\boldsymbol{h}\|}{500000}\right)\right] + 4\left[1 - exp\left(-\frac{|u|}{2{,}4}\right)\right] + 1{,}9 \\ &+ 1{,}3\left[1 - exp\left(-\frac{\sqrt{\|\boldsymbol{h}\|^2 + (22300)^2|u|^2}}{500000}\right)\right]\end{aligned} \tag{55}$$

A Figura 9 ilustra as superfícies dos variogramas espaço-temporal experimental e do modelo ajustado soma-métrico.

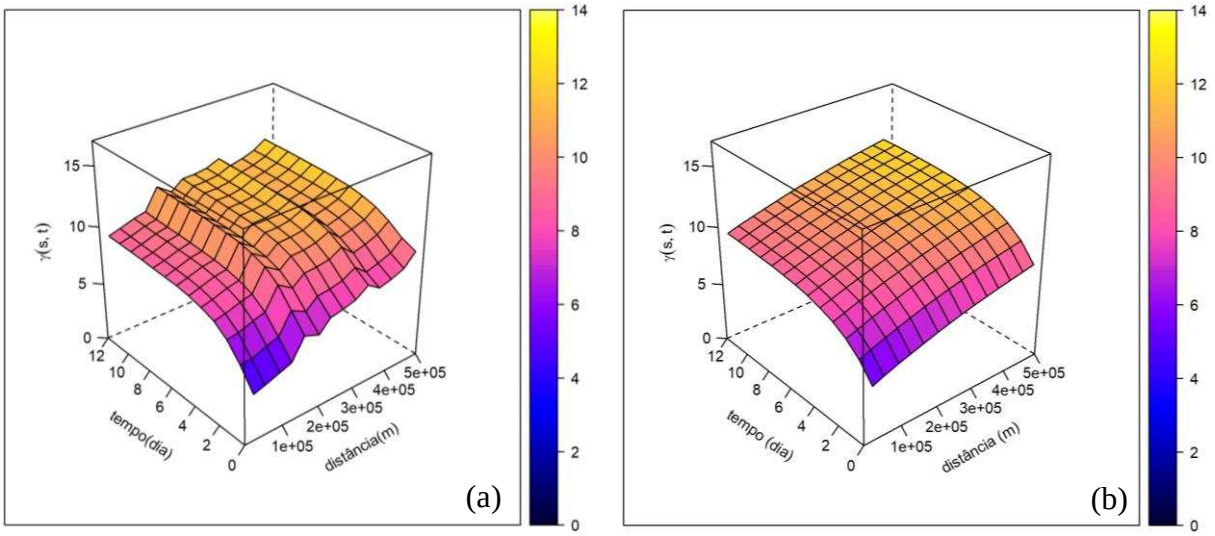


Figura 9 – (a) Superfície do variograma espaço-temporal experimental. (b) Superfície variograma espaço-temporal do modelo soma-métrico ajustado às semivariâncias estimadas Superfície dos variogramas espaço-temporais.

Então, após a caracterização do padrão espaço-temporal das temperaturas, aplicou-se o interpolador linear da Geoestatística, krigagem ordinária, usando o modelo de covariância soma-métrico, com 50 vizinhos espaço-temporais, para gerar campos interpolados em uma grade regular, com predição da temperatura máxima do ar para o estado de MG, nos dias 15 de cada mês, ao longo dos 21 anos analisados.

Ressalta-se que, os métodos aqui apresentados também podem ser aplicados para outros dias do mês ou, até mesmo, para todos os dias do ano. No entanto, o custo computacional para a obtenção das temperaturas preditas em uma grade regular com 5869

pontos, abrangendo todo o estado de Minas Gerais (1240 $km \times 1000\ km$) é relativamente alto. Em uma máquina com processador Intel core I7 de 2,40 GHz, o tempo de processamento, para cada dia, foi de aproximadamente 5,2 minutos.

Os valores médios preditos nos 5869 pontos de grade foram comparados com a média dos dados observados nas 61 estações (Figura 10).

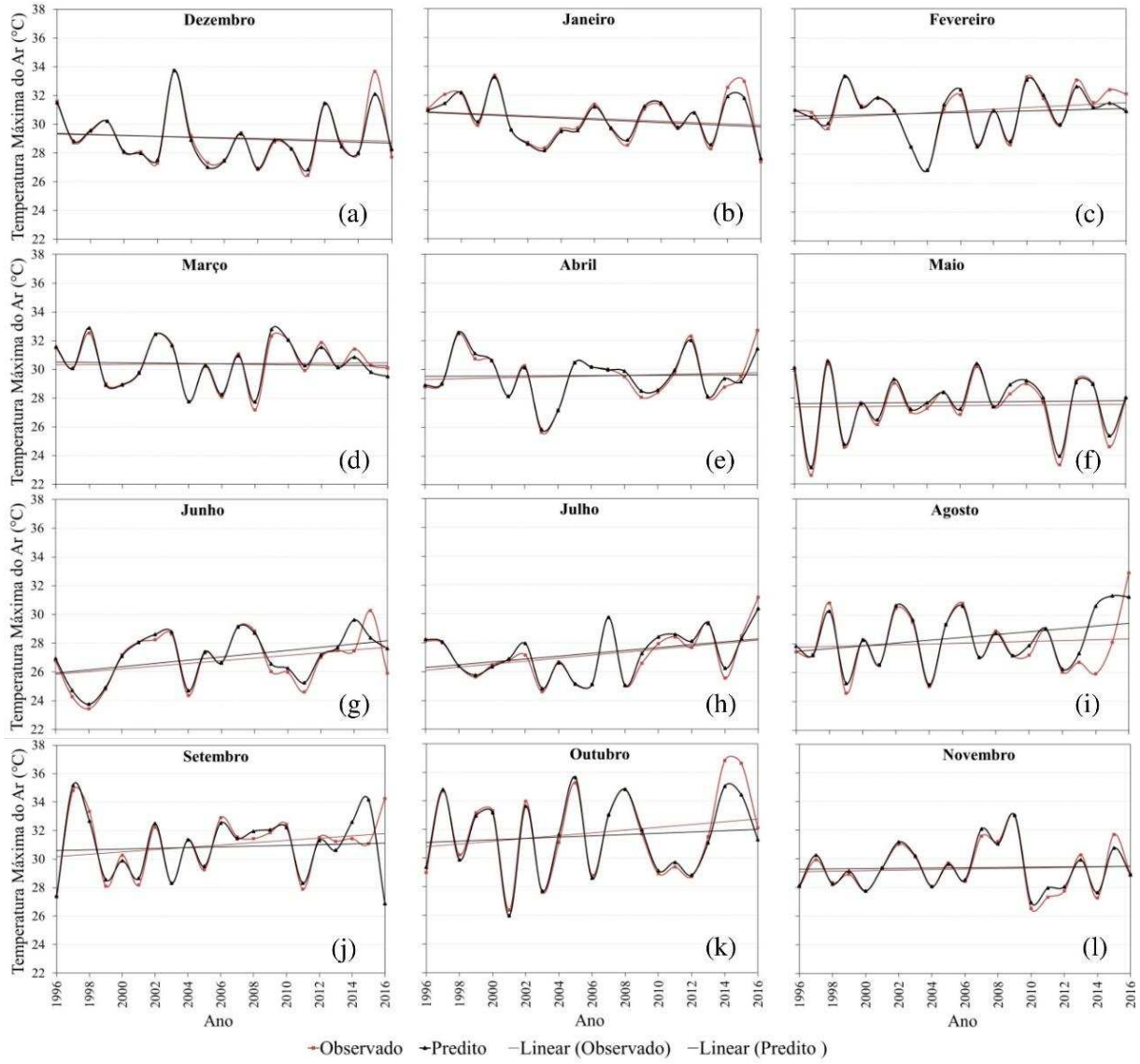


Figura 10 – Médias dos dados observados e das predições de temperatura máxima diária do ar no estado de MG, no período 1996 – 2016, para o dia 15 dos meses de (a) Dezembro, (b) Janeiro, (c) Fevereiro, (d) Março, (e) Abril, (f) Maio, (g) Junho, (h) Julho, (i) Agosto, (j) Setembro, (k) Outubro, (l) Novembro.

Observa na Figura 10, que há um bom ajuste entre as médias dos dados observados e preditos, embora nos últimos anos a diferença foi mais acentuada, provavelmente devido ao fato de que nos últimos anos aumentou consideravelmente o número de falhas, uma vez que algumas das estações foram desativadas. No último ano analisado (2016) ocorreram 34,8% de falhas nos dados observados. Essas falhas afetam tanto as médias preditas quanto as observadas.

Nota-se que nos meses de dezembro e janeiro há uma tendência de queda na temperatura máxima do ar no estado de MG, considerando os 21 anos analisados. No entanto, é possível observar que a partir de 2004 é evidenciado um aumento da média da temperatura máxima para toda a estação de verão (Figura 11).

Para os meses de março a maio (outono) não se observam tendências de aquecimento ou resfriamento (Figura 10d – 10f).

Nos meses de junho a agosto (inverno) observa-se uma tendência de aumento, aproximadamente de 1ºC, na média da temperatura máxima, a cada 10 anos. Isso pode ser um indício de que a estação mais fria do ano está ficando cada vez mais quente (Figura 10g - 10i).

Já na primavera, também notou-se tendência de aquecimento, principalmente nos meses de setembro e outubro (Figura 10j – 10l).

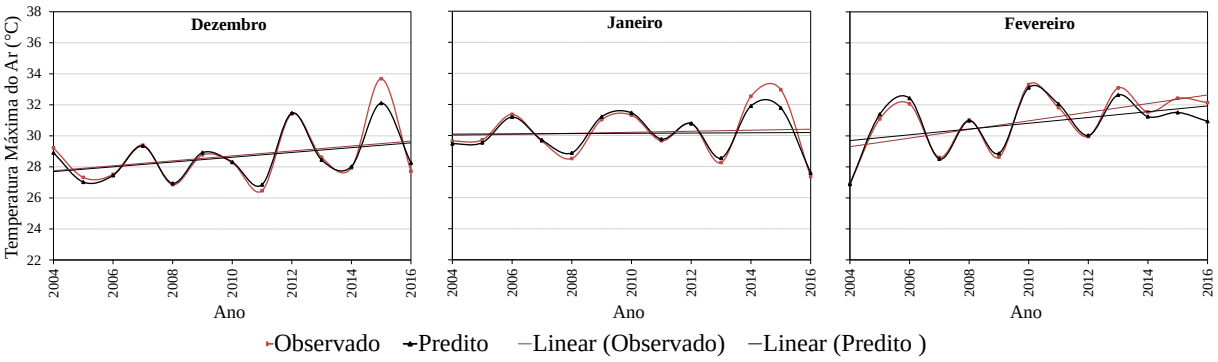


Figura 11 – Médias dos dados observados e das predições de temperatura máxima diária do ar no estado de MG, no período 2004 – 2016, para o dia 15 dos meses de (a) Dezembro, (b) Janeiro, (c) Fevereiro.

Ao considerar todos os meses do ano observa-se que prevalece leve tendência de aumento da média da temperatura máxima nas duas últimas décadas (Figura 12).

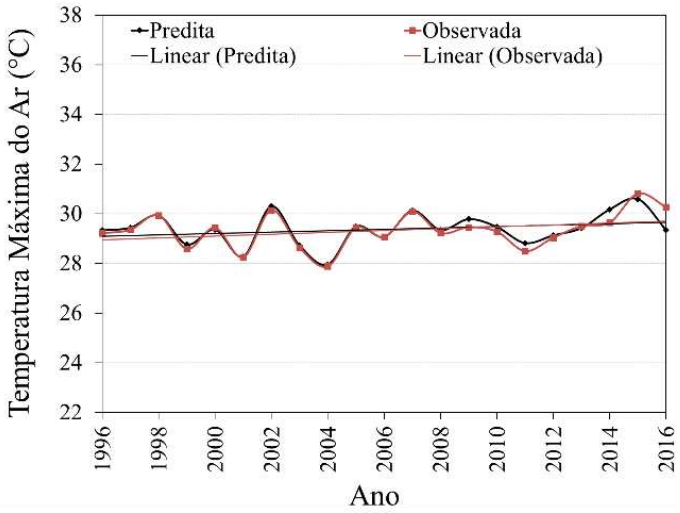


Figura 12 – Média anual dos dados observados e das predições de temperatura máxima diária do ar no estado de MG para o dia 15, no período 1996 – 2016.

Essa tendência de aumento no valor da temperatura máxima pode estar relacionada ao aquecimento global, onde estudos mostram que isso já vem ocorrendo nas últimas décadas (IPCC, 2014). Contudo, avaliando os gráficos das médias das predições dos valores de temperatura máxima diária do ar ao longo dos anos, seria necessário o estudo de um período maior para uma análise do aquecimento global.

A Figura 13 apresenta os mapas das predições espaço-temporais dos valores de temperatura máxima do ar no estado de MG para o dia 15 de janeiro, ao longo dos 21 anos analisados. Nota-se nos mapas que podem ocorrer altas variações espaciais nos valores de temperatura máxima do ar. Pode-se observar essa ocorrência, por exemplo, no dia 15 de janeiro de 2003, em que a região do Triângulo Mineiro apresentou temperatura máxima em torno de 32ºC e a região Metropolitana de Belo Horizonte abaixo de 26ºC.

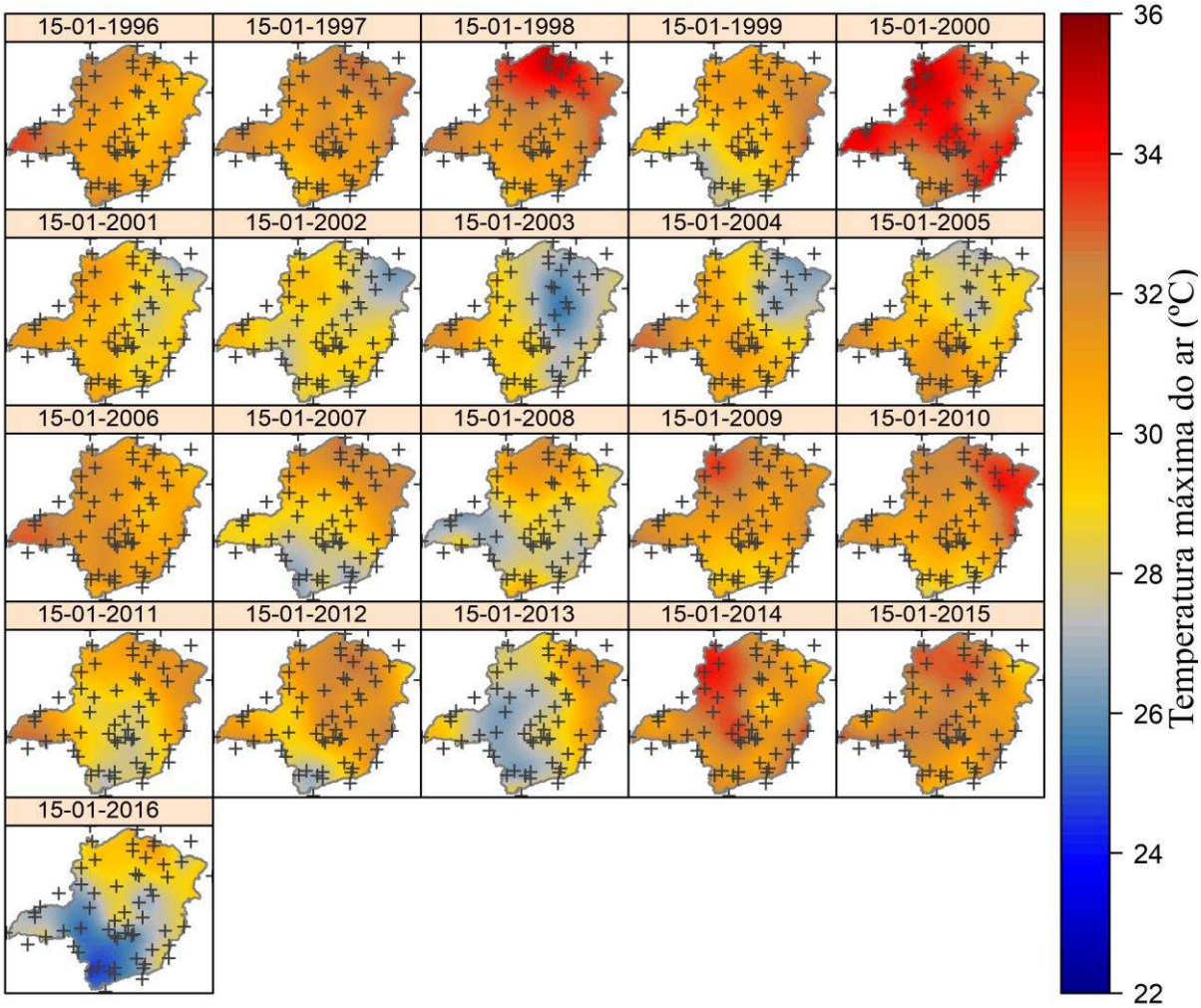


Figura 13 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de janeiro, no período de 1996 a 2016.

Observa-se ainda que, o dia 15 de janeiro de 2016 foi bastante frio, principalmente no sul do estado de MG. No entanto, ao analisar dias próximos a este, pode-se notar que este baixo valor de temperatura máxima provavelmente foi devido à chegada de uma frente fria, uma vez que no dia 10 o estado de MG estava com temperaturas altas e foi resfriando, do sul para o norte, até o dia 17 (Figura 14).

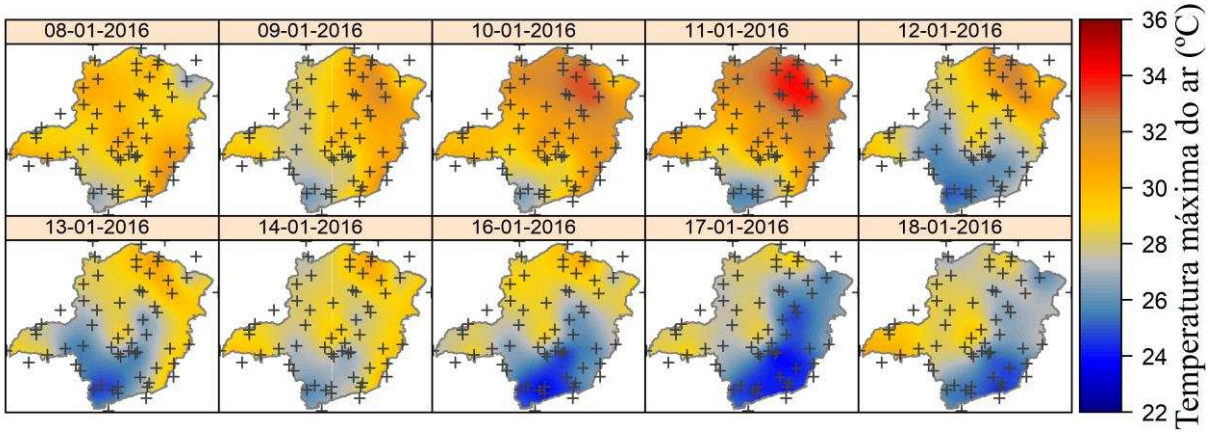


Figura 14 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG nos dias 08 a 18 de janeiro de 2016, exceto dia 15.

Então, devido a alta variabilidade da temperatura, tanto espacial quanto temporal, justifica-se a necessidade de utilizar grade com alta resolução e uma interpolação que considera as variações espaciais e temporais simultaneamente.

4 CONCLUSÕES

Com base no critério MSE, o modelo de covariância soma-métrico foi o que melhor se ajustou aos dados (MSE = 0,24), quando comparados com os demais modelos aqui avaliados. Já o modelo de covariância espaço-temporal produto não é adequado para o conjunto de dados aqui utilizados, pois ele não possui dependência espacial nem temporal.

As médias espaciais dos dados preditos ficaram muito próximas das médias dos dados observados nas 61 estações utilizadas, com exceção dos últimos três anos avaliados, por conter muitas falhas nas observações. Este resultado comprova que as predições realizadas foram bem sucedidas. Dessa forma, os campos espaciais gerados em grade regular possibilitaram visualizar o comportamento da temperatura máxima do ar sobre o estado de Minas Gerais. Assim, ficou evidenciada a adequabilidade do uso da geoestatística espaço-temporal, via funções de covariância, para estimar a temperatura máxima do ar no estado de MG no período de 1996 a 2016.

Foram observadas evidências do efeito do aquecimento global no estado de Minas Gerais, principalmente nos meses da estação de inverno, onde fica evidenciado uma tendência de aumento da temperatura máxima do ar ao longo dos anos analisados. No entanto, para uma melhor avaliação desse efeito é necessário um estudo de um período maior para certificar que as tendências encontradas não foram causadas por variações naturais do clima.

Este trabalho mostrou que é possível obter campos espaciais para a temperatura máxima diária do ar na região que engloba o estado de MG com alta qualidade ao utilizar a metodologia da geoestatística espaço-temporal que considera tanto as variabilidades espaciais quanto temporais. Ademais, esta metodologia não é restrita aos dados aqui utilizados. O código desenvolvido, utilizando o pacote spacetime, permite gerar campos espaço-temporais com grades regulares para qualquer resolução, localização ou variáveis. Logo, mostrou-se uma poderosa ferramenta adequada para a modelagem conjunta espaço-temporal, disponível para o programa de código aberto R.

REFERÊNCIAS BIBLIOGRÁFICAS

BIVAND, R.; LEWIN-KOH, N. Maptools: Tools for reading and handling spatial objects. **R package version 0.9-2**, 2017.

BIVAND, R.; KEITT, T.; ROWLINGSON, B. Rgdal: Bindings for the geospatial data abstraction library. **R package version 1.3-6**, 2018.

CRESSIE, N.; HUANG, H. Classes of nonseparable, spatio-temporal stationary covariance functions**. Journal of the American Statistical Association**, v. 94, n. 448, p. 1330–1339, 1999.

DE IACO, S.; POSA, D.; MYERS, D. E. Characteristics of some classes of space–time covariance functions. **Journal of Statistical Planning and Inference**, v. 143, n. 11, p. 2002-2015, 2013.

DOTY, B.; KINTER, J. L. III. The grid analysis and display system (GrADS): a desktop tool for earth science visualization. In: **American geophysical union 1993 fall meeting**, p. 6–10, 1993. Disponível em: <http://cola.gmu.edu/grads/>. Acesso em: 03 de jun. 2018.

GARCIA, C. H. **Tabelas para classificação do coeficiente de variação**. Piracicaba: IPEF. (Circular Técnica, 171). 12p, 1989.

GARCIA, A.; ANDRE, R. G. B. Variabilidade temporal da temperatura do ar em Jaboticabal-SP: **Nucleus**, v. 12, n. 1, p. 189-198, 2015.

GESCH, D. B.; VERDIN, K. L.; GREENLEE, S. K. New land surface digital elevation model covers the Earth. **EOS, transactions american geophysical union**, v. 80, n. 6, p. 69-70, 1999.

GNEITING, T. Nonseparable, stationary covariance functions for space-time data. **Journal of the American Statistical Association**, v. 97, n. 458, p. 590–600, 2002.

GNEITING, T.; RAFTERY, A. E. Strictly proper scoring rules, prediction,and estimation. **Journal of the American Statistical Association**, v. 102, n. 477, p. 359-378, 2007.

GOMES, D. P. et al. Estimativa da temperatura do ar e da evapotranspiração de referência no estado do Rio de Janeiro. **Irriga**, v. 19, n. 2, p. 302-314, 2014.

GRÄLER, B.; PEBESMA, E.; HEUVELINK, G. Spatio-temporal interpolation using gstat. **RFID Journal**, v. 8, n. 1, p. 204-218, 2016.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. 2016. Disponível em: <https://portaldemapas.ibge.gov.br/portal.php#homepage>. Acesso em: 23 jun. 2017.

INMET - INSTITUTO NACIONAL DE METEOROLOGIA. **Banco de Dados Meteorológicos para Ensino e Pesquisa** (BDMEP), 2018. Disponível em: <http://www.inmet.gov.br/portal>. Acesso em: 02 abr. 2018.

IPCC - INTERGOVERNMENTAL PANEL IN CLIMATE CHANGE. Summary for policymakers. In: **Climate Change 2014 - IPCC**: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T. E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1-32, 2014. Disponível em: <https://www.ipcc.ch/pdf/assessment-report/ar5/wg2/ar5_wgII_spm_en.pdf>. Acesso em: 21 jun. 2018.

KYRIAKIDIS, P. C.; JOURNEL, A. G. Geostatistical space–time models: a review. **Mathematical geology**, v. 31, n. 6, p. 651-684, 1999.

MATHERON, G. The internal consistency of models in geostatistics. In: **Geostatistics**. Springer, Dordrecht,. p. 21-38, 1989.

MEDEIROS, R. M. et al. Variabilidade da temperatura média do ar no Estado da Paraíba-Brasil. **Revista Brasileira de Geografia Física**, v. 8, n. 01, p. 128-135, 2015.

MG.GOV.BR - GOVERNO DE MINAS GERAIS. **Governo do Estado de Minas Gerais**. conheça Minas Gerais – geografia, 2017. Disponível em: <http://mg.gov.br/conheca-minas/geografia>. Acesso em: 12 jun. 2018.

MONTERO, J. M.; FERNÁNDEZ-AVILÉS, G.; MATEU, J. **Spatial and spatio-temporal geostatistical modeling and kriging**. Chennai, India: John Wiley e Sons, 2015.

MOREIRA, D. S. et al. Modeling the radiative effects of biomass burning aerosols on carbon fluxes in the Amazon region. **Atmospheric Chemistry and Physics**, v. 17, n. 23, p.14785-14810, 2017. Disponível em: <https://doi.org/10.5194/acp-17-14785-2017>. Acesso em: 12 de dez 2017.

OLIVEIRA, E. C. de; TYGEL, M. **Métodos matemáticos para engenharia**. São Carlos: SBMAC, 2005.

PERIN, E. B. et al. Interpolação das variáveis climáticas temperatura do ar e precipitação: revisão dos métodos mais eficientes. **Geografia**, v. 40, n. 2, 2015.

PEBESMA, E. J.; BIVAND, R. S. Classes and methods for spatial data in R. **R news**, v. 5, n. 2, p. 9-13, 2005.

PEBESMA, E. **Handling and analyzing spatial, spatiotemporal and movement data.** 2016. Disponível em: <https://edzer.github.io/UseR2016/>. Acesso em: 30 de março 2018.

PEBESMA, E.; GRÄLER, B. Gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation. **R package version 1.1-6**, 2018.

PEBESMA, E. Spacetime: Classes and methods for spatio-temporal data. **R package version 1.2-2**, 2018.

RAJA, N. B. et al. Space-time kriging of precipitation variability in Turkey for the period 1976–2010. **Theoretical and Applied Climatology**, v. 129, n. 1-2, p. 293-304, 2017.

REBOITA, M. S. et al. Aspectos climáticos do estado de Minas Gerais (climate aspects in minas gerais state). **Revista Brasileira de Climatologia**, v. 17, 2015. Disponível em: <http://revistas.ufpr.br/revistaabclima/article/view/41493/27319>. Acesso em: 14 de nov. 2018.

RIBEIRO JR, P. J.; DIGGLE, P. J. GeoR: Analysis of geostatistical data. **R package version 1.7-5.2**, 2016.

RYAN, J. A.; ULRICH, J. M. Xts: Extensible time series. **R package version 0.11-0**, 2018.

SARKAR, D. Lattice: Trellis Graphics for R. **R package version 0.20-35**, 2017.

SCHABENBERGER, O.; GOTWAY, C. A. **Statistical methods for spatial data analysis**. Chapman e Hall/CRC, 2005.
SHERMAN, M. **Spatial statistics and spatio-temporal data: covariance functions and directional properties**. John Wiley & Sons, 2011.

STEIN, M. L. Space time-covariance functions. **Journal of the american statistical association**, v. 100, n. 469, p. 310–321, 2005.

TEAM, R. C. R: **A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 23 de abril 2018.

VAROUCHAKIS, E. A. Spatiotemporal geostatistical modelling of groundwater level variations at basin scale: a case study at Crete's Mires Basin. **Hydrology Research**, v. 49, n.4, p. 1131-1142, 2018.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. São Paulo: Oficina de Textos, 2015.

CAPÍTULO 2 – O USO DA APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DA TEMPERATURA MÁXIMA DO AR

RESUMO

De posse de um conjunto de dados é possível utilizar algoritmos matemáticos e/ou computacionais para caracterizar esse conjunto e dessa forma poder obter informações relevantes a partir dele. Nesse contexto, serão utilizados modelos baseados em aprendizagem de máquina para inferir acerca da temperatura máxima do ar, baseando-se nas informações de outras variáveis medidas em estações meteorológicas instaladas em locais não equidistantes. Este trabalho utilizou dados de 61 estações meteorológicas localizadas no estado de Minas Gerais e regiões adjacentes, que foram coletados durante o ano de 2004. Esses dados foram assimilados pelos algoritmos de regressão linear múltipla, random forest e support vector machine, para obter modelos de predição da temperatura máxima do ar em locais onde não haviam medidas. Também foi utilizada uma interpolação ponderada pelo inverso da distância e a reanálise do European Centre for Medium-Range Weather Forecasts como comparação dos modelos. A reanálise também possibilitou obter predições dos modelos em pontos de grade regular, para gerar mapas espaciais da temperatura máxima do ar no estado de Minas Gerais. Normalmente a temperatura máxima do ar é medida juntamente com as demais variáveis meteorológicas e, portanto, foram realizadas outras predições, agora considerando como covariáveis somente as coordenadas geográficas e a altitude do ponto desejado. Observou-se que devido às altas correlações entre as variáveis meteorológicas o algoritmo de regressão linear múltipla pode ter sofrido influência em sua avaliação, mas ficou muito similar ao support vector machine, que foi considerado o melhor algoritmo de regressão para predizer a temperatura máxima do ar. No entanto, em um dia onde ocorreu a entrada de uma frente fria, que provocou grandes variações espaciais na temperatura do ar, o algoritmo random forest obteve-se a melhor performance.

Palavras-chave:.Random Forest, Support Vector Machine, Regressão.

ABSTRACT

With a set of data it is possible to use mathematical and / or computational algorithms to characterize this set and thus be able to obtain relevant information from it. In this context, models based on machine learning will be used to infer about the maximum air temperature, based on the information of other variables measured in meteorological stations installed in places not equidistant. This work utilized data from 61 meteorological stations located in the state of Minas Gerais and adjacent regions that were collected during the year 2004. These data were assimilated by the multiple linear regression algorithms, random forest and support vector machine, to obtain the prediction models of the maximum air temperature in places where there were no measurements. We also used an inverse time-weighted interpolation and reanalysis of the European Center for Medium-Range Weather Forecasts repository as a comparison of the models. The reanalysis also allowed to obtain predictions of the models in points of regular grid, to generate spatial maps of the maximum air temperature in the state of Minas Gerais. Usually the maximum air temperature is measured along with the other meteorological variables and therefore, predictions were also made considering as covariates only the geographic coordinates and the altitude of the desired point. It was observed that due to the high correlations between the meteorological variables the multiple linear regression algorithm may have influenced its evaluation, but its performance was very similar to the support vector machine, which was considered the best regression algorithm to predict the maximum air temperature. However, on a day where a cold front occurred, which caused large spatial variations in air temperature, the random forest algorithm achieved better performance.

Keywords: Random Forest, Support Vector Machine, regression.

1 INTRODUÇÃO

Além de vários benefícios que a ciência da computação promove nos tempos atuais, ela também está sendo amplamente utilizada para verificar padrões em um determinado conjunto de dados e assim reconhecer semelhanças ou estimar valores em localidades vizinhas, onde não possuem informações. Esta técnica é conhecida como aprendizagem de máquina (GUO, KELLY, GRAHAM, 2005). Ela atualmente é utilizada em diversas aplicações, como por exemplo, reconhecimento facial, onde é armazenado um grande conjunto de parâmetros e medidas do rosto de uma pessoa, como a distância entre os dois olhos, o comprimento do nariz, da boca e do queixo. Com estas informações o algoritmo é capaz de reconhecer se o rosto que está diante da câmera é ou não o que foi cadastrado previamente.

Para se obter o diagnóstico da atmosfera, são instaladas estações meteorológicas com instrumentos que medem as variáveis que caracterizam o estado e a dinâmica da nossa atmosfera terrestre, tal como, a temperatura, a pressão, a umidade, o vento, etc. No entanto, devido ao grande custo de instalação e manutenção de estações meteorológicas, elas são bem escassas e normalmente são instaladas em locais mais acessíveis, principalmente as estações convencionais, que necessitam de observadores para coletar os dados regularmente. Desta forma, o uso da aprendizagem de máquina pode ser empregado para amenizar o problema da falta de estações para se obter a estimativa de uma determinada variável em local onde não se tem medida, mas que provavelmente está sendo influenciado pelas localidades vizinhas onde se conhece o valor da respectiva variável.

O estudo da variável temperatura do ar é de suma importância devido a sua fundamental contribuição em diversas áreas do conhecimento, como climatologia, hidrologia, meteorologia, oceanografia e agricultura. A baixa temperatura pode causar nevascas ou geadas, provocando vários prejuízos econômicos e também morte de animais e até mesmo de pessoas. Por outro lado, as altas temperaturas também favorecem as queimadas, a desidratação e a mortalidade de animais, plantas e seres humanos.

Segundo o quinto relatório do Intergovernmental Panel on Climate Change (IPCC, 2014) o aumento da concentração de gases de efeito estufa na atmosfera está provocando a elevação da temperatura média do planeta que poderá causar grandes consequências tais

como: derretimento de grande parte das calotas polares; elevação do nível médio do mar; aumento de eventos extremos, como tornados, furacões, tempestades severas e secas prolongadas, incluindo desertificação de algumas regiões. Pesquisas recentes apontam que alguns destes efeitos já podem ser observados em certas regiões do planeta.

Na literatura, vários trabalhos fazem a predição da temperatura máxima do ar utilizando covariáveis (variáveis explicativas), via métodos tradicionais da estatística. Pode-se citar Dos Anjos Antonini et al. (2010), Ramos et al. (2011) e Gomes et al. (2014). Pelo fato de métodos paramétricos, como RLM, assumirem uma forma funcional para a função regressão, torna-os fáceis de ajustar, pois é necessário apenas estimar os coeficientes. No entanto, para que as propriedades desejáveis dos estimadores dos parâmetros desses métodos sejam atendidas é necessário assumir pressuposições que geralmente não são encontradas em conjunto de dados meteorológicos, por exemplo, que as covariáveis não devem ter multicolinearidade. Dessa forma, a utilização de técnicas de aprendizagem de máquina podem ser uma alternativa para trabalhar com dados dessa natureza.

A Figura 1 apresenta uma hierarquia de aprendizagem de máquina. Quando a variável resposta é quantitativa contínua, tem-se um problema de regressão e quando a variável resposta é qualitativa tem-se um problema de classificação.

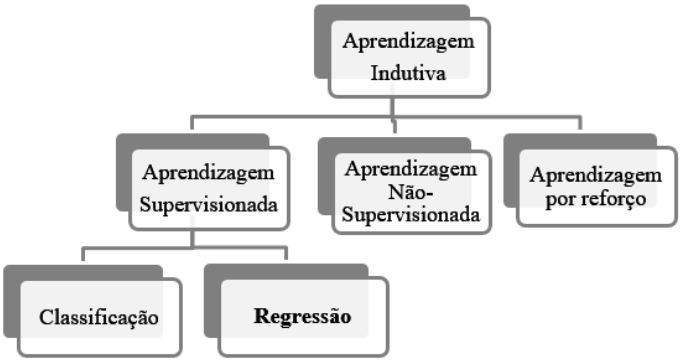


Figura 1 – Hierarquia de aprendizagem de máquina.
Fonte: Adaptada de Monard e Baranauskas (2003).

Os métodos paramétricos, como a regressão linear múltipla (RLM), assumem uma forma funcional linear para a função regressão $f(X)$. Isso os tornam fáceis de ajustar pois é necessário apenas estimar os coeficientes. Em aprendizagem de máquina supervisionado, os métodos não assumem explicitamente uma forma paramétrica para $f(X)$ e possuem

abordagens flexíveis para a realização de regressão. Em geral, os algoritmos de aprendizagem supervisionado são aplicados para aprender, a partir de dados amostrados, o mapeamento $Y = f(X)$ que representa um determinado fenômeno Y, no sentido de fazer classificações ou predições da variável resposta Y, a partir de um novo vetor de covariáveis X (JAMES et al., 2013).

Entre os algoritmos de aprendizagem de máquina para regressão, destacam-se Random Forests (BREIMAN, 2001; LIAW, WEINER, 2002) e Support Vector Machines (SMOLA, SCHÖLKOPF, 2004; VAPNIK, 2013).

O algoritmo Random Forest (RF), além de desempenhar a tarefa de classificação e regressão, estima a importância de uma variável avaliando o quanto a magnitude do erro de predição aumenta ao se permutar os dados daquela variável enquanto as demais permanecem inalteradas (BREIMAN, 2001).

Quanto ao Support Vector Machine (SVM), derivado da Teoria de Aprendizagem Estatística e Otimização Matemática, pode ser aplicado também a classificação e regressão. Segundo Hastie, Tibshirani, Friedman, (2009), essa é uma abordagem que demonstra ter bom desempenho em termos de generalizações nas bases de dados reais e com pequeno número de parâmetros a serem ajustados. No entanto, na literatura ainda existem poucos estudos acessíveis para a utilização desse método a problemas de regressão.

Nesse sentido, pretende-se como objetivo principal testar as capacidades dos algoritmos de regressão de aprendizado de máquina (RF e SVM) e comparar com o modelo tradicional de regressão linear múltipla (RLM) e o método clássico da interpolação Ponderada pelo Inverso da Distância (IDW), para inferir acerca da temperatura máxima do ar do estado de Minas Gerais, baseando-se nas informações de outras variáveis.

2 MATERIAL E MÉTODOS

Nas subseções que seguem, apresenta-se uma breve introdução dos fundamentos teóricos do modelo tradicional de Regressão Linear Múltipla (RLM) e da modelagem baseada nos algoritmos de aprendizagem de máquina supervisionado: Random Forest (RF) para regressão e Support Vector Machine para regressão com uso de funções lineares (SVM_Lin).

2.1 Modelos de Regressão

Em todo este estudo, denota-se por $X_1,\ X_2\ , \dots , X_p$ as p covariáveis distintas e supõe-se que Y é uma variável resposta que assume valores contínuos e numéricos.

Por uma amostra de n observações, $n > p + 1$, entende-se um conjunto de pares $(\boldsymbol{x_i}, y_i) \in \mathbb{R}^p \times \mathbb{R}$, em que o vetor p-dimensional $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots , x_{ip})$ têm componentes x_{ij} dada pelo valor da i-ésima observação da j-ésima covariável X_j e y_i a resposta correspondente observada de Y, $i = 1, 2, \cdots , n$ e $j = 1, 2, \cdots , p$.

Conforme descrito por James et al. (2013), dado um espaço de entrada $X = (X_1, X_2, \dots , X_n)$ e o espaço de saída Y, em aprendizagem de máquina supervisionado, os algoritmos de regressão são aplicados para encontrar um mapeamento f, porém desconhecida, que associa cada vetor de entrada $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots , x_{ip}) \in$ X a sua respectiva saída $y_i \in Y$.

Para análise dos modelos preditivos de regressão, uma técnica comum é separar, aleatoriamente, o conjunto de dados em dois subconjuntos disjuntos: um com aproximadamente 2/3 dos dados para ajuste dos modelos, denominado conjunto de treinamento e o outro para validação dos modelos, sendo denominado de conjunto de teste.

$$\overbrace{(\boldsymbol{x_1}, y_1\ ), (\boldsymbol{x_2}, y_2\ ), \dots , (\boldsymbol{x_m}, y_m)}^{\text{2/3 treinamento}}, \overbrace{(\boldsymbol{x_{m+1}}, y_{m+1}), \dots , (\boldsymbol{x_{m+n}}, y_{m+n})}^{\text{1/3 teste}} \quad (1)$$

De acordo com Izbicki (2018), pode ser de interesse investigar questões com objetivo inferencial e/ou preditivo, tais como: (i) quais são as covariáveis mais importantes, entre um conjunto possíveis de variáveis, que explicam a variável resposta Y? (ii) a relação entre Y e uma dada covariável é diretamente ou inversamente proporcional ou depende dos valores das outras covariáveis? (iii) a relação entre Y e cada covariável é linear ou não-linear? (iv) Como construir um estimador $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ de f, tal que $\hat{y}_1 = \hat{f}(\boldsymbol{x_1}) \approx y_1, \dots , \hat{y}_n = \hat{f}(\boldsymbol{x_n}) \approx y_n$, que

tenha um bom poder preditivo? ou seja, dadas novas observações $(\boldsymbol{x_{m+1}}, y_{m+1}), \dots, (\boldsymbol{x_{m+n}}, y_{m+n})$ tem-se que $\hat{y}_{m+1} = \hat{f}(\boldsymbol{x_{m+1}}) \approx y_{m+1}, \dots, \hat{y}_{m+n} = \hat{f}(\boldsymbol{x_{m+n}}) \approx y_{m+n}$.

Uma maneira de medir a qualidade de um preditor $f: \mathbb{R}^p \to \mathbb{R}$ é através da função erro ou risco esperado de f, denotada por $R(f)$, definida como a esperança de uma função positiva $L(\boldsymbol{x}, y, f(\boldsymbol{x}))$, denominada função de perda ou função de custo. Essa define a perda da aproximação do valor verdadeiro y pelo valor obtido pelo preditor $f(\boldsymbol{x})$ e, quanto menor o risco esperado, melhor é o preditor f (SMOLA, SCHÖLKOPF, 2004). Assim, encontra-se o melhor preditor f para representar os dados, minimizando $R(f)$.

No entanto, não é possível minimizar $R(f)$ diretamente, já que a distribuição de probabilidade dos dados geralmente não é conhecida e, usualmente, dispõe-se apenas de informações de uma amostra dos dados.

Segundo Faceli et al. (2011), em aprendizagem de máquina supervisionado, a estratégia adotada é utilizar o princípio da indução para inferir uma função $\hat{f}$ que minimize o erro sobre os dados de treinamento e que leve também a um menor erro de generalização (erro sobre novos dados).

De acordo com Izbicki (2018), qualquer que seja a função de perda L utilizada para definir o risco esperado têm-se, pela lei dos grandes números, que se n é suficientemente grande

$$\frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{x_{m+i}}, y_{m+i}, f(\boldsymbol{x_{m+i}})) \approx E\left[L(\boldsymbol{x}, y, f(\boldsymbol{x}))\right] = R(f) \tag{2}$$

Assim, um estimador para $R(f)$, obtido a partir de dados, que mede o desempenho de f, é o risco empírico $R_{emp}(f)$, definido por

$$R_{emp}(f) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x_i}, y_i, f(\boldsymbol{x_i})) \tag{3}$$

em que $T = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_m}, y_m)\}$ é uma amostra do conjunto de observações e $L(\boldsymbol{x_i}, y_i, f(\boldsymbol{x_i}))$ é uma função de perda, com $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$ e y_i o valor conhecido de resposta correspondente, para cada $i = 1, 2, \cdots, m$.

Segundo Schölkopf e Smola (2002), as funções de perda quadrática, perda absoluta e perda ε-insensível, usualmente são as mais utilizadas na literatura. São definidas, respectivamente, por

$$L(\boldsymbol{x}, y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2 \tag{4}$$

$$L(\boldsymbol{x}, y, f(\boldsymbol{x})) = |y - f(\boldsymbol{x})| \tag{5}$$

$$L(\boldsymbol{x}, y, f(\boldsymbol{x})) = \max_{(\boldsymbol{x},y)\in T}\{0, |y - f(\boldsymbol{x})| - \varepsilon\} \tag{6}$$

ou equivalentemente, para $\boldsymbol{r} = y - f(\boldsymbol{x})$,

$$\tilde{L}_{\varepsilon}(r) = \max_{(\boldsymbol{x},y)\in T}\{0, |r| - \varepsilon\} = \begin{cases} 0, & \text{se} \quad |r| \leq \varepsilon \\ |r| - \varepsilon & \text{se} \quad |r| > \varepsilon \end{cases}, \tag{7}$$

em que a distância ε é escolhida a priori. A ideia por trás da função de perda ε-insensível é que os desvios até ε, não devem ser penalizados e todos os demais desvios devem incidir apenas em uma penalidade linear. Utilizou-se essa função de perda na teoria do algoritmo de Support Vector Machine estabelecido por Vapnik (2013).

Já, ao utilizar a função de perda quadrática (Equação (4)), o risco empírico de f é o MSE de um estimador $\hat{f}$, ou seja,

$$R_{emp}(f) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2 \approx \mathrm{E}[(Y - f(X))^2] = R(f) \tag{8}$$

e, para a função de perda absoluta (Equação (5)), o risco empírico de f é o MAE de um estimador $\hat{f}$.

Segundo Faraway (1995), o MSE de um estimador $\hat{f}$ têm uma alta probabilidade de adequação aos dados (overfitting) e a técnica de separar o conjunto original nos conjuntos de treinamento e validação, ajuda a reduzir o overfitting. Assim, usa-se o conjunto de treinamento para estimar f e os dados de teste para estimar $R(f)$.

Outra forma de estimar $R(f)$ é através de uma medida $\mathcal{P}(f)$ de penalização para f que serve para corrigir a diferença entre $R_{emp}(f)$ e $R(f)$. Essa forma é utilizada na teoria do algoritmo de SVM para regressão (SCHÖLKOPF, SMOLA, 2002).

2.1.1 Regressão Linear Múltipla

A regressão linear múltipla (RLM) é um método de aprendizado estatístico útil e amplamente utilizado para prever uma variável resposta quantitativa. De acordo com James et al. (2013), muitas abordagens sofisticadas de aprendizado estatístico podem ser vistas como generalizações ou extensões de regressão linear.

Ainda de acordo com James et al. (2013), o modelo de RLM assume a forma:

$$\boldsymbol{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \boldsymbol{\varepsilon} \tag{9}$$

em que os parâmetros β_j, $j = 0, 1, 2, \cdots, p$, são denominados coeficientes de regressão e $\boldsymbol{\varepsilon}$ um vetor de erros não explicado pelo modelo

O modelo de regressão amostral correspondente a Equação (9) torna-se

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \cdots, n \tag{10}$$

em que $E\left(Y \middle| X_1 = x_{i1},\ X_2 = x_{i2},\ \ldots, X_p = x_{ip}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ é o valor esperado de Y dado que $X_1 = x_{i1},\ X_2 = x_{i2},\ \ldots, X_p = x_{ip}$.

Então, o parâmetro β_0, também denominado de intercepto, pode ser interpretado como o valor esperado em Y quando todas as covariáveis são nulas, ou seja, $E\left(Y \middle| X_1 = 0,\ X_2 = 0,\ \ldots, X_p = 0\right) = \beta_0$. Já o β_j, $j = 1, 2, \cdots, p$, além de quantificar a associação entre a variável Y e X_j, representa a variação esperada na resposta Y por uma unidade de mudança em X_j, mantendo-se constante as demais covariáveis, pois $\frac{\partial Y}{\partial X_j} = \beta_j$.

A Equação (10) consiste em um sistema de n equações da forma

$$\begin{cases} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \tag{11}$$

que, na forma matricial, é expressa por

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\boldsymbol{y}_{n\times 1}} = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}}_{X_{n\times(p+1)}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}_{(p+1)\times 1}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}_{n\times 1}} \tag{12}$$

ou seja,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{13}$$

em que $\boldsymbol{y}$ é um vetor $n \times 1$ das observações aleatórias da variável resposta Y, X é uma matriz $n \times (p + 1)$ dos valores das covariáveis, $\boldsymbol{\beta}$ é um vetor $(p + 1) \times 1$ dos coeficientes de regressão e $\boldsymbol{\varepsilon}$ é um vetor $n \times 1$ de erros aleatórios.

O vetor dos parâmetros $\boldsymbol{\beta}$ e a variância do erro σ^2 são desconhecidos e devem ser estimados a partir de dados de amostra. O método mais utilizado para obter um estimador de $\boldsymbol{\beta}$ da RLM é o dos mínimos quadrados ordinários (MQO). Ele consiste em adotar como vetor

de estimadores de mínimos quadrados, $\widehat{\boldsymbol{\beta}}$, o que minimiza a soma dos quadrados dos desvios (Z), dada por:

$$Z = \sum_{i=1}^{n} {\varepsilon_i}^2 = \boldsymbol{\varepsilon^t \varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})^t(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t X^t\boldsymbol{y} + \boldsymbol{\beta}^t X^t X\boldsymbol{\beta} \tag{14}$$

Note que, por Z ser uma soma de quadrados, a função Z tem mínimo que ocorre quando suas derivadas parciais em relação a $\boldsymbol{\beta}$ forem nulas. Então

$$\left.\frac{\partial Z}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^t\,\boldsymbol{y} + \mathbf{2}\,\boldsymbol{X}^t\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = 0 \tag{15}$$

ou seja,

$$\boldsymbol{X}^t\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{y} \tag{16}$$

A Equação (16) é denominada sistema de equações normais dos mínimos quadrados (SEN).

Note que, a matriz $\boldsymbol{X}^t\boldsymbol{X}$ é simétrica e seus elementos da diagonal principal são somas de quadrados dos elementos das colunas de **X** e, fora da diagonal, são somas dos produtos cruzados dos elementos das colunas de **X**.

Então, o SEN é um sistema de $p + 1$ equações, uma para cada um dos coeficientes de regressão, cuja solução são os estimadores de mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Em forma matricial, o SEN torna-se

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} {x_{i1}}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} {x_{ip}}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{bmatrix} \tag{17}$$

Então, se existe a matriz inversa $(\boldsymbol{X}^t\boldsymbol{X})^{-1}$, tem-se que o vetor

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \tag{18}$$

é o estimador de MQO do vetor $\boldsymbol{\beta}$ que minimiza a função soma de quadrados dos desvios (Equação (14)). Após obter as estimativas $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ faz-se previsões usando o modelo de regressão ajustado:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \tag{19}$$

ou, equivalentemente,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \tag{20}$$

A diferença entre o valor observado y_i e o valor ajustado correspondente $\hat{y}_i$, é o resíduo $e_i = y_i - \hat{y}_i$, $i = 1, 2, \cdots, n$. Os n resíduos podem ser escritos em notação matricial

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{21}$$

Para que as propriedades desejáveis de um estimador sejam atendidas, algumas pressuposições são necessárias para o modelo de RLM: (a) a variável resposta $\boldsymbol{y}$ deve ser função linear das covariáveis $X_1,\ X_2, \ldots, X_p$; (b) os valores de $X_1,\ X_2, \ldots, X_p$ devem ser fixos, ou seja, as covariáveis não devem ser variáveis aleatórias; (c) a média dos erros deve ser nula, ou seja, $E(\varepsilon_i) = 0,\ \forall\, i = 1, 2, \cdots, n$; (d) a variância dos erros deve ser sempre σ^2, ou seja, $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2,\ \forall\, i = 1, 2, \cdots, n$; (e) o erro de uma observação deve ser independente do erro em outra observação, ou seja, $cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0,\ \forall\, i \neq j$; (f) o número de observações n deve ser maior que o número de parâmetros a serem estimados, ou seja, $n > p + 1$; (g) os erros devem ter distribuição normal, ou seja, $\varepsilon_i \sim N(0, \sigma^2),\ \forall\, i$; (h) as covariáveis $X_1,\ X_2,\ \ldots,\ X_p$ devem ser linearmente independentes, ou seja, não devem ter multicolinearidade entre as covariáveis.

A pressuposição (g) é equivalente a assumir que a distribuição condicional de Y dado $X_1,\ X_2, \ldots, X_p$ é normal com média $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ e variância σ^2. Ela é necessária para testar hipóteses ou construir intervalos de confiança para os parâmetros.

A pressuposição (h) garante a existência da matriz $(X^t X)^{-1}$, já que nenhuma coluna da matriz X é uma combinação linear das outras colunas.

Das pressuposições d) e e) resultam que

$$\begin{aligned} E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{t}}) &= E\left(\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}[\varepsilon_1 \quad \varepsilon_2 \quad \ldots \quad \varepsilon_n]\right) = E\begin{pmatrix}{\varepsilon_1}^2 & \varepsilon_1\varepsilon_2 & \ldots & \varepsilon_1\varepsilon_n\\ \varepsilon_2\varepsilon_1 & {\varepsilon_2}^2 & \ldots & \varepsilon_2\varepsilon_n\\ \vdots & \vdots & \ldots & \vdots\\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \ldots & {\varepsilon_n}^2\end{pmatrix} \\ &= \begin{pmatrix}E({\varepsilon_1}^2) & E(\varepsilon_1\varepsilon_2) & \ldots & E(\varepsilon_1\varepsilon_n)\\ E(\varepsilon_2\varepsilon_1) & E({\varepsilon_2}^2) & \ldots & E(\varepsilon_2\varepsilon_n)\\ \vdots & \vdots & \ldots & \vdots\\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \ldots & E({\varepsilon_n}^2)\end{pmatrix} = \begin{pmatrix}\sigma^2 & 0 & \ldots & 0\\ 0 & \sigma^2 & \ldots & 0\\ \vdots & \vdots & \ldots & \vdots\\ 0 & 0 & \ldots & \sigma^2\end{pmatrix} = \sigma^2\boldsymbol{I}\end{aligned} \tag{22}$$

em que $\boldsymbol{I}$ a matriz identidade de ordem n.

Então, utilizando que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon} \tag{23}$$

e a Equação (22), resulta que

$$E(\widehat{\boldsymbol{\beta}}) = E[\boldsymbol{\beta} + (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon}] = \boldsymbol{\beta} + (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} \tag{24}$$

ou seja, $\widehat{\boldsymbol{\beta}}$ é um estimador não viesado de $\boldsymbol{\beta}$.

Em relação as variâncias e covariâncias dos estimadores de MQO, tem-se que

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}) &= E\left\{\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]^t\right\} = E\left\{\left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]\left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]^t\right\} \\ &= E\{(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}\,. \end{aligned} \tag{25}$$

Portanto, a variância dos estimadores dos parâmetros estão na diagonal principal da matriz $\sigma^2(X^tX)^{-1}$ e, fora da diagonal, as covariâncias.

O Teorema de Gauss-Markov assegura que, se as pressuposições (a) até (e) forem atendidas, o estimador $\widehat{\boldsymbol{\beta}}$ de MQO satisfaz a propriedade de melhor estimador linear não viesado e de variância mínima (BLUE).

Um estimador não viesado de σ^2 é dado pelo quadrado médio do resíduo (QMRes), definido por

$$\hat{\sigma}^2 = QMRes = \frac{SQRes}{n-p-1} \tag{26}$$

em que $SQRes$ é a soma de quadrados dos resíduos dada por

$$\begin{aligned} SQRes &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^t\boldsymbol{e} = \left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} + \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y}^t\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} + \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} + \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} = \boldsymbol{y}^t\boldsymbol{y} - \widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{y} \end{aligned} \tag{27}$$

Observe que a Equação (9), com uma única covariável, torna-se um modelo de regressão linear simples ($y = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$), em que a linha de regressão de mínimos quadrados é uma reta (Figura 2a) e, com duas covariáveis, $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{\varepsilon}$, a linha de regressão de mínimos quadrados se torna um plano (Figura 2b).

Então, procedimentos de regressão tradicionais são processos que originam uma função $f(\boldsymbol{x})$ que tem o menor desvio entre as respostas preditas e observadas experimentalmente.

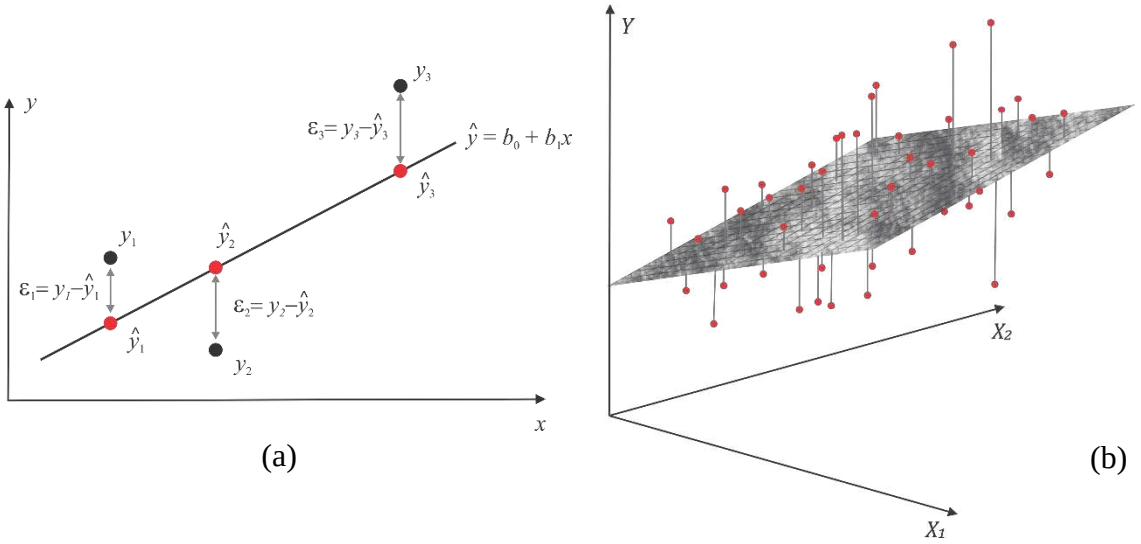


Figura 2 – Para cada observação (mostrada em vermelho), a curva de regressão de mínimos quadrados (a) no cenário bidimensional, com uma covariável e uma variável resposta, se torna uma reta. (b) no cenário tridimensional, com duas covariáveis e uma variável resposta, se torna um plano.
Fonte: Adaptada de James et al. (2013).

2.1.2 Random Forest

O algoritmo de Random Forest (RF) para regressão é uma técnica não paramétrica, que cria e combina desempenho de modelos de aprendizado de várias árvores de regressão para predizer o valor de uma variável resposta (BREIMAN, 2001; GUO, KELLY, GRAHAM, 2005; RODRIGUEZ-GALIANO et al., 2015).

Segundo James et al. (2013), o processo de construção de uma árvore de regressão consiste, através de um conjunto de regras, dividir recursivamente o espaço das covariáveis (conjunto de valores possíveis para X_1, X_2 , … , X_p) em J regiões distintas e não sobrepostas $\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_J$.

Dividir recursivamente significa particionar os dados em subconjuntos (ou ramos) menores e, cada um desses ramos menores, em ramos ainda menores e, assim sucessivamente, até um ponto de parada. Cada particionamento é denominado de nó e os nós terminais recebem o nome de folhas. A profundidade da árvore é medida pela quantidade de nós com ramos para outros nós da árvore

Assim, cada nó da árvore corresponde a um teste sobre uma única covariável. Para cada observação que cai na região R_j, faz-se a mesma previsão que é simplesmente a média

dos valores da variável resposta do conjunto de treinamento em R_j ou seja, para prever o valor de resposta de uma nova observação, verifica-se a região a qual a observação pertence e, então, atribui-se a média dos valores da variável resposta correspondente.

As regiões $R_1, R_2, \ldots, R_J$ são construídas de modo a minimizar a soma dos quadrados dos erros de previsão (SQR_{tree}), dada por

$$SQR_{tree} = \sum_{i:x_i \in R_1} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2} \left(y_i - \hat{y}_{R_2}\right)^2 + \cdots + \sum_{i:x_i \in R_J} \left(y_i - \hat{y}_{R_J}\right)^2 \quad (28)$$

em que $\hat{y}_{R_j}$ é a resposta média para as observações de treinamento dentro da região R_j, $j = 1, 2, \cdots, J$.

A Figura 3 ilustra uma partição do espaço de recurso bidimensional e sua árvore de regressão correspondente. A árvore possui quatro nós de decisão e cinco nós terminais ou folhas denotados por R_1, R_2, R_3, R_4 e R_5. Note que, cada nó interior da árvore corresponde a um teste sobre uma covariável. Para os nós restrito às condições $X_1 < t_1$ e $X_2 < t_2$, a observação é prevista pelo resultado da folha R_1 dada pela média dos valores da variável resposta das observações do conjunto de treinamento dentro da região R_1.

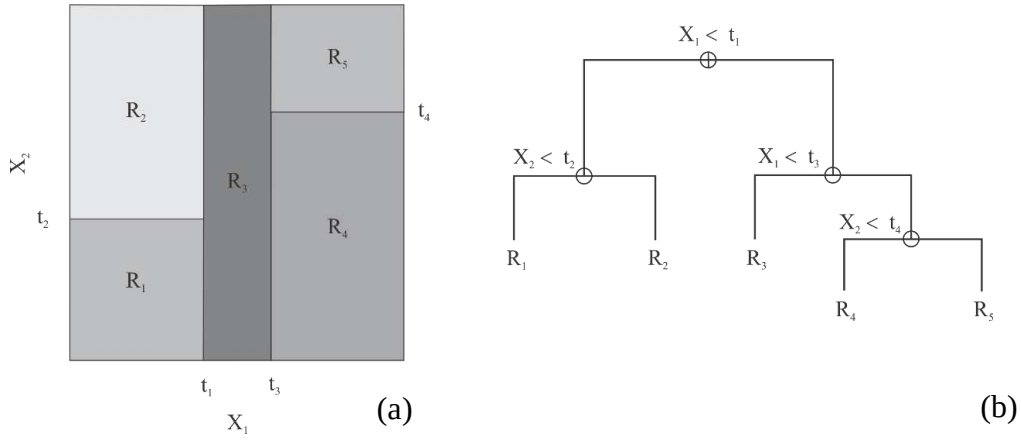


Figura 3 – (a) Uma partição do espaço de recurso bidimensional com uma divisão binária recursiva; (b) Uma árvore de regressão correspondente à divisão binária recursiva com quatro nós e cinco folhas.
Fonte: Adaptada de James et al. (2013).

No entanto, segundo James et al. (2013), considerar todas as J partições possíveis do espaço de covariáveis de modo a minimizar a soma dos quadrados dos erros, Equação (28), é computacionalmente inviável e, assim, adota-se uma abordagem denominada divisão binária

recursiva do espaço de covariáveis. Isto é, em cada nó, identifique a covariável X_j, dentre todos as covariáveis $X_1,\ X_2\ , \ldots , X_p$, e o ponto de corte t_s, dentre todos os possíveis valores das covariáveis, que divide o espaço de covariáveis em dois semi-planos $R_1(j,s) = \{X\ ;\ X_j < t_s\}$ e $R_2(j,s) = \{X\ ;\ X_j \geq t_s\}$ que minimize a soma do quadrado dos erros de predição naquele nó, isto é,

$$\min_{j,s}\left[\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2\right] \tag{29}$$

em que $\hat{y}_{R_1}$ é a resposta média para as observações de treinamento em $\mathrm{R}_1(j,s)$ e $\hat{y}_{R_2}$ é a resposta média para as observações de treinamento em $\mathrm{R}_2(j,s)$.

Uma vez estabelecida as regiões $R_1(j,s)$ e $R_2(j,s)$, o nó inicial da árvore é fixado. Em seguida, aplica-se novamente esse processo e busca-se a melhor covariável e o melhor ponto de corte que particiona $R_1(j,s)$ ou $R_2(j,s)$ em regiões menores, com menor erro quadrático. O processo de divisão de dados em cada nó interno de uma regra da árvore continua recursivamente e resulta em árvores adultas até que todos os pontos sejam cobertos ou que uma condição de parada especificada anteriormente seja atingida. Para cada um dos nós terminais (ou folhas), é anexado um modelo de regressão simples que se aplica apenas nesse nó.

Modelo baseado em árvore enfrenta também problemas de erros de variância e viés. Variância significa o quão diferente são as previsões do modelo, num mesmo ponto, se diferentes conjuntos de treinamento forem tomados da mesma população. Viés é o quanto, em média, os valores previstos são diferentes dos valores reais.

Ainda, segundo James et al. (2013), dois dos parâmetros que influenciam o viés são a profundidade da árvore e a quantidade de atributos que podem ser usados para medir a pureza do conjunto. Normalmente, árvore pequena diminui a complexidade do modelo, mas há um aumento do viés e, à medida que a árvore cresce o viés diminui, mas sofrerá com aumento da variância a ponto de causar o sobreajuste (overfitting) nos conjuntos de treinamentos, isto é, produz boa precisão nos conjuntos de treinamentos, mas uma performance preditiva ruim em novas observações. Optar por modelos com viés mais alto, mas variância baixa pode causar o underfitting.

Uma opção para lidar com sobreajuste é fixar o número mínimo de observações necessárias para dividir um nó; definir o número mínimo de observações necessárias em um nó terminal ou folha; definir o número máximo de folhas. Outra alternativa, é podar a árvore,

isto é, tornar a árvore de regressão menor, de modo a diminuir a variação no conjunto de teste e assim melhorar a capacidade de generalização da árvore. Para podar a árvore de regressão, retira-se cada nó e observa-se como o erro de predição varia no conjunto de teste (MSE_{teste}). Com base nisso, decide-se quais os nós irão permanecer na árvore. O número de casos nos nós pode ser considerado como critério de remoção. (JAMES et al., 2013).

Apesar de árvores de regressão serem de fácil interpretação, segundo James et al. (2013), é uma técnica de aprendizagem fraca, pois normalmente não possuem bom desempenho preditivo quando aplicados a novas observações. A causa disso poderá ser devido à metodologia de estimação, ou problema de overfitting ou, ainda, por ser sensível a mudanças nos conjuntos de treinamento.

Segundo o princípio de Breiman (2001), combinar várias técnicas de aprendizagem de baixa precisão pode gerar um modelo de aprendizagem altamente preciso.

Random Forest (RF) é considerada uma técnica de aprendizagem de máquina forte, com bom desempenho preditivo quando aplicado à observações não pertencentes aos conjuntos de treinamento, o que reduz o erro de generalização. Ela consiste em estimar K árvores de regressão, sem poda, utilizando K novos conjuntos de treinamentos distintos, de mesmo tamanho da amostra original, por meio de um sorteio aleatório com reposição (reamostragem bootstrap). Essa técnica é denominada de método bagging (BREIMAN, 1996).

Depois que K dessas árvores $\{T(\boldsymbol{x})\}_1^K$ são cultivadas, o preditor de regressão RF, segundo Breiman (2001), é dado por:

$$\hat{f}_{rf}^{k}(x) = \frac{1}{K}\sum_{k=1}^{K} T(x) \tag{30}$$

isto é, $\hat{f}_{rf}^{k}(x)$ é a média das previsões de cada árvore de regressão.

Na construção de cada árvore utiliza-se m covariáveis, escolhidas aleatoriamente, para a divisão em cada nó. As árvores são cultivadas até o tamanho máximo, sem poda. Os números K árvores e m covariáveis são dois parâmetros definidos pelo usuário, necessários para criar uma RF (BREIMAN, 2001).

Segundo Liaw e Weiner (2002), no processo bagging, alguns dados podem ser usados mais de uma vez nos K conjuntos de treinamento, enquanto outros podem nunca serem usados. As amostras que não são selecionadas para o treinamento na k-ésima árvore do processo bagging são incluídas como parte de outro subconjunto chamado amostras out-of-

bag (oob), usados pela k-ésima árvore para avaliar o desempenho. A partir das predições oob de todas as árvores da floresta, calcula-se o erro quadrático médio com as amostras oob (MSE_{oob}), por:

$$MSE_{oob} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{\,oob}\right)^2 \tag{31}$$

em que y_i é o i-ésimo valor observado da variável e $\hat{y}_i^{\,oob}$ é a média das predições oob para a i-ésima observação.

Já a porcentagem de variância explicada pelo modelo ($Var_{exp} = R^2$), em que utiliza as amostras oob, segundo Liaw e Weiner (2002), é dada por

$$Var_{exp} = 1 - \frac{MSE_{oob}}{\hat{\sigma}_y^2}. \tag{32}$$

O MSE_{oob} também fornece uma indicação da importância de uma covariável. Se uma determinada covariável for relevante, a permutação aleatória dela por outra deve fornecer muito efeito sobre a estimativa de MSE_{oob} calculada antes e após a permutação aleatória dessa específica covariável. Assim, um aumento no MSE_{oob} é proporcional à importância da variável preditiva. Assim, uma das vantagens de RF é poder manipular com conjuntos de dados extensos e com grande número de covariáveis, com possibilidade de identificar quais covariáveis são mais importantes para predizer o valor da variável resposta. A técnica de RF estima a importância de uma variável preditiva observando o quanto o MSE_{oob} aumenta quando os dados oob para essa variável são permutados, enquanto todas as outras variáveis são deixadas inalteradas. O aumento no MSE_{oob} é proporcional à importância da variável preditiva (LIAW e WEINER, 2002).

2.1.3 Support Vector Machine

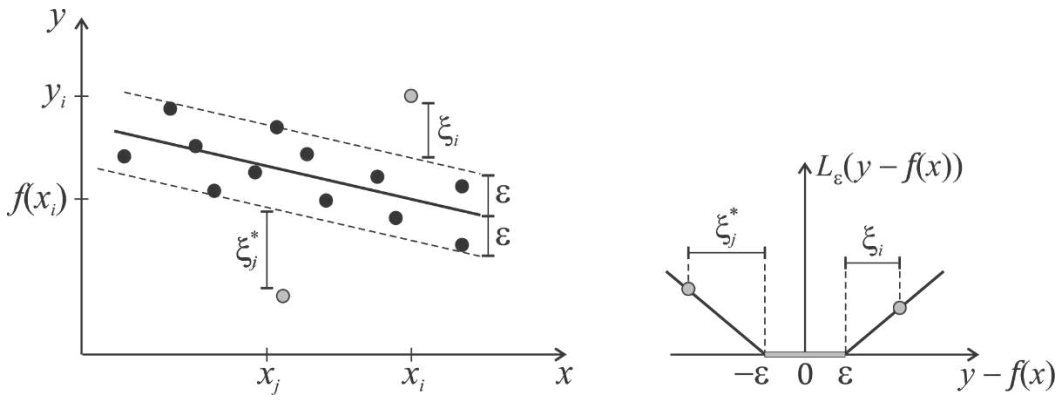
Segundo Vapnik (2013), as Support Vectors Machine (SVMs) são técnicas de aprendizado de máquina que deriva da Inteligência Artificial e foi inicialmente desenvolvida como um método de classificação linear, posteriormente generalizado para um classificado não linear e, por fim, expandido para problemas de regressão linear e não-linear.

Estabelecido por Vapnik (2013), o algoritmo Support Vectors Machine para regressão, denominado ε−Support Vector Regression (ε−SVR), têm como objetivo encontrar uma função contínua f que têm no máximo ε desvio para as respostas conhecidas y_i, $\forall i =$

$1, 2, \cdots, m$, e, ao mesmo tempo, seja a mais uniforme e regular possível (Figura 4a). E, para explicar o erro cometido por f e avaliar seu desempenho, utiliza-se a função de perda ε-insensível, definida por

$$L_\varepsilon(y - f(\boldsymbol{x})) = \max_{(\boldsymbol{x},y)\in T}\{0, |y - f(\boldsymbol{x})| - \varepsilon\} = \begin{cases} 0, & \text{se} \quad |y - f(\boldsymbol{x})| \leq \varepsilon \\ |y - f(\boldsymbol{x})| - \varepsilon & \text{se} \quad |y - f(\boldsymbol{x})| > \varepsilon \end{cases} \tag{33}$$

que ignora os erros associados a pontos do conjunto de treinamento que estão dentro de uma certa distância $\varepsilon > 0$ da função verdadeira (Figura 4b).



(a) (b)

Figura 4 – (a) Imagem do ε-tubo em torno de um preditor linear $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$. As variáveis ξ_i e ξ_i^* medem o custo de erros de treinamento correspondente a pontos fora do ε-tubo; (b) função de perda ε-insensível.
Fonte: Adaptada de Smola e Schölkopf (2004).

Nesse estudo, a procura de f foi restrita somente a classe de funções lineares, ou seja, funções pertencentes ao conjunto

$$\{f: X \subset \mathbb{R}^p \longrightarrow \mathbb{R}\,;\; f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b, \;\; \text{com} \;\; \boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^p, b \in \mathbb{R}\} \tag{34}$$

em que $\boldsymbol{x}$ é um vetor p-dimensional contendo dados de entrada, $\boldsymbol{w}$ é o vetor de pesos p-dimensional, de norma mínima, normal ao hiperplano de Equação $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, b é um escalar denominado bias e $\boldsymbol{w} \cdot \boldsymbol{x}$ é o produto interno entre os vetores $\boldsymbol{w}$ e $\boldsymbol{x}$. Para a regularidade, deve-se encontrar função com pequeno $\boldsymbol{w}$, isto é, de norma mínima. Assim, segundo Smola e Schölkopf (2004), busca-se encontrar um função linear f que aproxime os pontos de treinamento $(\boldsymbol{x_i}, y_i)$, com uma precisão de ε, ou seja, encontrar f reduz ao seguinte problema de otimização quadrática:

$$\underset{\boldsymbol{w},b}{\text{Minimizar}} \;\; \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{35}$$

$$\text{sujeito a } \begin{cases} y_i - f(\boldsymbol{x_i}) \leq \varepsilon \\ f(\boldsymbol{x_i}) - y_i \leq \varepsilon \end{cases}, \quad \forall\, i = 1, \cdots, m$$

equivalentemente,

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{Minimizar}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{sujeito a } \begin{cases} y_i - \boldsymbol{w} \cdot \boldsymbol{x_i} - b \leq \varepsilon \\ \boldsymbol{w} \cdot \boldsymbol{x_i} + b - y_i \leq \varepsilon \end{cases}, \quad \forall\, i = 1, \cdots, m\,. \end{aligned} \tag{36}$$

Dessa forma, o problema de aprendizagem de máquina utilizando o algoritmo de SVM_Lin reduz a uma formulação que pode ser analisada a partir da teoria de otimização de função quadrática denominada otimização convexa, (DO, 2009).

Assim, ao considerar uma distância $\varepsilon \geq 0$, deve-se escolher um função linear f em que todos os pontos de treinamento $(\boldsymbol{x_i}, y_i),\ i = 1, \cdots, m$, se encontram posicionados em torno do preditor f, dentro de uma região denominada ε-tubo, o mais delgado possível, ou seja, que $|y_i - f(\boldsymbol{x_i})| \leq \varepsilon$, $\forall\, i = 1, \cdots, m$ (Figura 4a).

Porém, se permitir que alguns pontos de treinamento estejam localizados fora desse ε-tubo, esses são vistos como erros de treinamento. Conforme descrito por Vapnik (2013), para explicar esses erros introduziu-se variáveis de folgas ξ_i ou ξ_i^*, relacionadas a cada ponto do conjunto de treinamento. As variáveis ξ_i ou ξ_i^* assumem valor zero se o ponto de treinamento estiver dentro do ε-tubo. Se o ponto estiver "acima" do ε-tubo, isto é, $y_i - f(\boldsymbol{x_i}) \geq 0$, ξ_i aumenta progressivamente para pontos fora do ε-tubo de acordo com a função de perda usada.

Por exemplo, utilizando a função perda $L_\varepsilon(y - f(\boldsymbol{x}))$ (Equação (33)) tem-se que, para pontos de treinamentos dentro do ε-tubo não há perda e, para os pontos fora do ε-tubo, a perda aumenta progressivamente, conforme a distancia ξ_i (relacionados com os pontos tais que $y_i - f(\boldsymbol{x_i}) \geq 0$) ou ξ_i^* (relacionados com os pontos tais que $y_i - f(\boldsymbol{x_i}) \leq 0$) (Figura 4b). Dessa forma, queremos encontrar $\boldsymbol{w} \in \mathbb{R}^p$ e $b \in \mathbb{R}$ que definirá o preditor linear $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$, de modo que a soma dos erros de treinamento $\sum_{i=1}^{m}(\xi_i + \xi_i^*)$ seja mínima, além de $\frac{1}{2}\|\boldsymbol{w}\|^2$. Acrescentando as variáveis de folga ξ_i ou ξ_i^* nas restrições e penalizando-as na função objetivo, tem-se uma nova formulação equivalente do problema de otimização da Equação (36):

$$\underset{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi^*}}{\text{Minimizar}} \ \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i:y_i-(\boldsymbol{w}\cdot\boldsymbol{x_i}+\mathrm{b})\geq 0} \xi_i \ + \sum_{i:y_i-(\boldsymbol{w}\cdot\boldsymbol{x_i}+\mathrm{b})\leq 0} \xi_i^* \right) \right\} \tag{37}$$

$$\text{sujeito a } \begin{cases} y_i - f(\boldsymbol{x_i}) \leq \varepsilon + \xi_i \\ f(\boldsymbol{x_i}) - y_i \leq \varepsilon + \xi_i^*, \ \forall\, i = 1, \cdots, m \\ \xi_i,\ \xi_i^* \geq 0 \end{cases}$$

em que $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$, $\boldsymbol{\xi}^* = (\xi_1^*, \xi_2^*, \ldots, \xi_m^*)$ e a constante $C = \frac{c}{m}$, escolhida a priori, impõe uma relação entre a regularidade de f e quanto de erros são permitidos. Se C for alto, mais ênfase será dada ao erro com uma alta probabilidade de adequação aos conjuntos de treinamento (overfitting). Nesse caso, o preditor f pode ter baixo desempenho de generalização. Enquanto, se C for pequeno, mais ênfase será dada à norma dos pesos e as predições podem ser ruins devido à falta de flexibilidade (underfitting).

O problema de otimização com restrições de desigualdade, Equação (37), é um problema quadrático com função objetivo a ser minimizada convexa. Esse problema possui um único mínimo global (BOYD, VANDENBERGHE, 2009) e para resolvê-lo é necessário encontrar o ponto de sela da função Lagrangiana primal

$$\begin{aligned} &\mathcal{L}(\mathbf{w}, \mathrm{b}, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*) \\ &= \frac{1}{2}\|\boldsymbol{w}\|^2 \ + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}\alpha_i[\boldsymbol{f}(\boldsymbol{x_i}) - y_i + \varepsilon + \xi_i] \\ &- \sum_{i=1}^{m}\alpha_i^*[y_i - f(\boldsymbol{x_i}) + \varepsilon + \xi_i^*] - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{38}$$

em que $\alpha_i \geq 0$, $\alpha_i^* \geq 0$, $\eta_i \geq 0$, $\eta_i^* \geq 0$, $\forall\, i = 1, \cdots, m$, são multiplicadores de Lagrange.

O mínimo é tomado em relação as variáveis $\boldsymbol{w}$, b, ξ_i, ξ_i^* e o máximo em relação aos multiplicadores de Lagrange α_i, α_i^*, η_i, η_i^* (DO, 2009).

Resolver o problema de minimização envolve tomar as derivadas parciais de $\mathcal{L}$ em relação a $\boldsymbol{w}$, b, ξ_i e ξ_i^* o que resulta em:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\boldsymbol{x_i} = 0 \tag{39}$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0 \ \Leftrightarrow \ \sum_{i=1}^{m}\alpha_i = \sum_{i=1}^{m}\alpha_i^* \tag{40}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \quad \text{e} \quad \frac{\partial \mathcal{L}}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0, \quad \forall\, i = 1, \cdots, m\,. \tag{41}$$

Substituindo as Equações (39) (40) (41) em (38) resulta a função denominada Lagrangeana dual de Wolfe, que depende apenas dos multiplicadores de Lagrange α_i, α_i^* e do conjunto de treinamento:

$$\mathcal{L}(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)\left(\alpha_j-\alpha_j^*\right)\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right) - \varepsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i-\alpha_i^*) \qquad (42)$$

O problema de otimização agora é denominado problema dual de Wolfe:

$$\underset{\boldsymbol{\alpha},\boldsymbol{\alpha}^*}{\text{Maximizar}}\left\{-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)\left(\alpha_j-\alpha_j^*\right)\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right) - \varepsilon\sum_{j=1}^{m}(\alpha_i+\alpha_i^*) + \sum_{j=1}^{m} y_i(\alpha_i-\alpha_i^*)\right\}$$

$$\text{sujeito a}\begin{cases}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)=0 \\ \quad 0\le\alpha_i\le C \\ \quad 0\le\alpha_i^*\le C, \qquad \forall\, i=1,\cdots,m\,.\end{cases} \qquad (43)$$

Deve-se encontrar N pares de multiplicadores (α_i,α_i^*). Assim, se $\widehat{\boldsymbol{\alpha}} = (\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_m)$ e $\boldsymbol{\alpha}^* = (\hat{\alpha}_1^*,\hat{\alpha}_2^*,\ldots,\hat{\alpha}_m^*)$ são estimadores maximizando o problema dual da Equação (43), utiliza-se a Equação (39) para encontrar um estimador de $\boldsymbol{w}$,

$$\widehat{\boldsymbol{w}} = \sum_{i=1}^{m}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)\boldsymbol{x_i} \quad \text{e, portanto,} \quad \hat{f}(\boldsymbol{x}) = \sum_{i=1}^{m}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)(\boldsymbol{x_i}\cdot\boldsymbol{x}) + \hat{b}. \qquad (44)$$

Porém, para que o método dos multiplicadores de Lagrange com restrições de desigualdade forneça uma solução ótima, segundo a teoria de otimização com restrições, as condições de Karush-Kuhn-Tucker (KKT) devem ser atendidas (BOYD, VANDENBERGHE, 2009).

Conforme descrito por Smola e Scholkopf (2004), as condições afirmam que, no ponto ótimo, o produto entre variáveis duais e as restrições deve ser nulo. Para o estudo em questão, significa que

$$\alpha_i(\boldsymbol{w}\cdot\boldsymbol{x_i} + \mathrm{b} - y_i + \varepsilon + \xi_i) = 0 \quad \forall\, i=1,\cdots,m \qquad (45)$$

$$\alpha_i^*(y_i - \boldsymbol{w}\cdot\boldsymbol{x_i} - \mathrm{b} + \varepsilon + \xi_i^*) = 0 \quad \forall\, i=1,\cdots,m \qquad (46)$$

$$(C-\alpha_i)\xi_i = 0 \qquad (47)$$

$$(C-\alpha_i^*)\xi_i^* = 0\,. \qquad (48)$$

Das Equações (47) e (48) tem-se que as variáveis de folga ξ_i ou ξ_i^* são não-nulas e correspondem as observações fora do ε-tubo somente quando $\alpha_i = C$ ou $\alpha_i^* = C$.

Além disso, como um ponto observado não pode estar simultaneamente em ambos os lados do ε-tubo, tem-se que se $\alpha_i > 0$ então $\alpha_i^* = 0$ ou se $\alpha_i^* > 0$ então $\alpha_i = 0$. Logo, o produto $\alpha_i\alpha_i^* = 0$. Para $\alpha_i = 0$ e $\alpha_i^* = 0$ indica que os pontos localizam dentro do ε-tubo e não interferem na construção do preditor. Dessa forma, os pontos de treinamento para os quais $0 < \alpha_i \le C$ (e, portanto $\alpha_i^* = 0$) ou $0 < \alpha_i^* \le C$ (e, portanto $\alpha_i = 0$) são os mais relevantes entre todos os dados de treinamento e esses são denominados vetores de suporte. Isso deu o

significado ao nome support vector machine, conforme descrito por Smola e Scholkopf (2004). Dessa forma, para uma escolha qualquer i em que $0 < \alpha_i \leq C$, pode-se determinar o viés b pela Equação (45). Nesse caso, tem-se que $\xi_i = 0$ e o valor de b é dado pela Equação (49).

$$b = y_i - \boldsymbol{w} \cdot \boldsymbol{x_i} - \varepsilon \tag{49}$$

Analogamente, para $0 < \alpha_i^* \leq C$, tem-se $\xi_i^* = 0$ e

$$b = y_i - \boldsymbol{w} \cdot \boldsymbol{x_i} + \varepsilon\,. \tag{50}$$

Segundo (DO, 2009), em vez de usar um vetor de suporte aleatório para determinar b, uma solução mais estável é tomar b como sendo o valor médio de todos os b's obtidos da Equação (49) e (50), ou seja,

$$b = \frac{1}{S}\sum_{i=1}^{S}(y_i - \boldsymbol{w} \cdot \boldsymbol{x_i} - \varepsilon - \xi_i) \tag{51}$$

em que S o número de vetores de suporte.

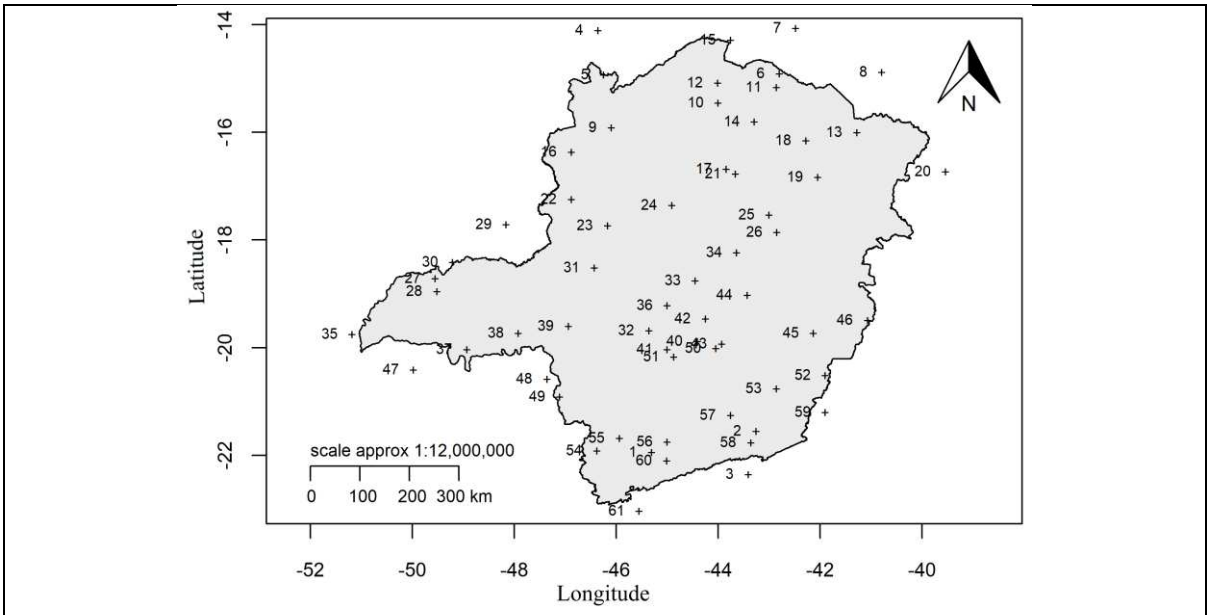
Maiores informações sobre SVM podem ser encontradas nos trabalhos de Burges (1998), Schölkopf e Smola (2002), Do (2009), Boyd e Vandenberghe (2009) e Vapnik (2013).

2.2 Descrição dos Dados

Foi selecionado um período de um ano do conjunto de dados obtidos do Banco de Dados Meteorológicos para Ensino e Pesquisa do Instituto Nacional de Meteorologia (INMET, 2018), no período de 01 de janeiro 1996 a 31 de dezembro de 2016. Dentre esse período notou-se que o ano de 2004 é o que possui o menor número de falhas nas medidas das variáveis meteorológicas. Dessa forma, optou-se nesse trabalho utilizar os dados desse ano. Quanto aos dados diários optou-se por utilizar somente o dia 15 de cada mês, por simples escolha.

Foram utilizados registros de 61 estações meteorológicas convencionais de observações de superfícies, das quais 48 dessas estações estão localizadas no estado do Minas Gerais e as outras 13 em estados circunvizinhos (2 no Rio de Janeiro, 3 em São Paulo, 1 em Mato Grosso do Sul, 3 em Góias e 4 na Bahia). O Estado de Minas Gerais (MG), localizado na região sudeste do Brasil, entre 14º13´58´´ e 22º54´00´´ S e 39º51´32´´ e 51º02´35´´ W possui área total de 586.520,732 km^2. Detalhes demográficos e outros detalhes da área de estudo são apresentados em (ÁVILA et al, 2014; REBOITA et al., 2015).

A distribuição espacial dessas 61 estações meteorológicas utilizadas no presente estudo, numeradas de 1 a 61 estão descritas na Figura 5.



1. Lambari-MG (878,5)
2. Coronel Pacheco-MG (435,0)
3. Avelar (Paty do Alferes)-RJ (507,0)
4. Posse-GO. (825,6)
5. Formoso-MG. (840,0)
6. Espinosa-MG. (569,6)
7. Caetité-BA. (882,5)
8. Vitoria da Conquista-BA (874,8)
9. Arinos-MG (519,0)
10. Januária-MG (473,7)
11. Monte Azul-MG (625,0)
12. Mocambinho-MG (452,0)
13. Pedra Azul-MG (648,9)
14. Janaúba-MG (516,0)
15. Carinhanha-BA (450,2)
16. Unai-MG (460,0)
17. Montes Claros-MG(652,0)
18. Salinas-MG (471,3)
19. Araçuaí-MG (289,0)
20. Guaratinga-BA (194,8)
21. Juramento-MG (648,0)
22. Paracatu-MG (712,0)
23. João Pinheiro-MG (760,4)
24. Pirapora-MG (505,2)
25. Carbonita-MG (736,4)
26. Itamarandiba-MG (914,0)
27. Capinópolis-MG (620,6)
28. Ituiutaba-MG (560,0)
29. Ipameri-GO (773,0)
30. Itumbiara-GO (449,0)
31. Patos de Minas-MG (940,3)
32. Bom Despacho-MG (695,0)
33. Curvelo-MG (672,0)
34. Diamantina-MG (1296,1)
35. Paranaíba-MS (331,3)
36. Pompeu-MG (690,9)
37. Frutal-MG (543,7)
38. Uberaba-MG (737,0)
39. Araxá-MG (1023,6)
40. Florestal-MG (760,0)
41. Bambuí-MG (661,3)
42. Sete Lagoas-MG (732,0)
43. Belo Horizonte-MG (915,0)
44. Conceição do Mato Dentro-MG (652,0)
45. Caratinga-MG (609,7)
46. Aimorés-MG (82,7)
47. Votuporanga-SP (502,5)
48. Franca-SP (1026,2)
49. São Sebastião do Paraíso-MG (820,0)
50. Ibirité-MG (814,5)
51. Divinópolis-MG (788,4)
52. Caparaó-MG (843,2)
53. Viçosa-MG (712,2)
54. Poços de Caldas-MG (1150,0)
55. Machado-MG (873,4)
56. Lavras-MG (918,8)
57. Barbacena-MG (1126,0)
58. Juiz de Fora-MG (940,0)
59. Itaperuna-RJ (123,6)
60. São Lourenço-MG (953,2)
61. Taubaté-SP (577,0)

Figura 5 – Mapa do Estado de Minas Gerais mostrando as localizações de 61 estações meteorológicas convencionais de MG e estados circunvizinhos, utilizadas neste estudo. O número entre parêntese na legenda refere-se a altitude (m) da estação.

As variáveis meteorológicas utilizadas neste estudo foram: precipitação (mm), temperatura mínima do ar (°C), duração do brilho solar (horas), evaporação do piche (mm), temperatura compensada média diária (°C), umidade relativa média diária (%), velocidade do vento média diária (m/s**)** e temperatura máxima do ar (°C). A variável resposta é a

temperatura máxima diária do ar e as outras 7 variáveis meteorológicas, juntamente com as coordenadas geográficas e altitude (m) atuaram como covariáveis.

A Tabela 1 apresenta o formato padrão de um conjunto com n observações e p covariáveis. A linha i refere-se a i-ésima observação ($i = 1, 2, \cdots, n$), a entrada $x_{ij}^{(m)}$ é o valor da j-ésima covariável X_j ($j = 1, 2, \cdots, p$) da observação i, no mês m e a última coluna $y_i^{(m)} = f(\boldsymbol{x}_i)$ é a função desconhecida que tenta-se predizer a partir das covariáveis.

Tabela 1 – Formato de um conjunto com n observações.

	Covariáveis				Resposta
Observação i	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\cdots$	$\boldsymbol{X_p}$	Y
1	$x_{11}^{(m)}$	$x_{12}^{(m)}$	$\cdots$	$x_{1p}^{(m)}$	$y_1^{(m)}$
2	$x_{21}^{(m)}$	$x_{22}^{(m)}$	$\cdots$	$x_{2p}^{(m)}$	$y_2^{(m)}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	$x_{n1}^{(m)}$	$x_{n2}^{(m)}$	$\cdots$	$x_{np}^{(m)}$	$y_n^{(m)}$

Fonte: Adaptada de Monard e Baranauskas (2003).

Neste estudo, $n = 61$ (número de estações), $p = 10$ (número de covariáveis), x_{ij} medida da i-ésima observação da j-ésima covariável X_j, $i = 1, 2, \cdots, 61$, $j = 1, 2, \cdots, 10$, $m = 1, 2, \cdots, 12$.

Refere-se como Conjunto Anual das covariáveis ($\overline{x_{ij}}$) e da variável resposta ($\overline{y_i}$) o conjunto obtido a partir da média do dia 15 de cada mês do ano 2004 (Equação (52).

$$\overline{x_{ij}} = \frac{1}{12}\sum_{m=1}^{12} x_{ij}^{(m)} \qquad \text{e} \qquad \overline{y_i} = \frac{1}{12}\sum_{m=1}^{12} y_i^{(m)} \tag{52}$$

Foi selecionada aleatoriamente, da base de dados para o dia 15 de cada mês, do ano de 2004, uma amostra com 41 pares $(\boldsymbol{x_i}, y_i) \in \mathbb{R}^{10} \times \mathbb{R}$, $i = 1, 2, \cdots, 61$, para o conjunto de treinamento e os outros 20 pares para o conjunto de teste.

Para execução das análises descritivas e o desenvolvimento dos modelos de regressão foi utilizado o software estatístico R, versão 3.5.2 (TEAM, 2018) com alguns pacotes principais como, geoR (RIBEIRO JR, DIGGLE, 2016), corrplot (WEI, SIMKO, 2017), train

(KUHN, 2018), dplyr (WICKHAM et al., 2018), rgdal (BIVAND, KEITT, ROWLINGSON, 2018) e caret (KUHN et al., 2018).

O pacote caret fornece uma interface para vários pacotes do R que são utilizados nas técnicas de aprendizagem de máquina para classificação e regressão, como o randomForest (BREIMAN, CUTLER, 2018), baseado no algoritmo de RF e o e1071 (MEYER et al., 2018), baseado no algoritmo SVM. Este pacote possui várias funções que automatiza a escolha dos valores para alguns parâmetros dos modelos de predição com o intuito de simplificar o processo de construção e avaliação do modelo.

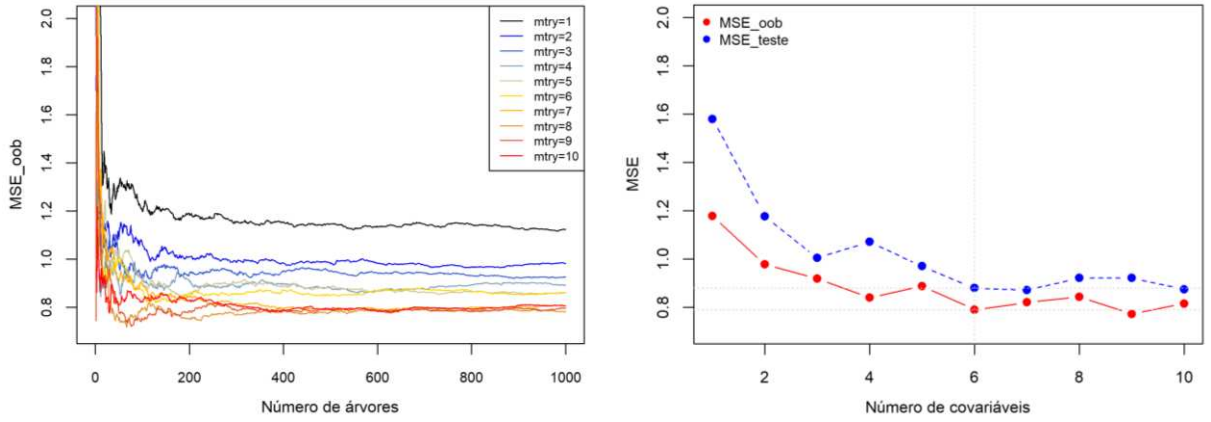
Para avaliar os desempenhos dos modelos de regressão utilizou-se as métricas de erro (MSE_{oob}, RMSE e MAE) e o R^2.

Para a obtenção dos mapas espaciais utilizou-se o software Grid Analysis and Display System (GrADS: DOTY, KINTER, 1993) e o software estatístico R. A reanálise ERA-Interim (DEE, UPPALA, 2009) do European Centre for Medium-Range Weather Forecasts (ECMWF) foi utilizada para obter as coordenadas geográficas e altitudes de uma grade regular e seu mapa espacial foi utilizado na comparação com os resultados dos modelos estudados. Foram extraídos dessa reanálise a temperatura máxima do ar a 2 m de altura e a altitude para o mês de janeiro e abril de 2004. Foi selecionada a região que engloba o Estado de MG (13ºS a 24ºS; 39ºW a 52ºW) com resolução espacial de 0,5° × 0,5°, ou seja, 27 pontos em longitude e 23 pontos em latitude, totalizando uma grade com 621 pontos.

A reanálise ECMWF/ERA-Interim é amplamente utilizada na comunidade cientifica para validar modelos que predizem variáveis meteorológicas, como no trabalho de Miralles et al. (2012). Também existem outras reanálises que poderiam ser utilizadas, no entanto, estudos mostram que a ERA-Interim tem proporcionado resultados mais realísticos (MOONEY, MULLIGAN, FEALY, 2011).

De acordo com Liaw e Wiener (2002), para a utilização de uma RF é necessário definir três principais parâmetros: o número de covariáveis utilizadas em cada árvore (mtry), o número de árvores construídas pelo algoritmo (ntree) e o número mínimo de observações em cada nó terminal (nodesize). Neste trabalho, utilizou-se o nodesize = 5, que é o padrão do pacote randomForest do software estatístico R. Quanto ao número de árvores, foram realizados vários testes variando o número de árvores de 1 a 1000. Nota-se na Figura 6a que existe uma convergência dos erros a partir de certo número de árvores. Pode-se observar que a partir de aproximadamente 100 árvores já não existem ganhos significativos de desempenho,

mas, nesse estudo optou-se por utilizar 400 árvores por praticamente não apresentar variações no MSE a partir desse número.



(a) (b)

Figura 6 – Taxa de erros do algoritmo Random Forest, para diferentes número de covariáveis, aplicado ao Conjunto Anual. (a) Erro out-of-bag em função do número de árvores utilizadas. (b) Erros out-of-bag (linha vermelha) e de teste (linha azul) calculado para 400 árvores.

Quanto ao mtry, o valor-padrão estipulado para problemas de regressão é de 1/3 do número total de covariáveis. O objetivo é minimizar tanto o MSE_{teste} quanto o MSE_{oob} (LIAW, WIENER, 2002).

Na Figura 6b, nota-se que tanto o MSE_{oob} quanto o MSE_{teste} possuem alto erros quando o número de covariáveis é baixo. Ao aumentar o número de covariáveis o erro vai decaindo até um limite. Priorizando inicialmente o MSE_{teste}, verificou-se que os menores valores ocorreram para 6, 7 e 10 covariáveis. Nesses valores, o melhor MSE_{oob} foi obtido com 6 covariáveis. Dessa forma, optou-se nesse trabalho por utilizar mtry = 6.

Neste trabalho serão apresentados os resultados obtidos com RF empregando ajuste manual (mtry = 6, nodesize = 5 e ntree = 400), sendo denominado de “RF_manual” e também o RF ajustado automaticamente por meio da função train do pacote caret do R, denominado de “RF_auto”.

Durante o processo de organização dos dados foram observadas ocorrências de algumas falhas nas medidas das variáveis meteorológicas, em determinadas estações. Entretanto, os pacotes utilizados não trabalham com valores indefinidos. Desta forma, para completar as falhas utilizou-se a Interpolação Ponderada pelo Inverso da Distância (IDW), considerando a média dos valores das três estações mais próximas, ponderada pela distância.

Após o tratamento dessas falhas, foram calculadas estatísticas descritivas da temperatura máxima diária do ar dos conjuntos de dados utilizados (Tabela 2).

Tabela 2 – Estatística Descritiva dos valores Observados da Temperatura Máxima do Ar (ºC) no dia 15 de cada mês do ano de 2004 e no Conjunto Anual, em relação às 61 estações de estudo.

	$\bar{y}$	S	min	max	Q_1	m_d	Q_3
jan	29,7	2,5	24,0	33,8	28,1	30,0	31,4
fev	26,9	3,4	20,8	33,0	24,4	26,6	29,7
mar	27,8	2,6	21,8	32,3	25,8	28,0	29,6
abr	27,2	2,2	19,8	31,1	25,4	27,3	28,8
mai	27,3	3.7	17,0	34,0	25,8	27,8	29,7
jun	24,4	3,3	16,8	29,6	22,0	25,4	26,8
jul	26,7	2,6	20,1	31,8	25,2	26,7	28,7
ago	25,0	2,6	18,4	29,3	23,5	25,1	27,2
set	31,4	2,7	25,3	36,1	29,6	31,6	33,6
out	31,1	4,3	21,2	37,9	27,5	31,9	34,7
nov	28,1	3,5	20,1	34,8	25,5	28,2	30,6
dez	29,2	2,5	24,3	35,0	27,2	29,5	30,9
Conjunto Anual	27,8	2,1	22,4	31,5	26,5	28,3	29,4

$\bar{y}$: média, S: desvio padrão, min: mínima, max: máxima, Q_1: primeiro quartil, m_d: mediana, Q_3: terceiro quartil.

A Figura 7 representa uma visão espacial dos valores da temperatura máxima do ar (Tmax) no dia 15 de cada mês, do ano de 2004. Os tamanhos dos círculos (ou as cores) representam os quartis da Tmax. Por exemplo, de acordo com a Tabela 2, para janeiro os círculos de cores azuis indicam Tmax inferiores a 28,1°C, os verdes indicam Tmax entre 28,1°C a 30,0°C, os amarelos valores entre 30,0°C a 31,4°C e os vermelhos valores de Tmax acima de 31,4°C.

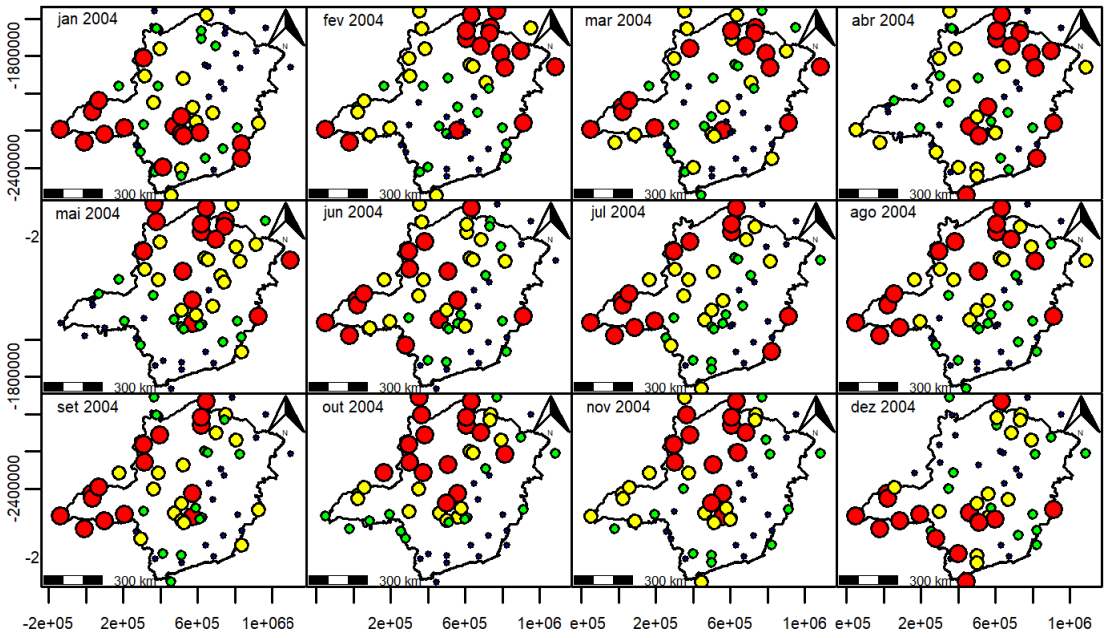


Figura 7 – Mapas espaciais indicando a variação da temperatura máxima observada em cada estação meteorológica conforme os valores quartis da temperatura máxima. Os tamanhos (ou as cores) dos círculos correspondem aos quartis da temperatura máxima no dia 15 de cada mês de 2004.

Na prática, o importante é estimar a temperatura máxima do ar em locais onde não se têm medidas de variáveis meteorológicas. Pois, é rara a existência de estação meteorológica que possui medidas das covariáveis, mas não faz medida da temperatura máxima do ar. Dessa forma, nesse estudo também se aplicou os modelos de aprendizado de máquina para os conjuntos de treinamentos contendo somente a temperatura máxima do ar (variável resposta) e as coordenadas geográficas e altitude das estações meteorológicas em estudo (covariáveis). Assim, após o treinamento, é possível estimar a temperatura máxima do ar em qualquer local desejado, desde que conheça sua longitude, latitude e altitude.

Então, neste estudo foram realizadas duas etapas para se obter a Tmax.. Na primeira etapa foram executadas as análises de regressão para os modelos de RLM, RF_auto, RF_manual e SVM_Lin, tendo como covariáveis as outras 7 variáveis meteorológicas e as coordenadas geográficas e altitude. Na segunda etapa foram executadas as análises para os modelos RLM, RF_auto, SVM_Lin e IDW, utilizando como covariáveis apenas a longitude, latitude e altitude. Nessa etapa, o modelo manual de RF não foi analisado, pois ao refazer o teste manual para os parâmetros do RF verificou-se que o número de variáveis que

proporcionou o menor MSE foi obtido com todas as três covariáveis. Dessa forma, o ajuste manual se identificou com o automático, que também utiliza as três covariáveis. Optou-se também comparar a temperatura máxima do ar obtida pelo método IDW, por ele ser um método muito utilizado (PERIN et al., 2015).

3 RESULTADOS E DISCUSSÃO

Inicialmente, para a predição da temperatura máxima do ar (Tmax), foi utilizado o Conjunto Anual com todas as covariáveis disponíveis (longitude (lon), latitude (lat), altitude (alt), precipitação (prec), temperatura mínima do ar (Tmin), insolação (insol), evaporação (evap), temperatura compensada (Tcomp), umidade relativa (UR), velocidade do vento (vento)).

Nota-se na Figura 8, que a variável Tcomp tem altas correlações positivas com a maioria das variáveis. Já a prec e o vento possuem baixas correlações com o conjunto de variáveis. Quanto à UR, normalmente apresenta correlações negativas fortes.

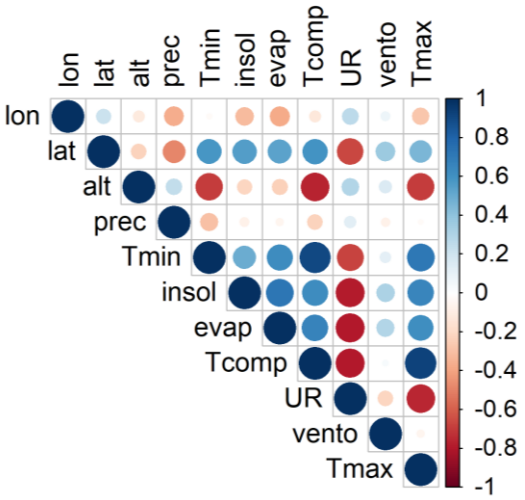


Figura 8 – Correlação entre as variáveis do banco de dados do Conjunto Anual.

Foi realizado um ajuste da média da temperatura máxima do ar do dia 15, nos meses do ano de 2004, observadas nas 21 estações do conjunto de teste e plotadas em função de cada covariável (curva de linhas pretas da Figura 9). O mesmo ajuste foi realizado para a predição de três dos modelos aqui analisados. No entanto, optou-se por apresentar para a RLM em função da alt, evap e insol (linhas azuis), o RF_manual em função da lat, lon e prec (linhas laranjas) e o SVM_Lin em função da Tmin, UR e vento (linhas cinzas).

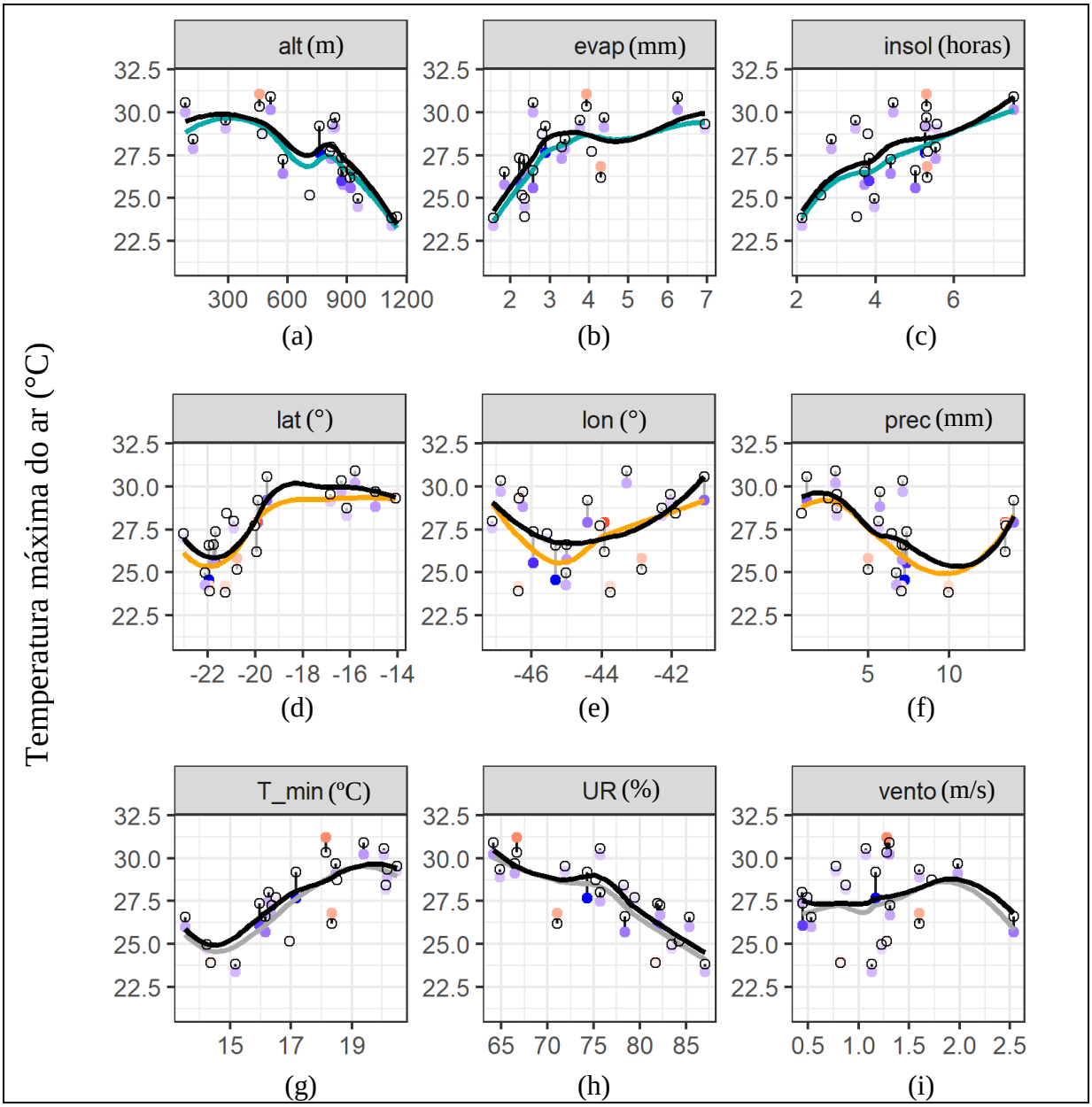


Figura 9 – Temperatura máxima do ar em função das covariáveis utilizando o Conjunto Anual. O tamanho do segmento é proporcional ao valor do viés entre a temperatura máxima observada no conjunto de teste e a predita pelos modelos: de RLM (gráficos a, b, c), RF_manual (gráficos d, e, f), SVM_Lin (gráficos g, h, i). Pontos com cores em tons azuis representam subestimativas da predição e vermelhos superestimativas. As linhas pretas representam as tendências das observações e as coloridas representam as tendências das predições.

Conforme esperado, a temperatura máxima do ar apresenta tendência decrescente com o aumento da altitude (Figura 9a) e também, com o aumento da umidade relativa do ar (Figura

9h). Já para a evaporação e insolação apresenta tendência crescente, uma vez que altas temperaturas favorecem a evaporação e maior tempo de brilho solar favorece o aumento da temperatura do ar (Figura 9b, 9c). Nota-se também que a precipitação e a velocidade do vento não evidenciaram a existência de tendências positivas ou negativas (Figura 9f, 9i).

Na análise de regressão para os modelos de RLM, RF_auto, RF_manual e SVM_Lin, em que utilizou todas as covariáveis disponíveis, é possível notar que o modelo com menor viés foi o SVM_Lin e o que mais distanciou dos dados observados foi o RF_manual, com predições geralmente mais frias do que a observada.

Na Figura 9 também é possível ter uma visão descritiva das covariáveis em estudo para o conjunto de teste. Por exemplo, é possível observar que em grande parte das estações ocorreu, em média, insolação em torno de 5 horas e apenas uma estação do conjunto de teste teve insolação média acima de 6 horas (Figura 9c).

A performance de cada modelo estudado foi analisada em função das métricas RMSE, R^2 (Figura 10) e MAE (Tabela 3).

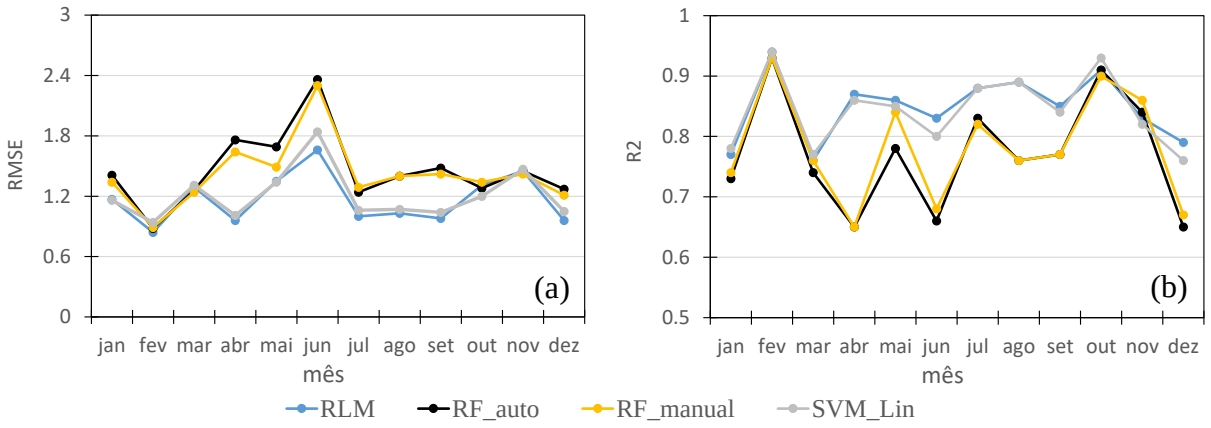


Figura 10 – Validação das predições dos modelos RLM (curva azul), RF_auto (curva preta), RF_manual (curva laranja) e SVM_Lin (curva cinza), realizadas para o dia 15 de cada mês do ano de 2004 utilizando todas as covariáveis disponíveis. (a) Raiz do erro quadrático médio. (b) Coeficiente de determinação.

O algoritmo RLM apresentou bons resultados, já que no período estudado o valor máximo do RMSE da temperatura máxima do ar foi de 1,66°C (Figura 10a), com variabilidade sempre acima de 77 % (Figura 10b) e MAE abaixo de 1,30°C (Tabela 3). A boa performance desse modelo pode ser explicada pelas covariáveis utilizada. Devido às altas

correlações entre as covariáveis (foi detectado fator de inflação da variância maior do que 5), pode existir multicolinearidade, o que gera efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo. Assim, como não foi realizada a correção da multicolinearidade, o coeficiente de determinação (R^2) tende a ficar próximo da unidade, ou seja, os valores do R^2 para RLM, apresentados na Figura 10b, podem ter sido influenciados pela multicolinearidade. Em relação ao RF, o com ajuste manual apresentou resultados ligeiramente melhores do que o de ajuste automático. No entanto, ambos apresentaram resultados inferiores aos outros modelos dois modelos analisados.

Tabela 3 – Comparação de precisão em termos de valores do MAE para os modelos utilizados nesse estudo, aplicado ao conjunto de dados com todas as covariáveis, para o dia 15 de cada mês do ano de 2004.

	RLM	RF_auto	RF_manual	SVM_Lin
jan	1,04	1,09	1.04	1,06
fev	0,74	0,66	0,68	0,80
mar	0,95	0,96	0.92	0,97
abr	0,78	1,26	1,16	0,84
mai	1,06	1,14	1,09	1,06
jun	1,30	1,80	1,77	1,47
jul	0,83	1,00	1,03	0,89
ago	0,88	1,14	1,14	0,90
set	0,73	1,18	1,07	0,80
out	1,03	1,03	1,09	0,91
nov	1,05	1,05	1,03	1,10
dez	0,75	1,09	1,02	0,81

Considerando o comportamento médio dos modelos ao longo do ano de 2014, pode-se observar na Tabela 4 que, ao utilizar todas as covariáveis, o SVM_Lin foi o que obteve os menores erros e o maior R^2, seguido do RLM. Já o RF_auto foi o que teve a pior predição da temperatura máxima do ar, ao considerar essas três métricas estatísticas.

Tabela 4 – Comparação de precisão em termos de valores das métricas RMSE, R^2 e MAE para os modelos RLM, RF_auto, RF_manual, SVM_Lin aplicado ao Conjunto Anual com todas as covariáveis.

	RLM	RF_auto	RF_manual	SVM_Lin
RMSE	0,72	1,18	0,95	0,65
R^2	0,93	0,77	0,84	0,93
MAE	0,61	0,99	0.76	0,53

Na segunda etapa passou-se a análise dos modelos RLM, RF_auto, SVM_Lin e IDW, com conjuntos de dados constituídos somente dos valores da temperatura máxima do ar, das coordenadas geográficas e da altitude das 61 estações meteorológicas em estudo. A Figura 11 apresenta o desempenho dos modelos RLM, RF_auto, SVM_Lin e IDW via as métricas RMSE e R^2.

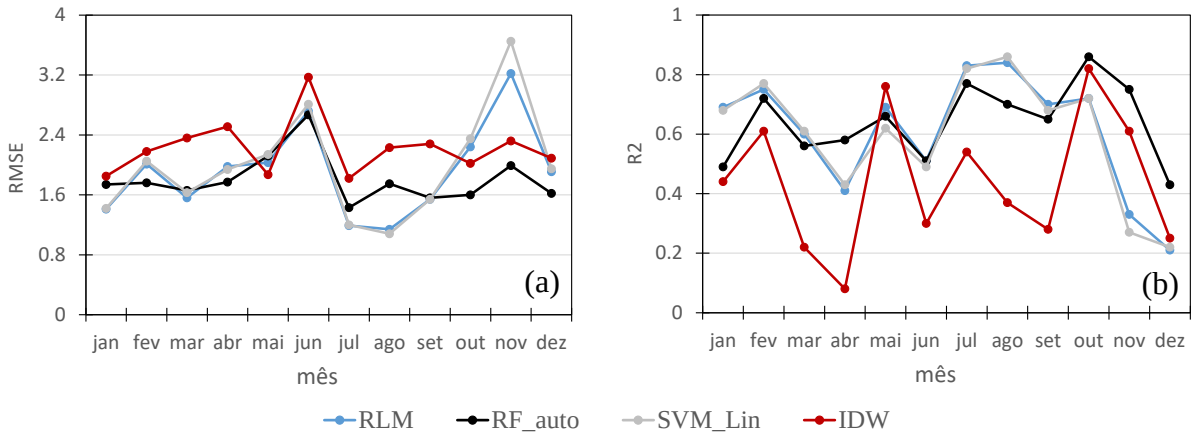


Figura 11 – Validação dos modelos RLM (curva azul), RF_auto (curva preta), RF_manual (curva laranja), SVM_Lin (curva cinza) e IDW (linha vinho), utilizando somente as covariáveis lon, lat e alt, no dia 15 de cada mês do ano de 2004. (a) Raiz do erro quadrático médio. (b) Coeficiente de determinação.

Já na Tabela 5 é apresentado o desempenho desses quatro modelos em relação à métrica MAE:

Tabela 5 – Comparação de precisão em termos de valores do MAE para os modelos utilizados nesse estudo aplicado ao conjunto de dados utilizando somente as covariáveis latitude, longitude e altitude, no dia 15 de cada mês do ano de 2004.

	RLM	RF_auto	SVM_Lin	IDW
jan	1,27	1,34	1,29	1,41
fev	1,56	1,48	1,59	1,41
mar	1,33	1,33	1,38	1,66
abr	1,54	1,40	1,52	1,96
mai	1,69	1,60	1,67	1,5
jun	2,26	2,15	2,28	2,28
jul	0,88	1,12	0,89	1,38
ago	0,91	1,37	0,87	1,83
set	1,34	1,34	0,28	1,80
out	1,76	1,18	1,85	1,63
nov	2,60	1,55	2,97	1,98
dez	1,54	1,32	1,52	1,52

Conforme pode-se observar na Figura 11 e também na Tabela 5, no dia 15 de novembro de 2004, os modelos RLM e SVM_Lin tiveram elevados erros e baixo R^2. Nesse dia havia uma frente fria entrando no Estado de Minas Gerais na direção de sudoeste para sudeste, conforme pode-se observar na Figura 12a. Essa frente fez com que nesse dia houvesse forte gradiente de temperatura no estado de Minas Gerais (Figura 12b) A grande diferença de temperatura entre a região sudeste e noroeste de Minas pode ter sido a causa do elevado erro apresentado por estes dois modelos. Já o RF apresentou-se erros de predições mais regulares, de forma que durante esse episódio teve apenas um pequeno acréscimo em seus erros. Quanto ao IDW, foi o que apresentou a menor performance entre os modelos avaliados, porém no dia 15 de novembro de 2004 ele foi bem melhor que o RLM e o SVM_Lin.

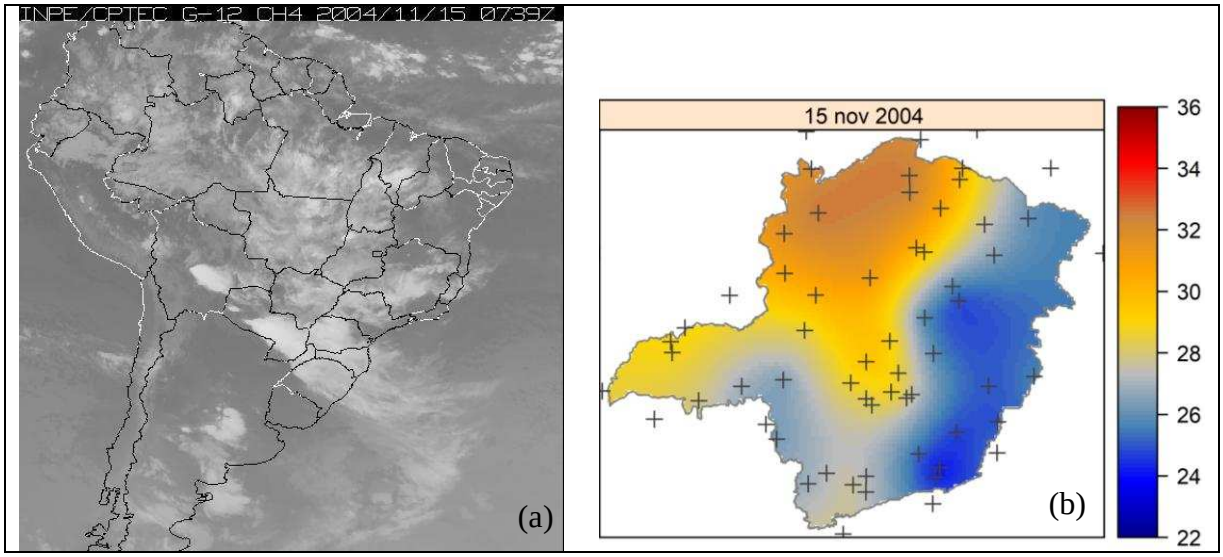


Figura 12 – (a) Imagem do canal 4 do satélite GOES-12 em 15/11/2004 às 07:39 (GMT); (b) Campo espacial da temperatura máxima do ar do estado de MG no dia 15 de novembro 2004 gerado pela metodologia apresentada no capítulo 1.
Fonte: (a) DSA/CPTEC/INPE.

Avaliando o Conjunto Anual, quando utilizou-se somente três covariáveis, observa-se na Tabela 6 que o algoritmo SVM_Lin e RLM apresentaram resultados muito similares, e bem superiores ao RF e IDW.

Tabela 6 – Comparação de precisão em termos de valores das métricas RMSE, R^2 e MAE para os modelos RLM, RF_auto, RF_manual, SVM_Lin aplicado ao Conjunto Anual utilizando somente as covariáveis latitude, longitude e altitude.

	RLM	RF_auto	RF_manual	SVM_Lin
RMSE	1,1	1,51	1,11	1,89
R^2	0,83	0,58	0,83	0,34
MAE	0,88	1,29	0,89	1,41

As Figuras 13 e 14 comparam as estimativas dos modelos em estudo com a reanálise ERA-Interim do ECMWF e apresentam o seu viés em relação ao valor observado da temperatura máxima do ar em cada estação de estudo nos dias 15 de janeiro e 15 de abril de 2004, respectivamente. Os números negativos nas Figuras 13a e 14a mostram que os valores estimados pela reanálise do ECMWF são menores que os valores observados.

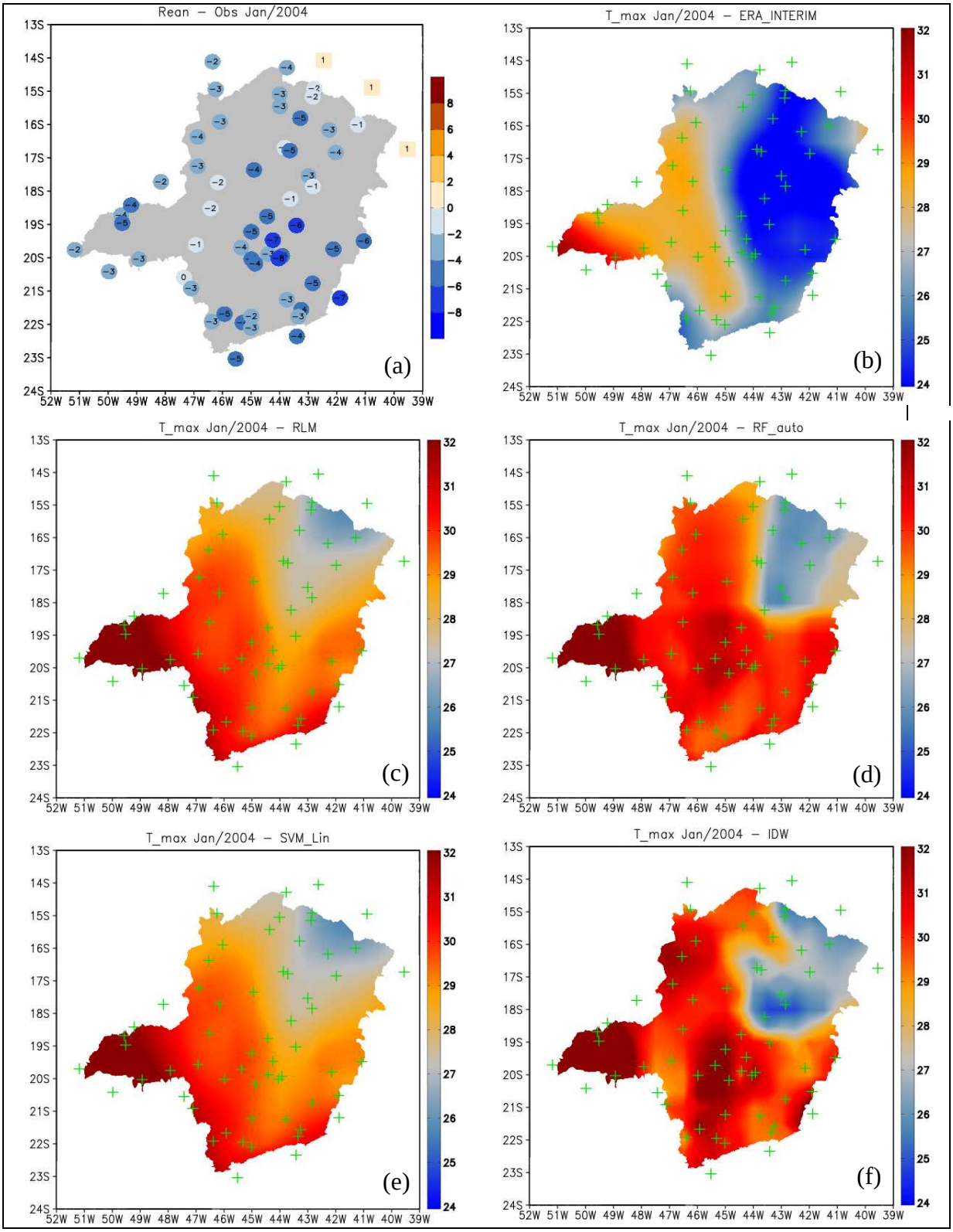


Figura 13 – (a) Viés da temperatura máxima do ar (diferença entre reanálise do ECMWF e observação), no dia 15/01/2004. As cores e os números dentro dos círculos/quadrados referem-se aos valores aproximados do viés; (b-f) Campo espacial da Tmax do estado de MG no dia 15/01/2004, gerado pela: (b) reanalise do ECMWF; (c) metodologia RLM; (d) metodologia RF; (e) metodologia SVM_Lin; (f) metodologia IDW.

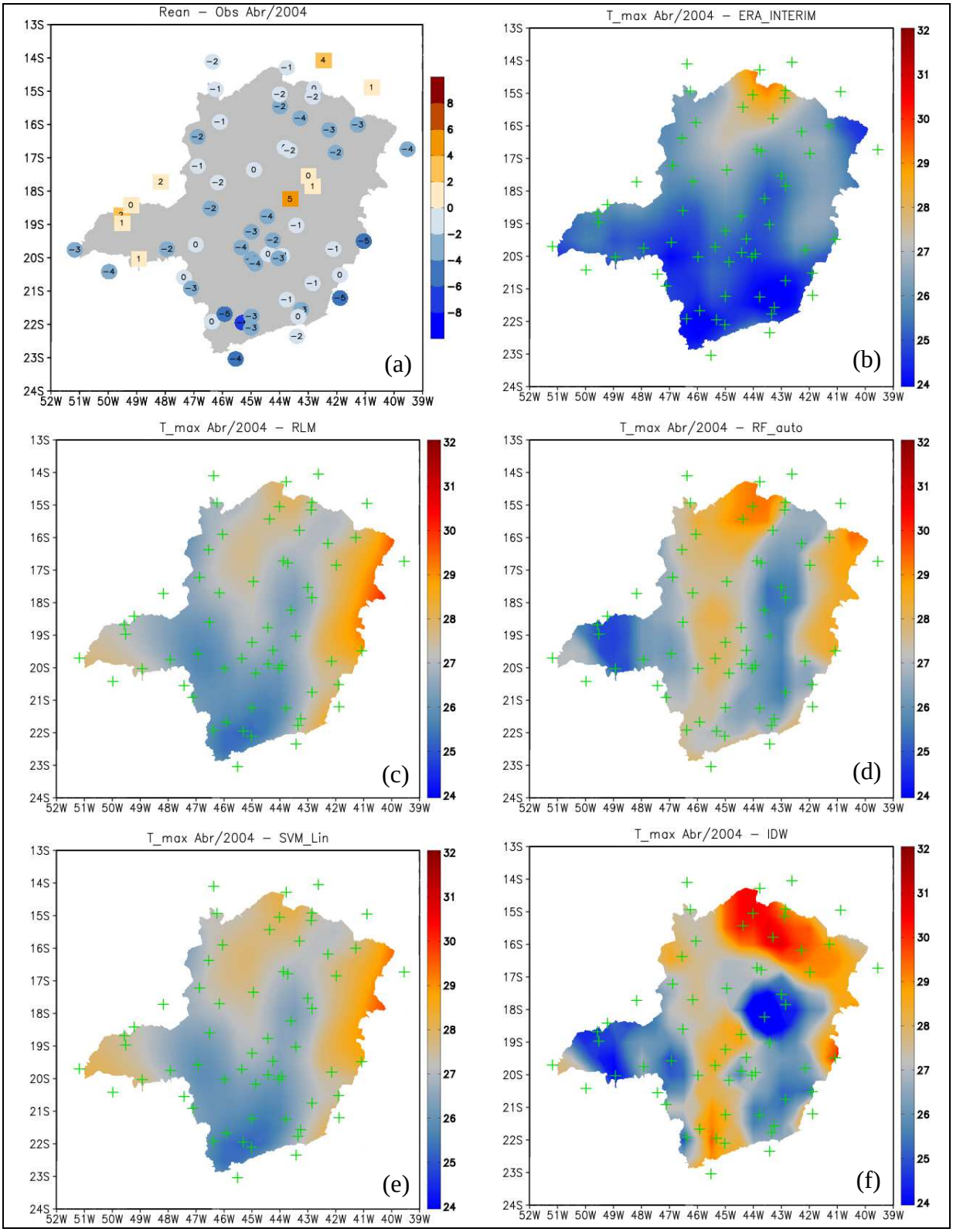


Figura 14 – (a) Viés da temperatura máxima do ar (diferença entre reanálise do ECMWF e observação), no dia 15/04/2004. As cores e os números dentro dos círculos/quadrados referem-se aos valores aproximados do viés; (b-f) Campo espacial da temperatura máxima do ar do estado de MG no dia 15/04/2004, gerado pela: (b) reanalise do ECMWF; (c) metodologia RLM; (d) metodologia RF; (e) metodologia SVM_Lin; (f) metodologia IDW.

Nota-se na Figura 13c-13f que todos os modelos apresentados neste estudo estimaram, para o dia 15 de janeiro de 2004, temperaturas mais frias ao nordeste de Minas e mais quente a oeste. Esse padrão também é observado na reanálise do ECMWF (Figura 13b).

No entanto, no geral, esses modelos simularam temperaturas máximas do ar mais elevadas do que a reanálise. O modelo RLM e SVM_Lin forneceram resultados muito semelhantes (Figuras 13c e 13e), conforme mostrado também na validação desses dois modelos ao longo do ano de 2004 (Figura 11). O algoritmo de RLM (ou SVM_Lin) proporcionou campos mais suavizados do que o RF_auto e o IDW. O RF_auto mostrou um gradiente forte de temperatura no nordeste de Minas Gerais e tanto o RF_auto quanto o IDW simularam um núcleo com altas temperaturas máximas ao centro do estado, o que não é observado na reanálise. Mas, pode-se notar na Figura 13a que, justamente no local onde ocorre esse núcleo, o viés da reanálise é muito negativo, ou seja, os dados observados estavam em torno de 5ºC mais quente do que o da reanálise. Portanto, a Figura 13a mostra que as incertezas no ERA-Interim podem afetar também algumas das análises apresentadas.

A Figura 15 foi obtida com a metodologia apresentada no capítulo 1 deste trabalho, na qual utiliza-se a geoestatística espaço-temporal para predizer a temperatura máxima do ar no estado de Minas Gerais.

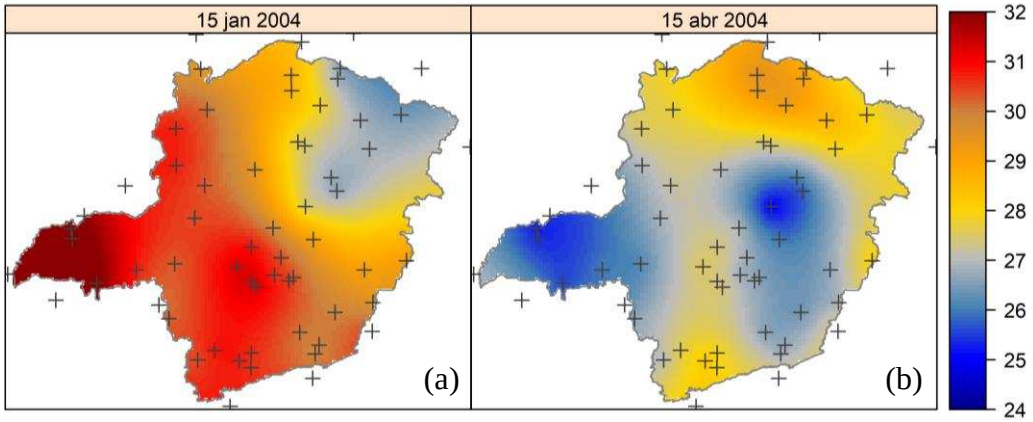


Figura 15 – Campo espacial da temperatura máxima do ar do estado de MG gerado pela metodologia geoestatística espaço-temporal apresentada no capítulo 1. (a) predição para o dia 15 de janeiro 2004; (b) predição para o dia 15 de abril 2004.

Observa-se que os campos obtidos com a metodologia da geoestatística espaço-temporal (Figura 15) identifica padrões similares aos obtidos pelos modelos RLM, RF,

SVM_Lin e IDW (Figuras 13 e 14). No entanto, ao comparar os campos espaciais nota-se que, para janeiro, a metodologia espaço-temporal produziu campos mais próximos do modelo SVM_Lin e RLM. Já para abril, ficou mais semelhante aos padrões apresentados pelo IDW. Nota-se, também, que os campos produzidos pela geoestatística espaço-temporal são visualmente mais suavizados do que os encontrados com os modelos de aprendizagem de máquina e o IDW.

4 CONCLUSÕES

Com o ajuste manual do algoritmo random forest, em que utilizou-se 6 covariáveis, 400 árvores e no mínimo 5 observações em cada nó, obteve-se um ligeira melhora em relação ao seu ajuste automático. No entanto, ao utilizar somente as variáveis latitude, longitude e altitude, o ajuste manual passou a ser idêntico ao automático, uma vez que o menor MSE_{teste} foi obtido ao utilizar todas as três covariáveis.

A temperatura máxima do ar apresenta forte dependência com a evaporação, insolação, temperatura mínima, altitude e umidade relativa. Nas três primeiras, essa relação é diretamente proporcional e, nas duas últimas, possui uma relação inversa. Já o vento e a precipitação não mostraram grandes influências na temperatura máxima do ar.

O modelo de regressão linear múltipla e o support vector machine linear apresentaram resultados muito similares, mas esse resultado pode não ser conclusivo uma vez que é reduzido o número de covariáveis utilizado nesse estudo.

Dados meteorológicos são coletados em estações que normalmente possuem instrumentos capazes de medir simultaneamente todas as variáveis meteorológicas apresentadas neste estudo. Portanto, na prática, o uso delas como covariáveis para obter a temperatura máxima é indevido, uma vez que quando se tem o valor dessas covariáveis já se têm também a temperatura máxima do ar. Assim, ao utilizar somente as covariáveis longitude, latitude e altitude foi possível calcular a temperatura máxima do ar em uma grade regular com resolução de 0,5ºx0,5º, cobrindo todo o estado de Minas Gerais e comparar o campo espacial produzido pelos modelos estudados com a reanálise do ECMWF. No entanto, comparando as métricas estatística das predições obtidas com os dois conjuntos de covariáveis, o RMSE e o MAE aumentaram, em média, ~0,4°C e ~0,30ºC, respectivamente. Já o R^2, reduziu, em média, em torno de 13%, quando utilizou-se somente as três covariáveis. Os modelos regressão linear múltipla e support vector machine Linear continuaram apresentando desempenhos melhores do que o random forest e o modelo ponderado pelo Inverso da Distância. No entanto, em um dia em que ocorreu entrada de frente fria no estado de Minas Gerais, os modelos de regressão linear múltipla e o support vector machine linear tiveram performance bem inferior aos outros dois.

Os campos produzidos pela geoestatística espaço-temporal apresentaram mais suavizados do que os encontrados com os modelos de aprendizagem de máquina e o IDW, mas tanto a aprendizagem de máquina quanto a geoestatística espaço-temporal reproduziram os principais padrões apresentados na reanálise do ECMWF.

REFERÊNCIAS BIBLIOGRÁFICAS

ÁVILA, L. F. et al. Tendências de temperaturas mínimas e máximas do ar no Estado de Minas Gerais. **Pesquisa Agropecuária Brasileira**, v. 49, n. 4, p. 247-256, 2014.

BIVAND, R.; KEITT, T.; ROWLINGSON, B. Rgdal: Bindings for the geospatial data abstraction library. **R package version 1.3-6**, 2018.

BOYD, S.; VANDENBERGHE, L. **Convex optimization**. New York: Cambridge university press, 2009. 716p.

BREIMAN, L. Bagging predictors. **Machine learning**, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BREIMAN, L.; CUTLER, A. Package Random Forest: Breiman and Cutler's Random Forests for Classification and Regression. n. **R package version 4.6-14**, 2018.

BURGES, C. J. C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, v. 2, n. 2, p. 121-167, 1998.

DEE, D. P.; UPPALA, S. Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. **Quarterly Journal of the Royal Meteorological Society**, v. 135, n. 644, p. 1830-1841, 2009.

DO, C. B. **Convex Optimization Overview (cnt'd)**. 2009. Disponível em <http://conglang.github.io/img/cs229-cvxopt2.pdf>. Acesso em: 16 nov. 2018.

DOS ANJOS ANTONINI, J. C. et al. **Modelo de Estimativa da Temperatura Média Diária do Ar no Estado de Goiás**. Planaltina, DF: Embrapa Cerrados, 2010, 27 p, 2010.

DOTY, B.; KINTER, J. L. III. The grid analysis and display system (GrADS): a desktop tool for earth science visualization. In: **American geophysical union 1993 fall meeting**, p. 6–10, 1993. Disponível em: <http://cola.gmu.edu/grads/>. Acesso em: 03 de jun. 2018.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LCT, 2011.

FARAWAY, J. J. **Data splitting strategies for reducing the effect of model selection on inference**. Technical Report 259, University of Michigan, 1995. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.902&rep=rep1&type=pdf>. Acesso em: 09 nov. 2018.

GOMES, D. P. et al. Estimativa da temperatura do ar e da evapotranspiração de referência no estado do Rio de Janeiro. **Irriga**, v. 19, n. 2, p. 302-314, 2014.

GUO, Q.; KELLY, M.; GRAHAM, C. H. Support vector machines for predicting distribution of Sudden Oak Death in California. **Ecological Modelling**, v. 182, n. 1, p. 75-90, 2005.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning. data mining, inference, and prediction**. Second edition, New York: Springer series in statistics, 2009. 739p.

IPCC - INTERGOVERNMENTAL PANEL IN CLIMATE CHANGE. Summary for policymakers. In: **Climate Change 2014**: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T. E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1-32, 2014. Disponível em: <https://www.ipcc.ch/pdf/assessment-report/ar5/wg2/ar5_wgII_spm_en.pdf>. Acesso em: 21 jun. 2018.

INMET - INSTITUTO NACIONAL DE METEOROLOGIA. **Banco de Dados Meteorológicos para Ensino e Pesquisa** (BDMEP), 2018. Disponível em: <http://www.inmet.gov.br/portal>. Acesso em: 02 abr. 2018.

IZBICKI, R.; DOS SANTOS, T. M. **Machine Learning sob a ótica estatística**. 2018. Disponível em: < http://www.rizbicki.ufscar.br/sml.pdf>. Acesso em: 08 nov. 2018.

JAMES, G. et al. **An introduction to statistical learning with applications in R**. Springer New York Heidelberg Dordrecht London, 2013. 441p.

KUHN et al., M. Caret: Classification and Regression Training. **R package version 6.0-81**. 2018.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.

MEYER, D. et al. e1071: misc functions of the department of statistics, probability theory group (formerly: e1071), TU Wien. **R package version 1.7-0.1**. 2018.

MIRALLES, D. G. et al. Soil moisture-temperature coupling: A multiscale observational analysis. **Geophysical Research Letters**, v. 39, n. 21, 2012.

MONARD, M. C.; BARANAUSKAS, J. A. Capítulo 4. Conceitos sobre Aprendizado de Máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, n. 1, p. 39-56, 2003.

MOONEY, P. A.; MULLIGAN, F. J.; FEALY, R. Comparison of ERA-40, ERA-Interim and NCEP/NCAR reanalysis data with observed surface air temperatures over Ireland. **International Journal of Climatology**, v. 31, n. 4, p. 545-557, 2011.

PERIN, E. B. et al. Interpolação das variáveis climáticas temperatura do ar e precipitação: revisão dos métodos mais eficientes. **Geografia**, v. 40, n. 2, 2015.

RAMOS, C. et al. Modelagem da variação horária da temperatura do ar em Petrolina, PE, e Botucatu, SP. **Revista Brasileira de Engenharia Agrícola e Ambiental**, p. 959-965, 2011.

REBOITA, M. S. et al. Aspectos climáticos do estado de Minas Gerais (climate aspects in minas gerais state). **Revista Brasileira de Climatologia**, v. 17, 2015. Disponível em: <http://revistas.ufpr.br/revistaabclima/article/view/41493/27319>. Acesso em: 14 de nov. 2018.

RIBEIRO JR, P. J.; DIGGLE, P. J. GeoR: Analysis of geostatistical data. **R package version, 1.7-5.2,** 2016.

RIOS, E.S. O **efeito de borda na geoestatística**. 2018. 43 p. Dissertação (Mestrado em Estatística em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, MG, 2018. Disponível em: <http://www.locus.ufv.br/bitstream/handle/123456789/18405/texto%20completo.pdf?sequence=1>. Acesso em: 05 abr. 2019.

RODRIGUEZ-GALIANO, V. et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804-818, 2015.

SCHÖLKOPF B.; SMOLA, A. J. Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, pages110-146, 2002. 645p.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199-222, 2004.

TEAM, R. C. R: **A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 23 de abril 2018.

VAPNIK, V. N. **The nature of statistical learning theory**. Springer science & business media, New York, 2nd ed, 2013.

WEI, T.; SIMKO, V. corrplot: Visualization of a correlation matrix. **R package version 0.84**, v. 56, n. 231, p. 316-324, 2017.

WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. **R package version 0.7. 8**. 2018.

CONCLUSÕES GERAIS

Esta pesquisa utilizou a Geoestatística espaço-temporal e a aprendizagem de máquina com o principal objetivo de obter predições, de uma variável resposta contínua, com dependência espacial, a partir de um conjunto de dados medidos ao longo da região de estudo. Dessa forma, é possível obter valores preditos em qualquer ponto interior à região de estudo ou para um conjunto de pontos equidistantes que permite construir uma grade regular para gerar campos espaciais da variável resposta.

Neste trabalho utilizou-se como variável resposta a temperatura máxima do ar no Estado de Minas Gerais. O estudo da variável temperatura do ar é de suma importância devido a sua fundamental contribuição em diversas áreas do conhecimento. O aumento da concentração de gases de efeito estufa na atmosfera está provocando a elevação da temperatura média do planeta. Portanto, esse trabalho visa contribuir com desafios tais como, aquecimento global, urbanização descontrolada, escassez de recursos naturais, epidemias e catástrofes naturais.

Na geoestatística espaço-temporal, via modelagem por estrutura de covariâncias espaço-temporal, foram utilizados 5 modelos de covariâncias em que o modelo de covariância soma-métrico foi o que melhor se ajustou aos dados. As médias espaciais dos dados preditos por esse modelo ficaram muito próximas das médias dos dados observados nas 61 estações utilizadas, no período de 1996 a 2016, com exceção dos últimos três anos avaliados, por conter muitas falhas nas observações. Os resultados comprovaram que as predições foram bem sucedidas e mostrou que é possível obter campos espaciais para a temperatura máxima diária do ar na região que engloba o estado de Minas Gerais, com alta qualidade, ao utilizar a metodologia da geoestatística espaço-temporal que considera tanto as variabilidades espaciais quanto temporais.

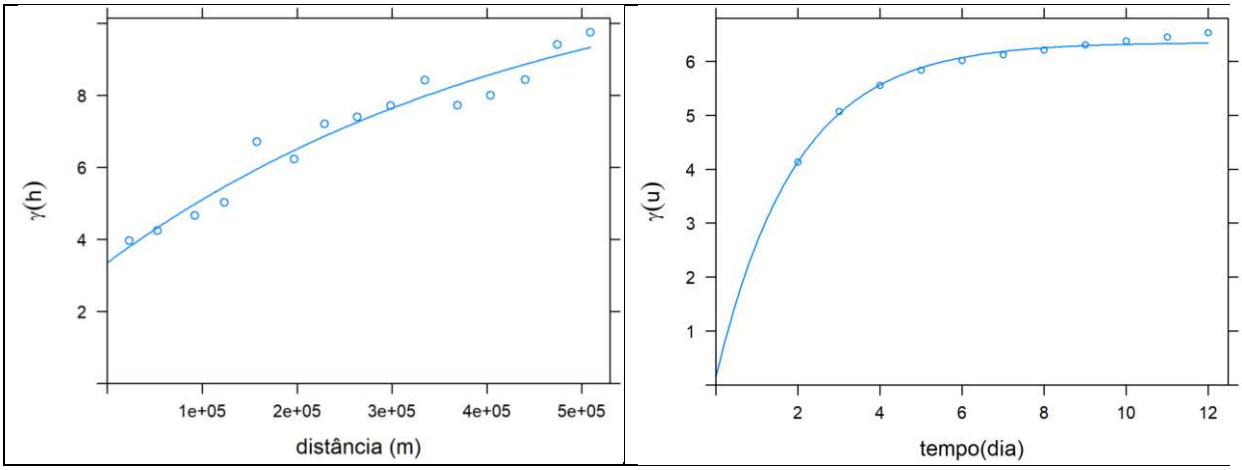
Com os modelos de aprendizagem de máquina foram geradas predições utilizando todas as covariáveis disponíveis no banco de dados utilizado e também predições empregando apenas as covariáveis longitude, latitude e altitude. O modelo de regressão linear múltipla e o support vector machine linear apresentaram resultados muito similares. Ao utilizar somente as três covariáveis foi possível calcular a temperatura máxima do ar em uma grade regular com resolução de 0,5ºx0,5º, cobrindo todo o estado de Minas Gerais e comparar o campo espacial

produzido pelos modelos estudados e com a reanálise do ECMWF. Os modelos regressão linear múltipla e support vector machine apresentaram desempenhos melhores do que o random forest e o modelo ponderado pelo Inverso da Distância. No entanto, em um dia em que ocorreu entrada de frente fria no estado de Minas Gerais, os modelos de regressão linear múltipla e o support vector machine tiveram performance bem inferior ao random forest e o modelo ponderado pelo Inverso da Distância. Mas, em geral, o modelo preditivo de aprendizagem de máquina supervisionado baseado no algoritmo support vector machine para regressão, com uso de funções lineares, foi o de melhor performance para predição da temperatura máxima do ar no estado de Minas Gerais em 2004.

Os campos produzidos pela geoestatística espaço-temporal apresentaram mais suavizados do que os encontrados com os modelos de aprendizagem de máquina e o IDW.

Com base nas considerações acima, pode-se concluir que este trabalho atingiu os objetivos propostos, disponibilizando para a comunidade científica metodologias para trabalhar com dados em pontos de estação meteorológica.

APÊNDICE



(a) (b)

Figura 1 – (a) Variograma puramente espacial com o modelo exponencial ajustado às semivariâncias estimadas. (b) Variograma puramente temporal com o modelo exponencial ajustado às semivariâncias estimadas.

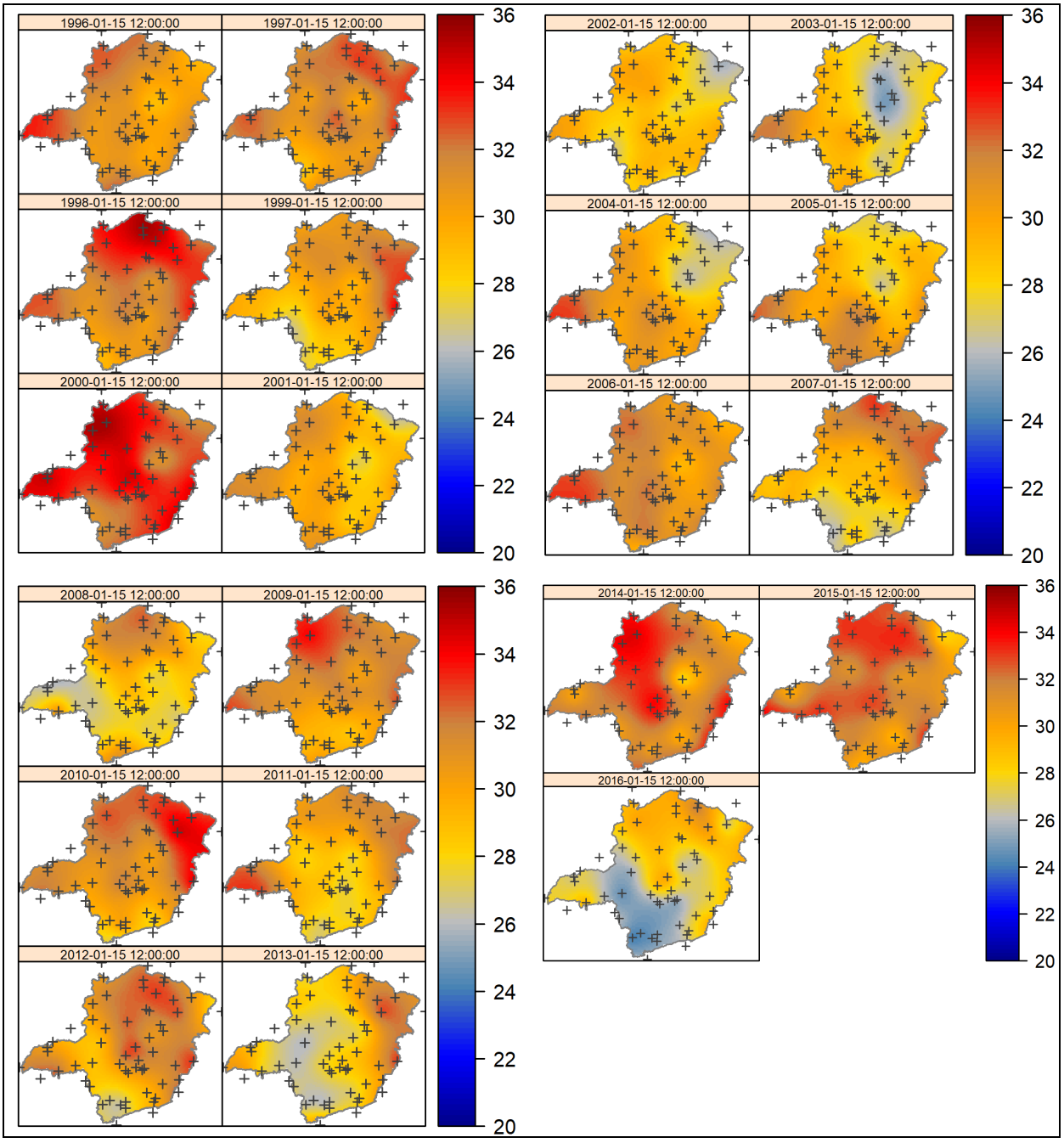


Figura 2 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de janeiro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

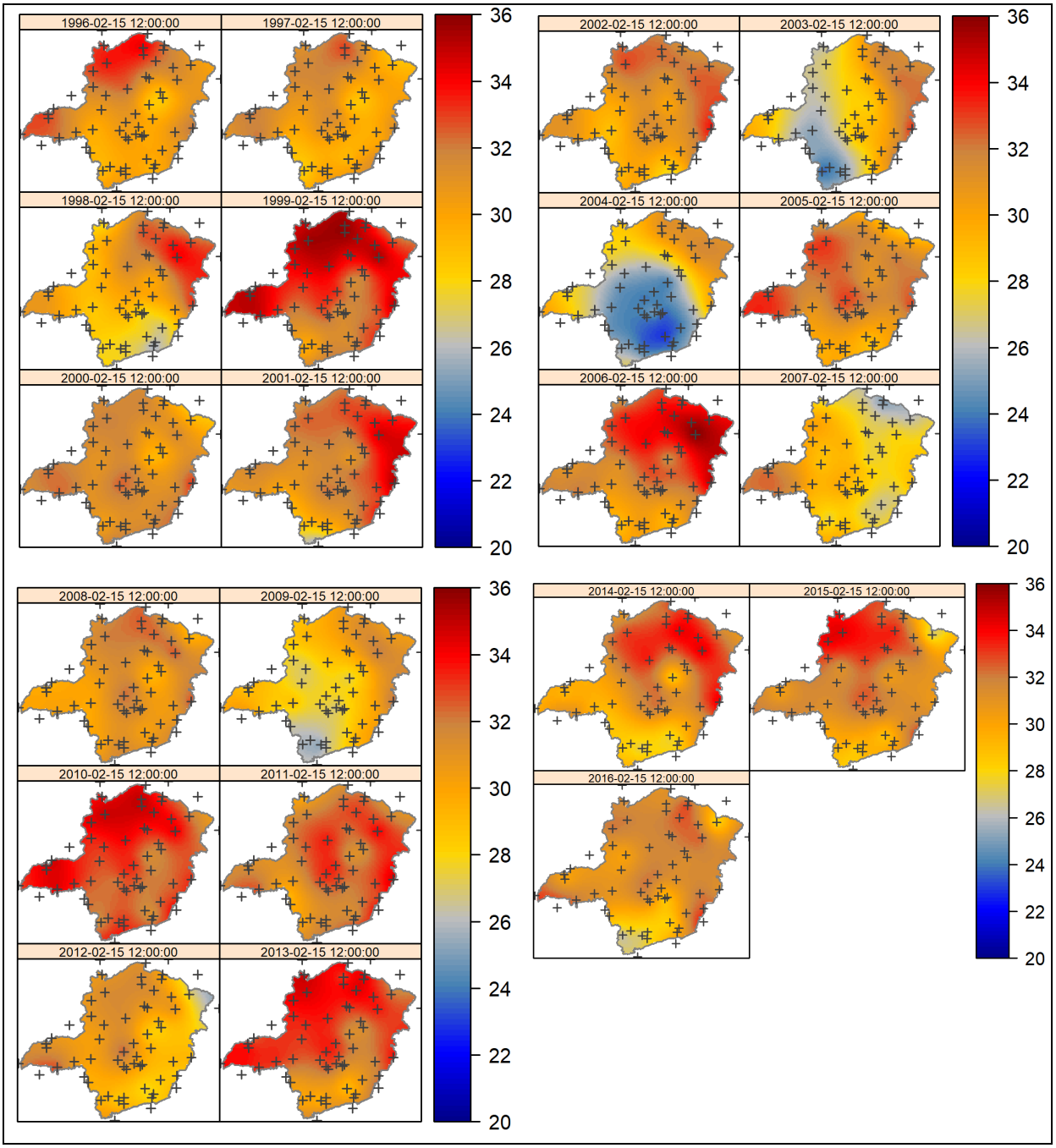


Figura 3 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de fevereiro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

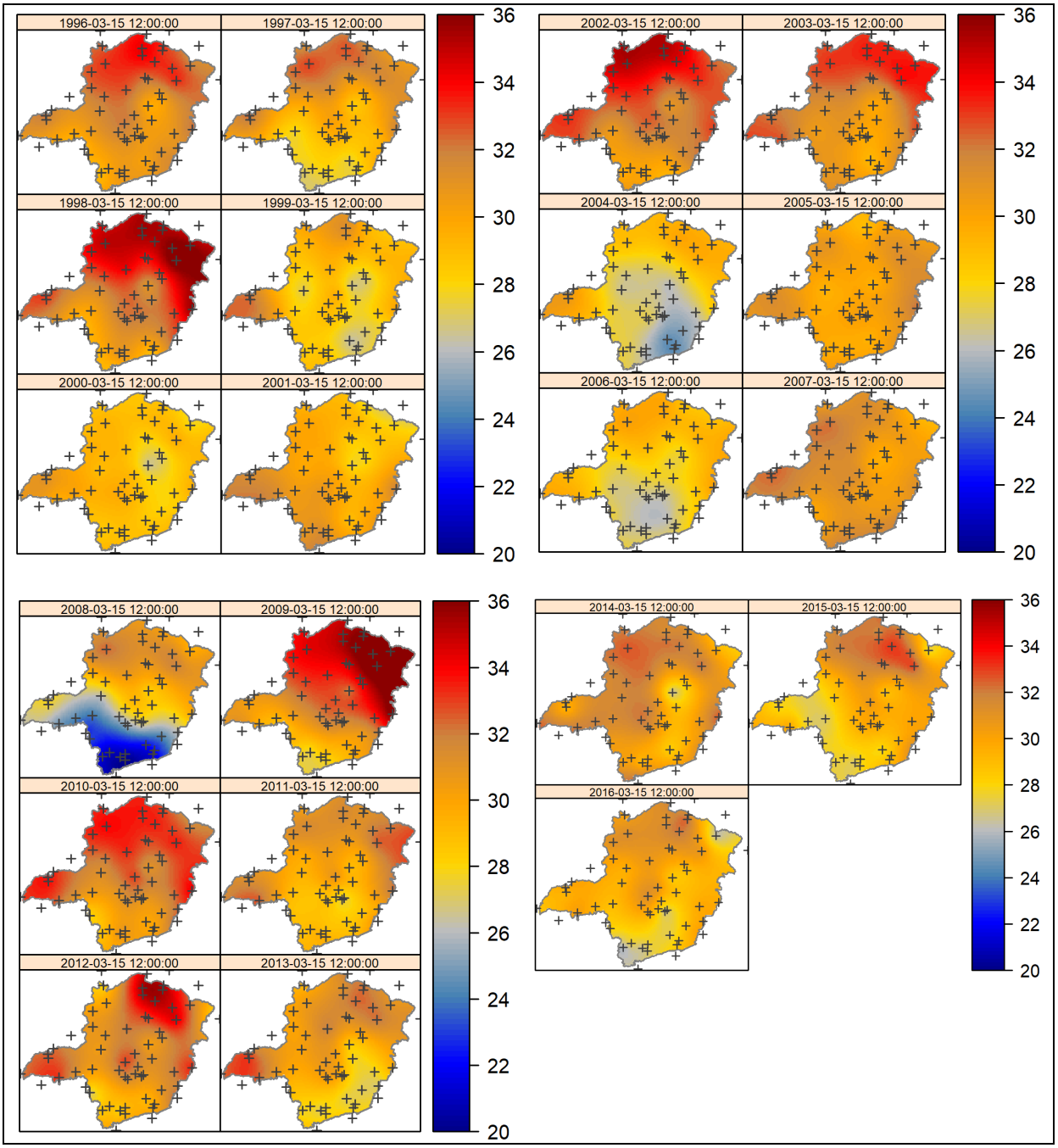


Figura 4 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de março, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

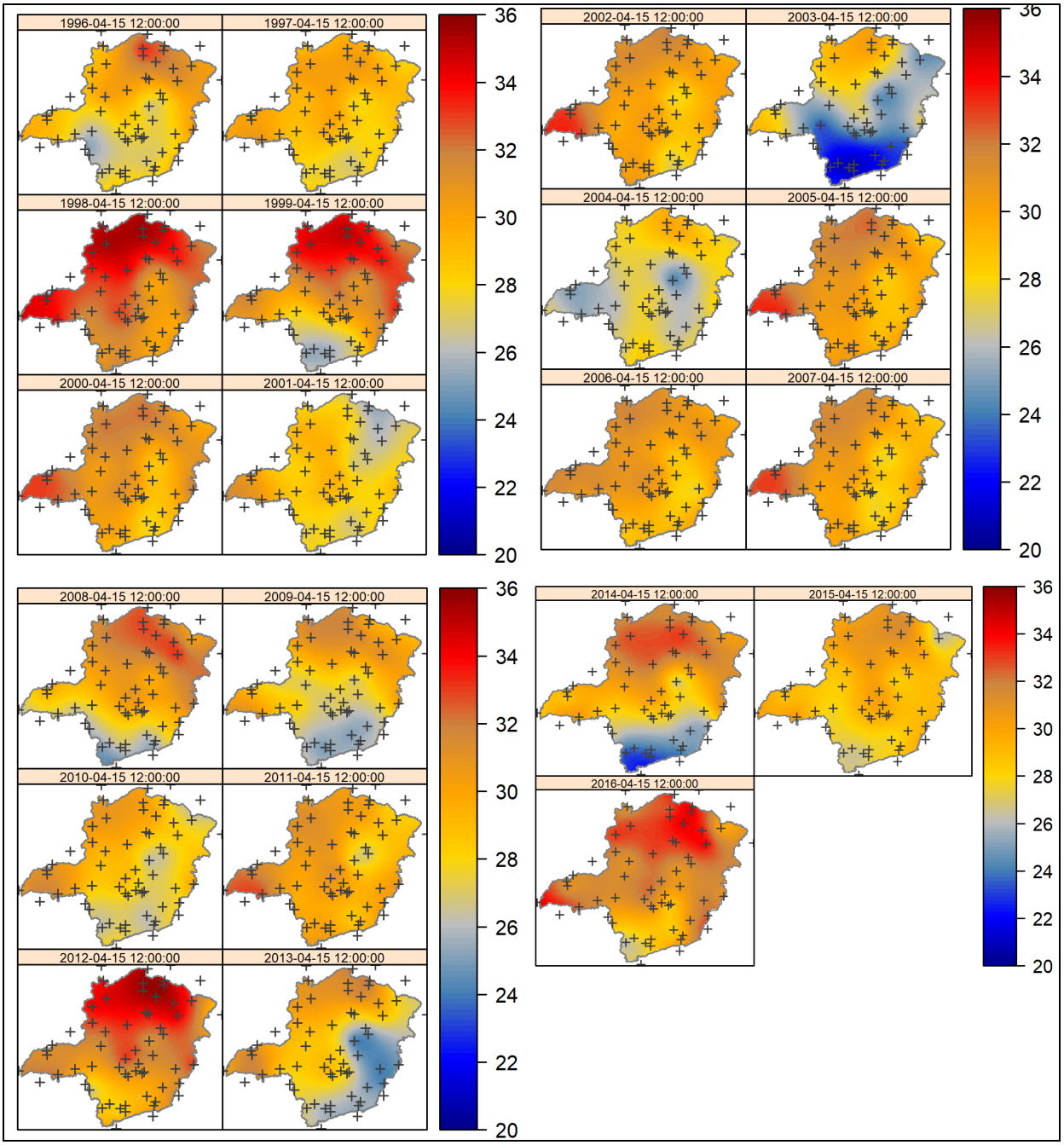


Figura 5 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de abril, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

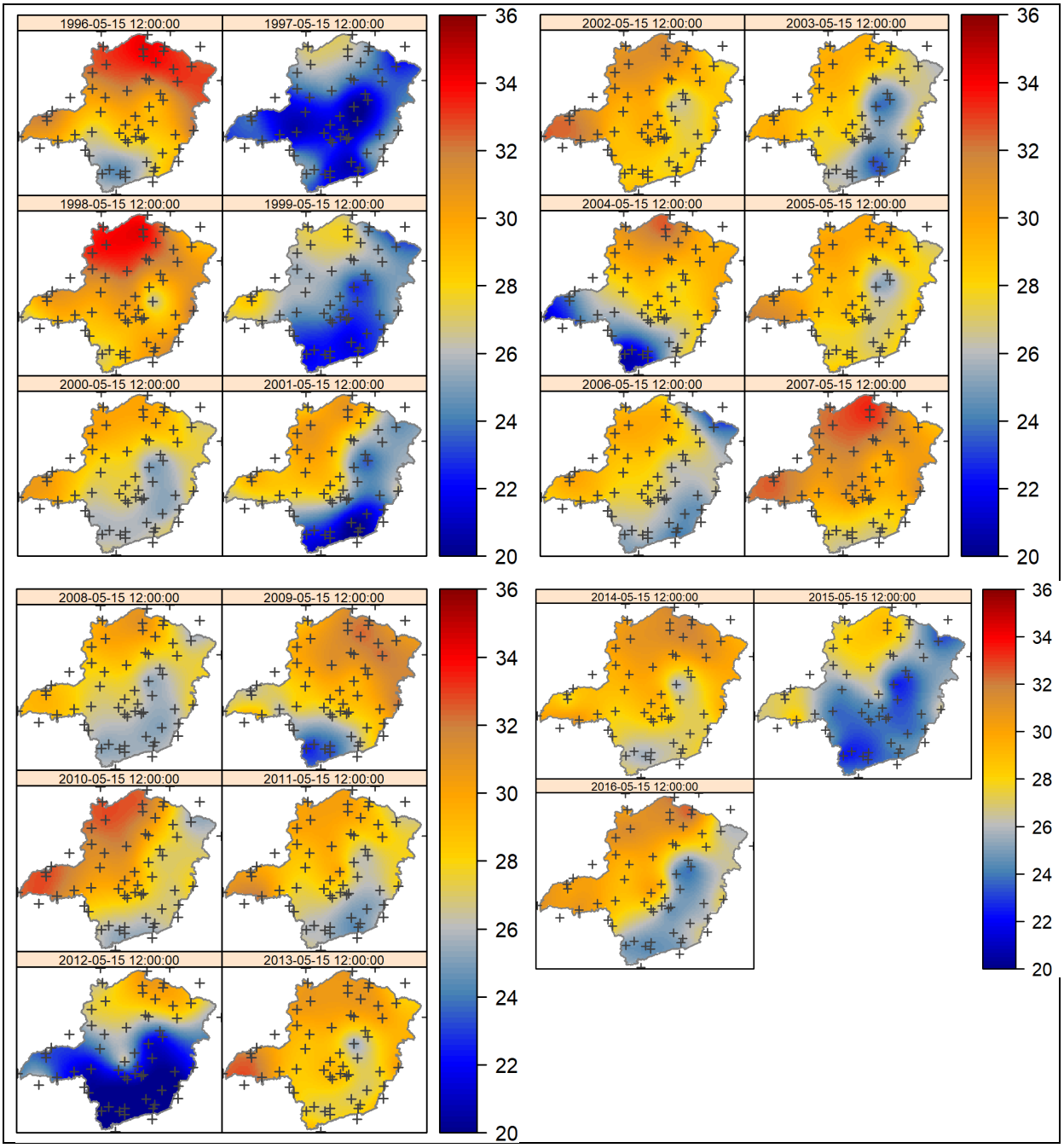


Figura 6 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de maio, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

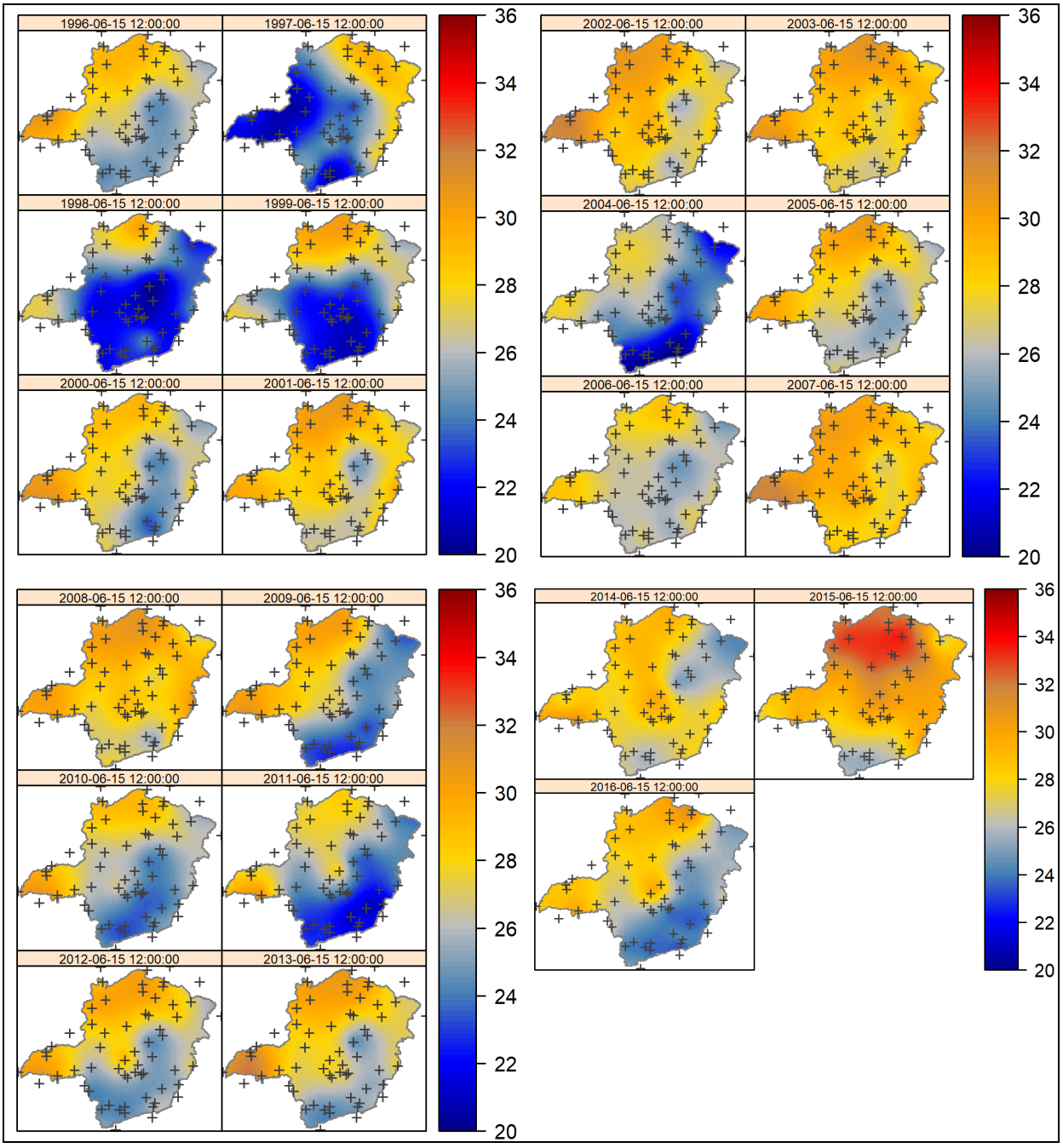


Figura 7 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de junho, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

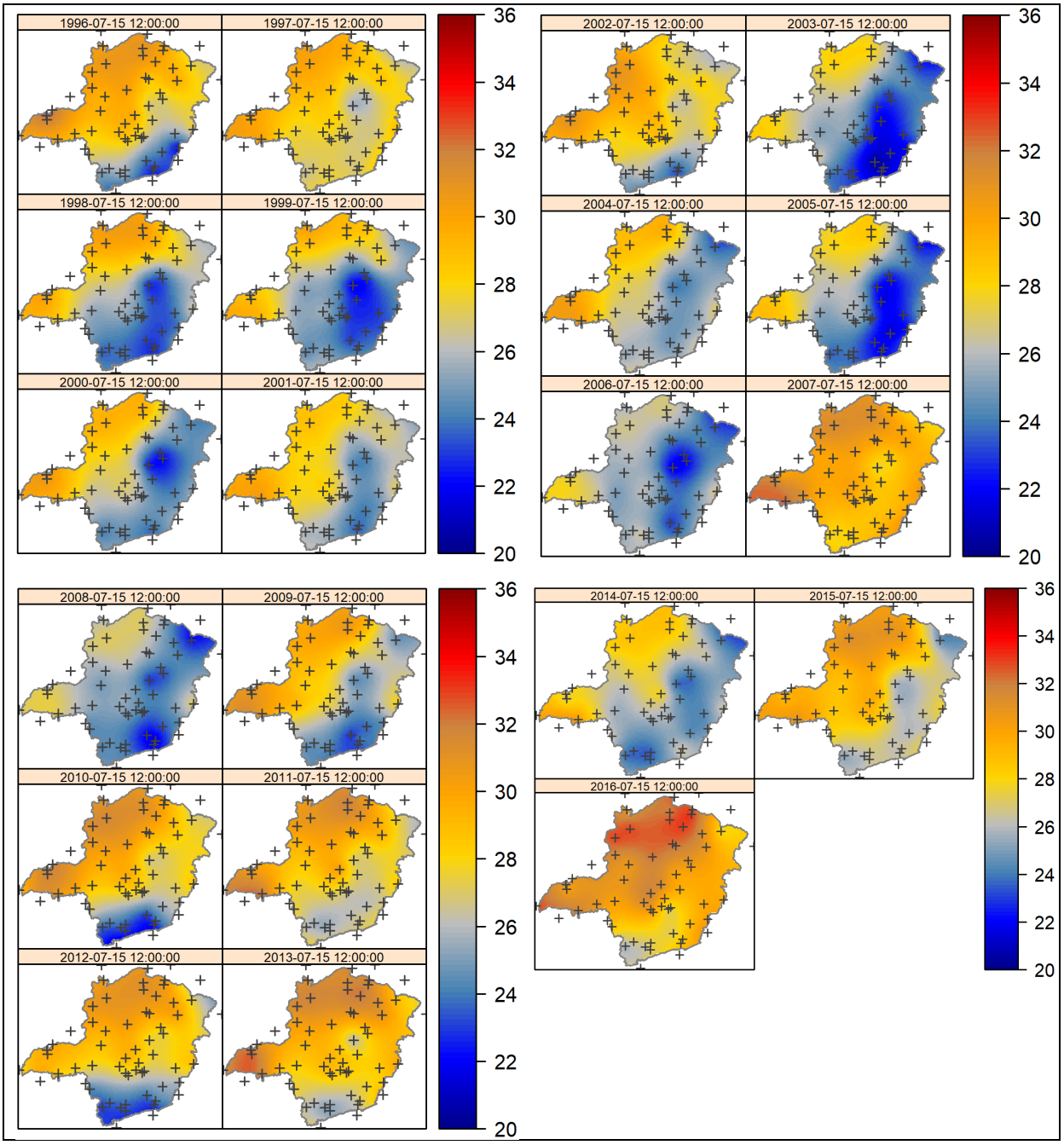


Figura 8 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de julho, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

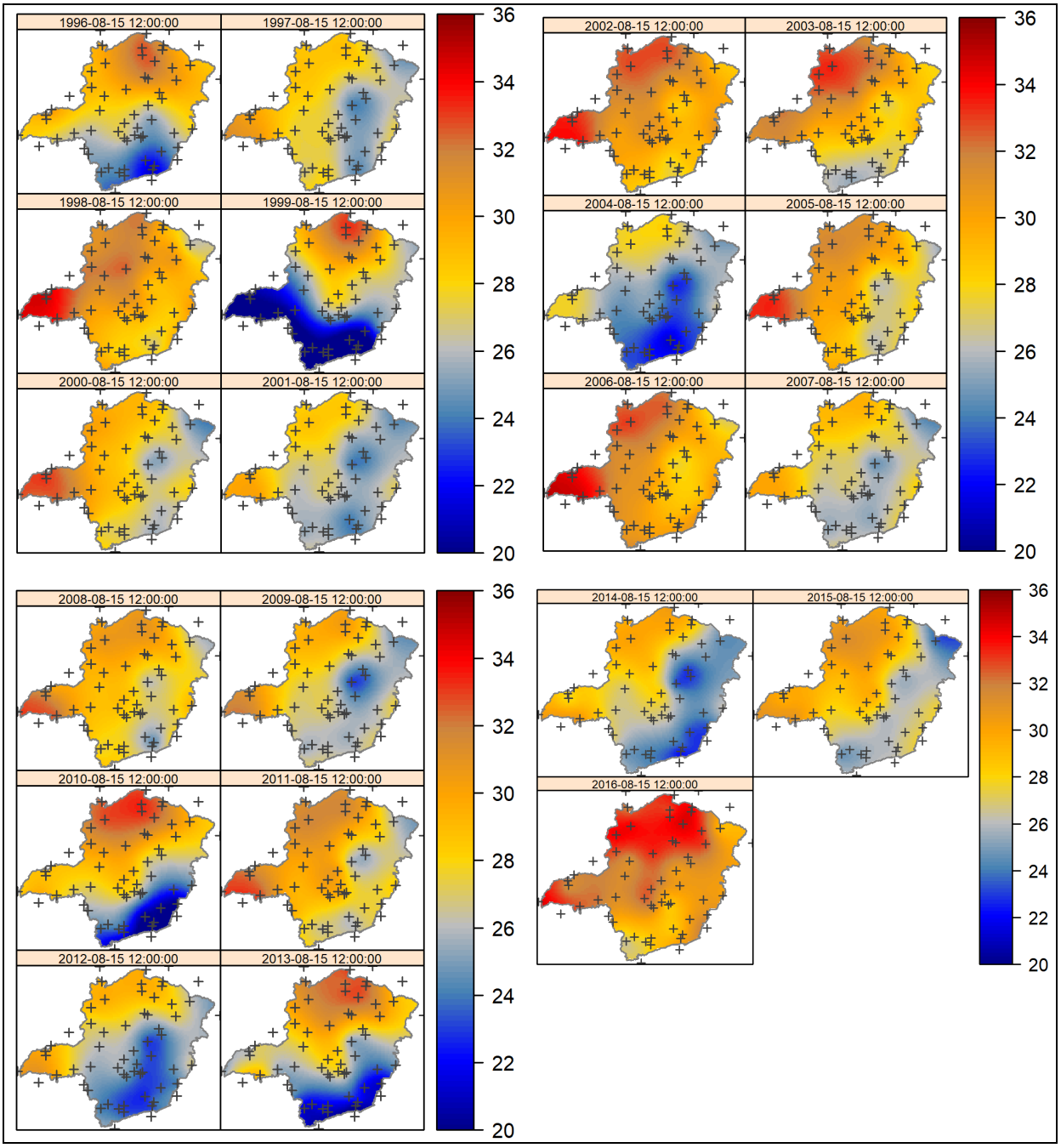


Figura 9 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de agosto, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

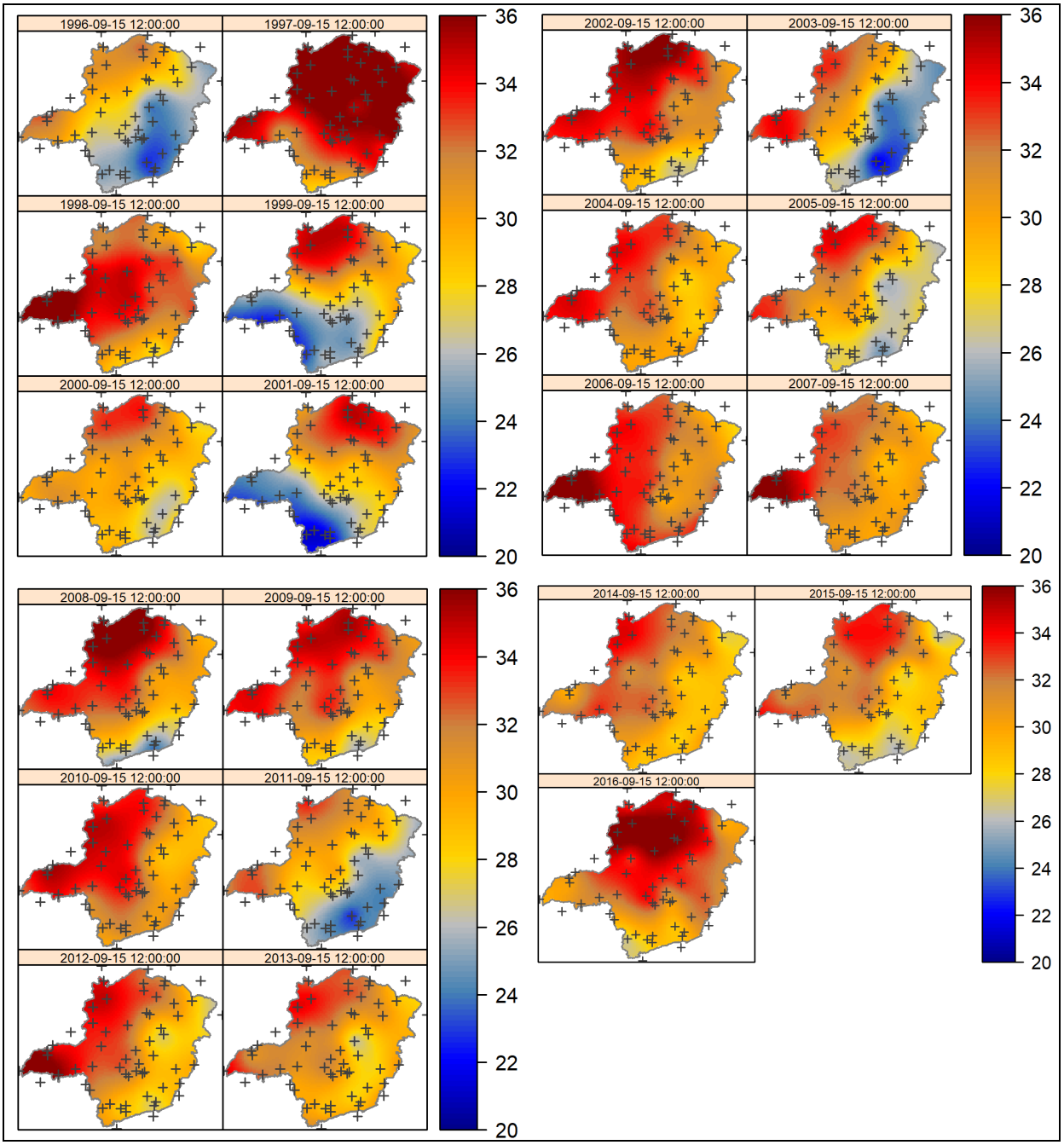


Figura 10 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de setembro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

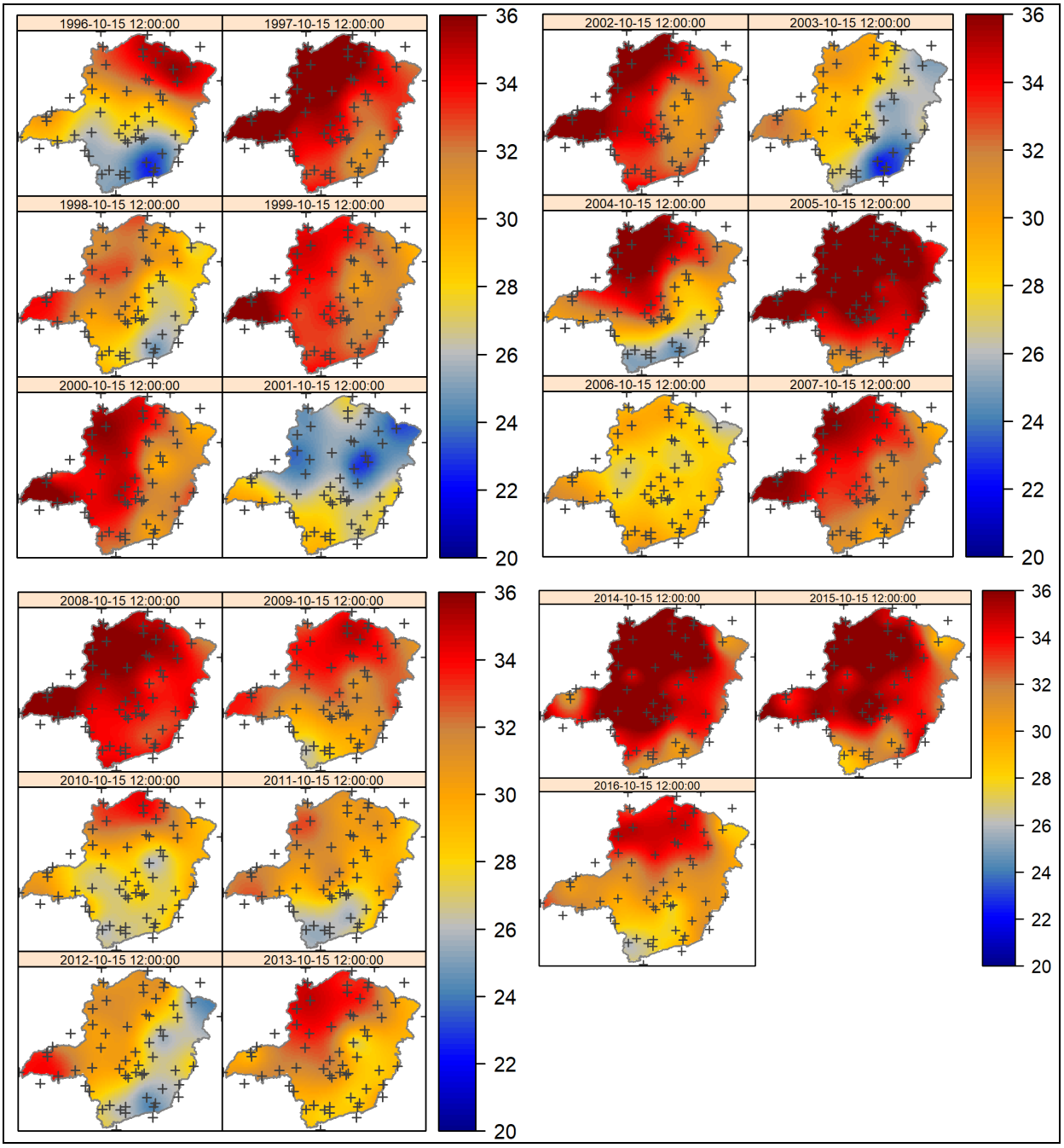


Figura 11 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de outubro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

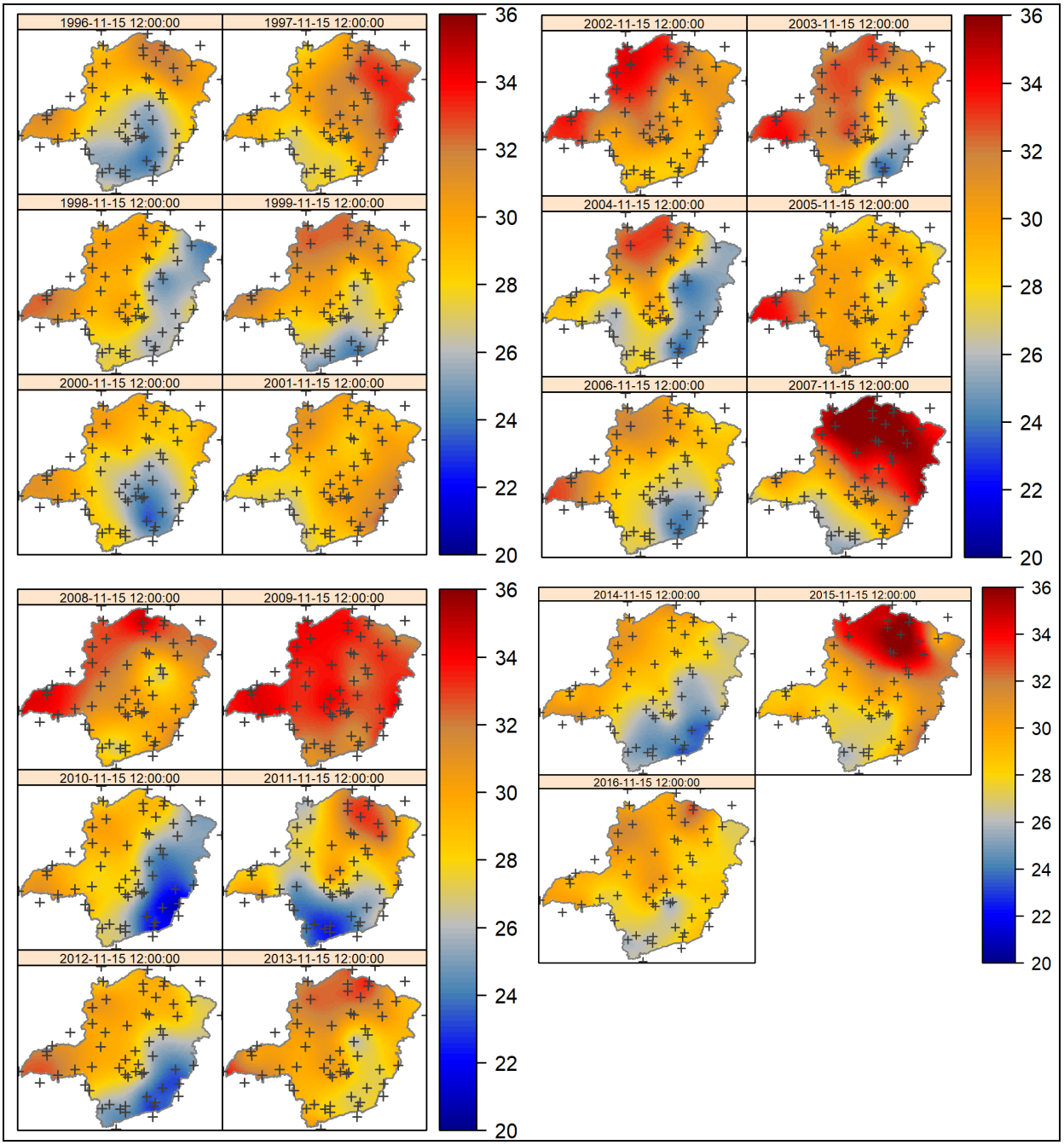


Figura 12 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de novembro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).

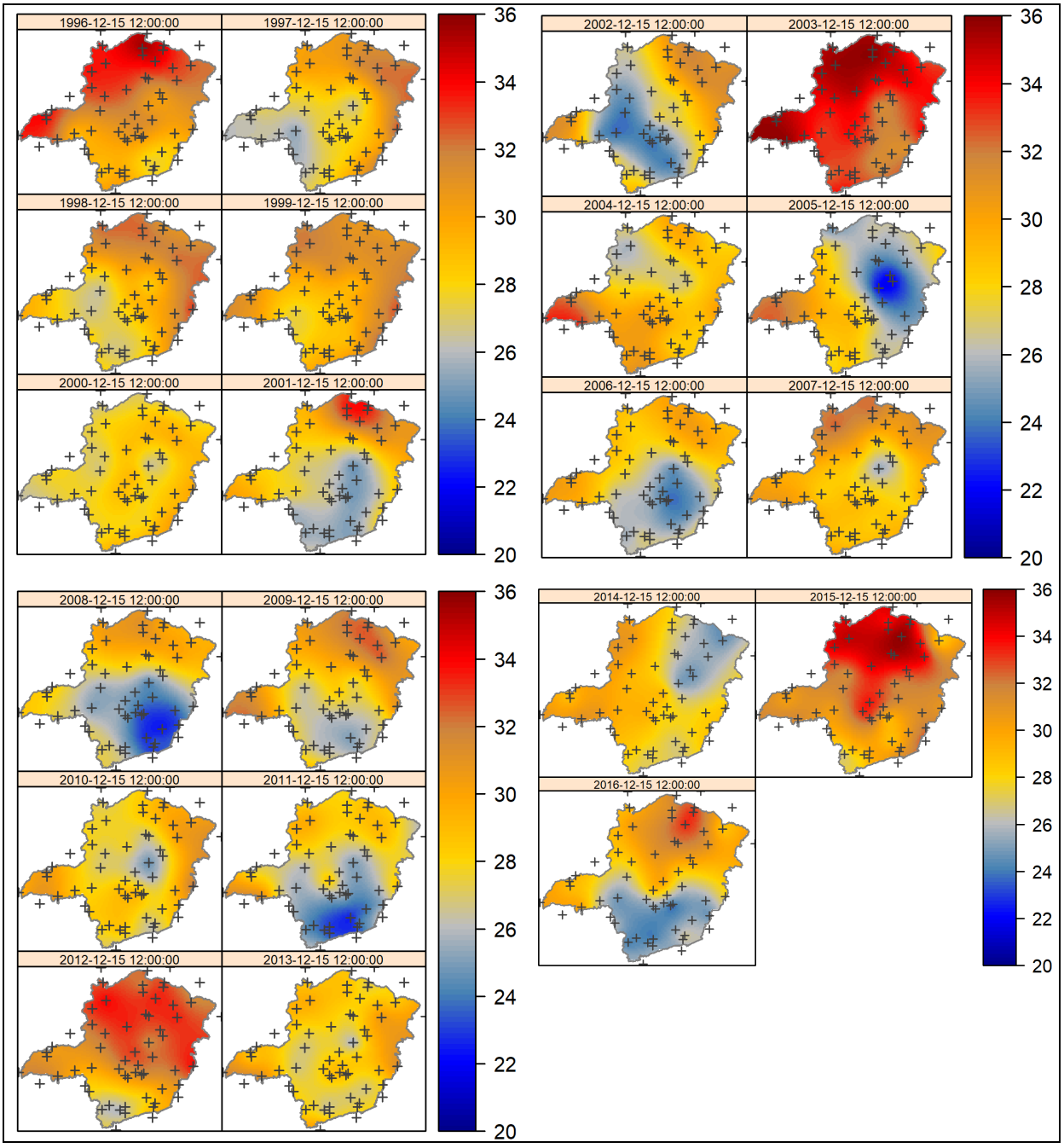


Figura 13 – Predição espaço-temporal dos valores de temperatura máxima do ar do estado de MG para o dia 15 de dezembro, no período de 1996 a 2016. As escalas indicam a variação da temperatura máxima do ar (°C).