Análise de séries climáticas temporais das previsões sazonais do modelo Eta

RELATÓRIO DAS ATIVIDADES DESENVOLVIDAS DE INICIAÇÃO CIENTÍFICA
(PIBIC/CNPq/INPE)
Período: agosto 2017 -julho 2018

Caio Bastos Iracema (Universidade do Estado do Rio de Janeiro, Bolsista PIBIC/CNPq)
E-mail: caio-b-iracema@hotmail.com
Chou Sin Chan (CPTEC/DMD, Orientador)
E-mail: chou.sinchan@cptec.inpe.br

COLABORADOR
Prof Dr. Michel Pompeu Tcheou
(Universidade do Estado do Rio de Janeiro)
E-mail: mtcheou@uerj.br

Sumário

Listas de figuras e tabelas

Resumo

O modelo Eta/INPE é um modelo atmosférico, estado da arte baseado em equações de conservação de massa, energia e momentum e é utilizado pelo CPTEC para produzir previsões meteorológicas em diferentes prazos de antecedência e em diferentes resoluções espaciais. Contudo, o estado real da atmosfera e o previsto através de modelagem numérica tende a ter desigualdades, interferindo no tempo de integração e, portanto, gerando erros. O objetivo deste trabalho é desenvolver um arcabouço para análise e ajustes dos dados de previsão numérica produzidos pelo modelo Eta, através da correção de padrões espectrais das previsões sazonais em diferentes resoluções espaciais. Para isso é necessário que se realize um levantamento das estatísticas a respeito dos desvios entre os dados do Eta em relação aos dados observacionais e que se conheça as regiões do domínio de estudo que apresentam alto graus similaridade estatística.

1 Introdução

O modelo Eta/INPE é um modelo atmosférico, estado da arte baseado em equações de conservação de massa, energia e momentum[1-8]. O modelo representa os principais processos atmosféricos que incluem a geração de nuvens e chuva, a turbulência atmosférica, os processos de transferência radiativa na atmosfera pelas ondas curtas e longas, os processos de interação entre a atmosfera-vegetação-solo e interação entre atmosfera e oceano, etc. O modelo Eta/INPE é utilizado pelo CPTEC para produzir operacionalmente previsões meteorológicas em diferentes prazos de antecedência e em diferentes resoluções espaciais, desde o horizonte de 3, 7, 11 dias até 4,5 meses, nas resoluções de 5, 15 e 40 km. São fornecidas as seguintes variáveis prognósticas: componentes zonal e meridional do vento, temperatura do ar, umidade e temperatura do solo, água líquida ou gelo das nuvens, umidade específica, pressão à superfície e energia cinética turbulenta. Dados atmosféricos oriundos de modelos numéricos são por si volumétricos; temos a resolução sobre a superfície terrestre e em função da altura de forma a fornecer células atmosféricas em função de latitude, longitude e altitude para as quais obtêm-se as variáveis prognósticas. Além disso, essas variáveis são fornecidas para um intervalo de tempo, em geral fixo de algumas horas. Temos assim dados volumétricos discretos no tempo.

O estado real da atmosfera e o previsto através de modelagem numérica tende a ter desigualdades, interferindo no tempo de integração e, portanto, gerando erros[9,10]. Os erros podem ser definidos em dois tipos: i) erros devidos às aproximações físicas do modelo numérico e ii) erros nas condições iniciais, que aumentam com o tempo de integração do modelo, devido as não-linearidades nas equações do mesmo.

Neste trabalho, obtêm-se inicialmente séries temporais de desvios entre os dados observacionais e os dados produzidos pelo modelo Eta. Em seguida, calculam-se as estatísticas desses desvios através das métricas de média, variância, simetria e curtose. Enfim, agrupam-se essas séries através do algoritmo K-means em regiões geográficas que possuam alguma similaridade estatística entre si. É esperado que esse agrupamento permita melhor desempenho de métodos de ajuste do padrão espectral das variáveis prognósticas de modelos de previsão numérica, como o Eta.

O período do conjunto de previsão considerado é de 2008, compreendendo intervalo de previsões entre 13 de dezembro a 30 de abril do ano seguinte. Os dados de previsão numérica são produzidos pelo modelo Eta de 40 km, e os dados observacionais correspondem aos dados de reanálise no NCEP de 38 km[11]. As séries utilizadas apresentam 139 dias de previsão de horizonte sazonal com resolução temporal de seis horas, portanto, por dia há quatro valores de previsão (às 00:00, 06:00, 12:00 e 18:00 UTC). As variáveis prognósticas analisadas são as componentes zonal e meridional do vento(em m/s), altura geopotencial (em mgp - metro geopotencial) e a umidade específica (em kg de massa de vapor d'água por kg de massa de ar). Além disso, o nível de pressão atmosférica cujos resultados são apresentados neste relatório corresponde ao nível de 300hPa. Porém, as análises também compreendem os níveis de pressão atmosférica de 250hPa, 350hPa, 500 hPa e 800 hPa.

2 Objetivo

O objetivo deste trabalho é desenvolver um arcabouço para análise e ajustes dos dados de previsão numérica produzidos pelo modelo Eta, através da correção de padrões espectrais das previsões sazonais em diferentes resoluções espaciais. Para isso é necessário que se realize um levantamento das estatísticas a respeito dos desvios entre os dados do Eta em relação aos dados observacionais e que se conheça as regiões do domínio de estudo que apresentam alto grau similaridade estatística entre as séries de desvios.

3 Fundamentação teórica

3.1 Métricas Estatíticas

Neste trabalho, utilizam-se as métricas de média, variância, assimetria (*skewness*) e curtose (*kurtosis*), referentes ao erro entres as séries temporais do vento meridional, vento zonal, altura geopotencial e umidade específica produzidas pelo Eta em comparação aos

dados observados do NCEP das mesmas variáveis para diferentes alturas, discernidas em hPa. As métricas em questão estão descritas na Tabela 1.

Tabela 1: As quatro estatísticas que descrevem uma distribuição.

Média	$\mu = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n-1}}$
Variância	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{(n-1)}$
Assimetria	$Skewness = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3(n-1)}$
Curtose	$Kurtosis = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4(n-1)}$

Através dessas métricas, referenciadas em latitude e longitude, será possível utilizar o método de agregação (ou clustering) conhecido como o algoritmo de K-means. Isso permite melhor desempenho de ajuste e de correção dos erros de previsão numérica de modelos regionais através de filtragem adaptativa, por exemplo.

3.2 K-means

O K-means permite classificar as informações por meio de comparações entre os valores numéricos dos dados, sem a necessidade de nenhuma supervisão humana. Para realizar a classificação, uma função objetivo avalia a qualidade do particionamento de modo que os objetos dentro de um grupo sejam semelhantes entre si, mas diferentes de objetos de outros grupos. Ou seja, a função objetivo visa uma alta similaridade dentro dos clusters e uma baixa similaridade entre os clusters[12]. Assim, esse critério tem por objetivo separar as regiões em função do comportamento dos dados. Seja um conjunto de dados D, contendo q objetos. O k-means particiona os objetos em função de suas distâncias aos centroides dos clusters. O algoritmo k-means distribui os objetos de D em k clusters, C1, ..., Ck,C $\subset$ D, Ci $\cap$ Cj = $\emptyset$ para (1 $\leq$ i,j $\leq$ k). O centroide pode ser definido como a média dos objetos (ou pontos atribuídos ao cluster). A distância entre o objeto v $\in$ Ci e ci como centro de massa ou centroide do cluster é medida através de dist(ci, v), onde ci = (cx, cy) e vi = (vx, vy). Por exemplo, dist(ci, v) pode ser a distância euclidiana entre o objeto v, e o centroide ci, que é dada por

$dist(ci, v) = ((cx - vx)^2 + (cy - vy)^2)^{1/2}$

onde cx, vx, cy, vy são os elementos respectivamente dos vetores v e ci. Assim, a função objetivo é dada como a soma dos erros para todos os objetos no conjunto de dados [12]

$$E = \sum_{i=1}^{k} \quad \sum_{ci \in Ci; v \in D} \mathrm{dist}(ci, v)^2$$

o parâmetro k é o número de clusters, v é um ponto no espaço que representa um dado objeto, e ci é o centroide de um cluster Ci (tanto v e ci são multidimensionais) e dist(ci, v) é a distância entre o objeto v, e o centroide mais próximo a ele, ci. Inicialmente escolhe-se aleatoriamente k distintos objetos em D para representar os centros dos clusters. Depois associa-se cada objeto de D ao cluster de centro mais próximo, com base na distância euclidiana. Interativamente reduz-se E em cada cluster. Para cada cluster, calcula-se um novo centroide usando os objetos atribuídos ao cluster na iteração anterior. Todos os objetos são, então, redistribuídos utilizando os centroides atualizados. As iterações continuam até que a redistribuição fique estável, isto é, os clusters formados na rodada atual são os mesmos formados na rodada anterior ou até que uma quantidade de iterações pré-determinada seja realizada[12]. Neste trabalho definiu-se previamente um total de 200 iterações. Diferentemente dos outros critérios de seleção de regiões que separam as regiões pela sua forma, o k-means separa as regiões com base no comportamento dos dados, podendo ser utilizado para agrupar regiões referente a séries climáticas que apresentam comportamento estatístico semelhantes. Agrupando assim os dados da região estudada em função da evolução da(s) grandeza(s) escolhida(s). Neste trabalho as 4 grandezas que descrevem as distribuições (média, variância, curtose e assimetria) das séries foram usadas, para cada variável estudada (vento meridional, vento zonal, altura geopotencial e umidade específica).

O número de clusters (k) é pré-determinado e deve ser escolhido de forma que maximize a inércia entre os clusters (BSS) e minimize a inércia dentro dos clusters (WSS). Existem vários métodos para se determinar o número de clusters e o que foi adotado neste trabalho é bem simples.

Quanto mais se aumenta o número de clusters maior fica a inércia entre os eles, porém este aumento fica cada vez menor e este método consiste em escolher um valor n de k tal que o valor de BSS quando k = n+1 é irrelevantemente maior que o valor de BSS quando k = n. Geralmente se admite um valor n de k tal que o valor de BSS é, no mínimo, algo próximo de 70% da variância total.

4 Materiais e Métodos Utilizados

A metodologia empregada neste trabalho é composta pelas seguintes etapas:

1. Inicialmente, realiza-se uma interpolação bidimensional do Eta para ajustar a grade em relação ao NCEP. A grade do primeiro era de 40km e a do segundo de 38 km.
2. Em seguida, calculam-se os erros entre as séries temporais das variáveis prognósticas do Eta e NCEP.
3. Calculam-se as métricas estatísticas média, variância, assimetria e curtose das séries temporais dos erros, onde cada série está associada a uma coordenada geográfica. São estas métricas que darão forma aos agrupamentos que serão feitos na etapa seguinte.
4. Aqui é feito o agrupamento das séries temporais dos erros de cada variável através do algoritmo k-means a partir desses parâmetros estatísticos.
5. Por fim, a correção dos desvios no modelo Eta através de ajustes dos padrões espectrais por filtragem adaptativa por região agrupada.

5 Análise de resultados

O período considerado é o do ano de 2008. As previsões do modelo Eta 40km e as reanálises do NCEP apresentam resolução temporal de seis horas, com isso, para cada dia há quatro saídas de previsões e com isso para cada coordenada há 553 previsões. Os parâmetros climáticos considerados nos testes foram o vento meridional, vento zonal, umidade especifica e a altura geopotencial e a região estudada engloba a América do Sul e América Central.

5.1 Vento Meridional

A figura 1 é um gráfico de dispersão em que no eixo X estão os possíveis valores de k (número de clusters de 2 a 20) enquanto que no eixo Y estão, devidamente calculadas, as proporções (%) dos respectivos BSS em relação a variância total. É possível notar que o aumento do BSS já começa a ficar pequeno a partir de k=4 para k=5 mas somente quando k=10 é que o BSS assume um valor aceitável (igual ou próximo a 70% no mínimo).

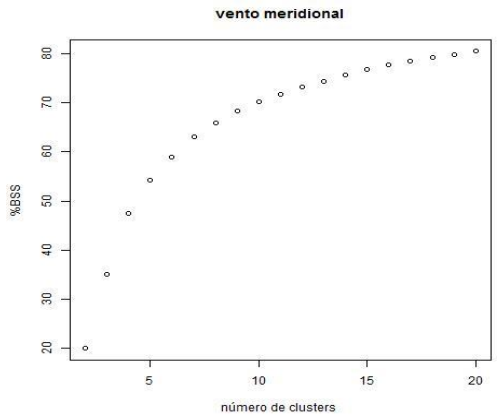

Figura 1: Escolhendo um número adequado para k (Vento Meridional).

Na figura 2, é apresentado o mapa dividido em regiões (10 clusters) de acordo com as informações de média, variância, assimetria e curtose do erro entre a variável prognóstica de vento meridional produzida pelo Eta em relação ao observado registrado pelo NCEP ao nível de pressão de 300hPa. O agrupamento resultou em uma concentração na parte central do mapa dos clusters de menores variâncias das series temporais, no geral. Os grupos com as maiores variâncias tendem a ficar, principalmente, no norte do mapa, embora marquem presença no Sul também. O cluster 9, talvez o mais compacto entre os grupos, é o que possui as menores médias de séries. A maioria dos clusters possuem uma mediana das curtoses das séries em torno de 3 (valor encontrado em uma distribuição normal). Os clusters 1 e 4, os que mais se espalham pelo mapa, são os que possuem as maiores medidas de curtose das séries (significa que é mais fácil nesses grupos obter valores que não se aproximam da média a vários múltiplos do desvio padrão). Alguns outliers do cluster 4 possuem assimetria superior a 1 (ou seja, estas séries temporais possuem uma assimetria notadamente positiva). Como muitas das medidas de assimetria variam de -0,5 a 0,5 não necessariamente são assimetrias fortes (positivas ou negativas) e talvez tendam a normalidade. Alguns outliers do cluster 4 possuem medida de assimetria superior a 1 (assimetria notadamente positiva).

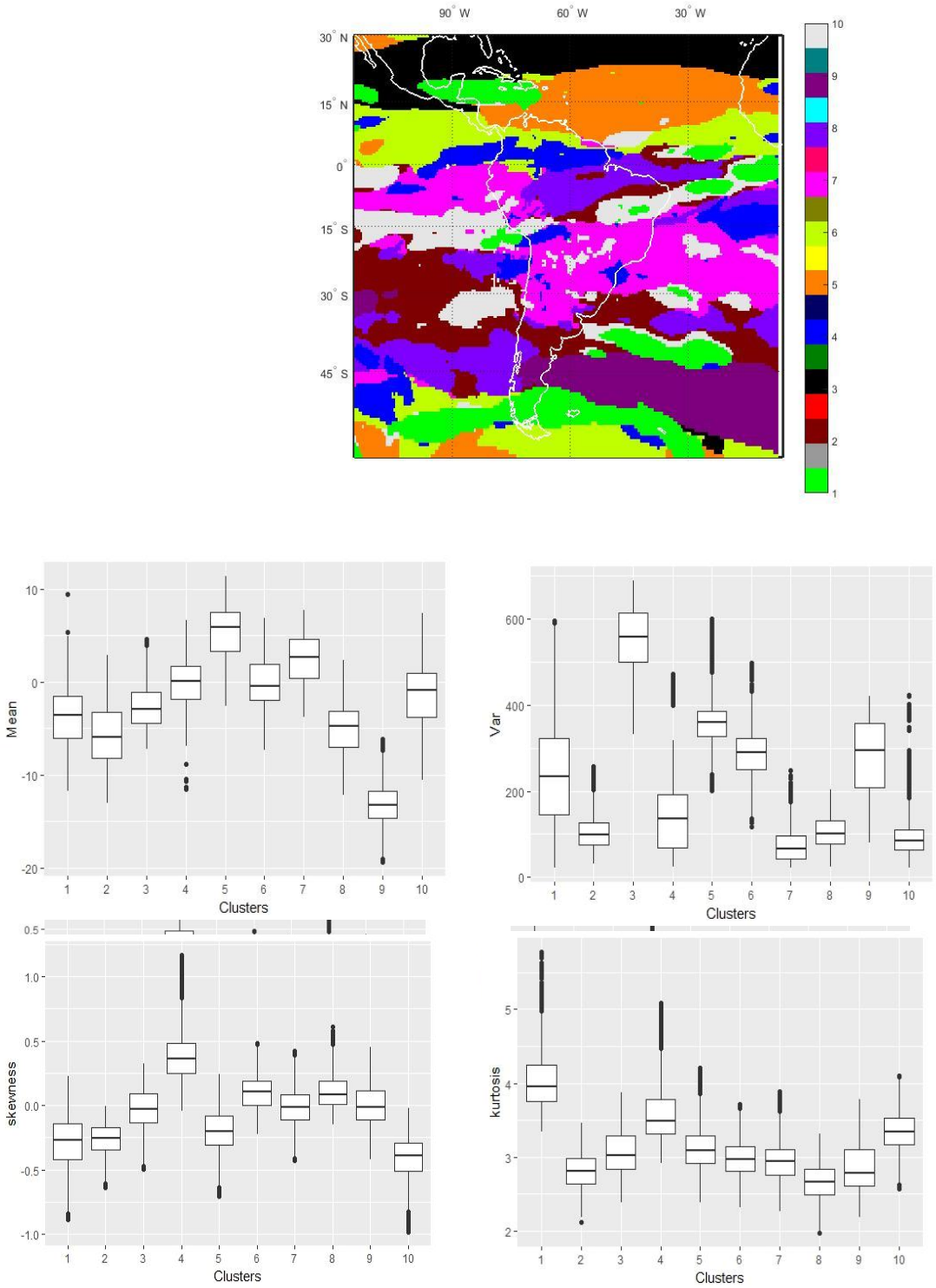

Figura 2: Agrupamento pelo K-means para a variável vento meridional a 300 hPa no ano de 2008; Boxplot das medidas de médias, variâncias, assimetrias e curtoses das séries temporais em cada cluster.

5.2 Vento Zonal

Na figura 3 é possível notar que o aumento do BSS já começa a ficar pequeno a partir de k=4 para k=5 mas somente quando k=9 é que o BSS assume um valor aceitável.

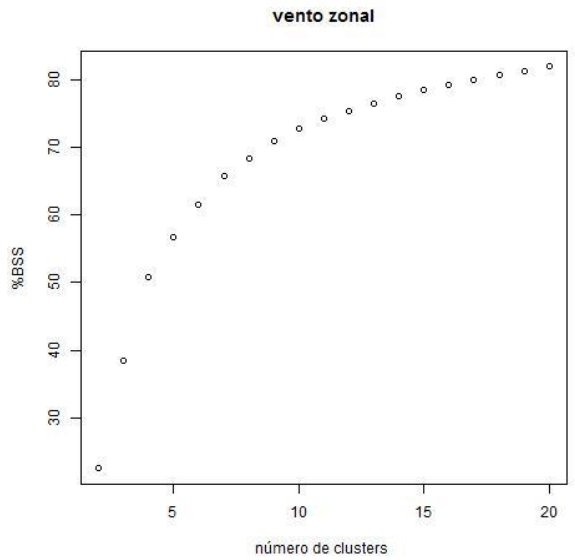

Figura 3: Escolhendo um número adequado para k (Vento Zonal).

Na figura 4, é apresentado o mapa dividido em regiões (9 clusters) de acordo com as informações de média, variância, assimetria e curtose do erro entre a variável prognóstica de vento zonal produzida pelo Eta em relação ao observado registrado pelo NCEP ao nível de pressão de 300hPa. É possível perceber novamente o fenômeno das variâncias (clusters com menores medidas de variância no centro e os com as maiores no sul e, principalmente, norte). Dos 4 grupos que não se concentram no centro do mapa (1, 6, 7 e 9), 2 deles possuem as maiores medidas de médias (1 e 6), além do cluster 9 que possui as menores medidas de médias das séries. As medidas de curtose das séries são , no geral, maiores que as das séries do vento meridional, porém ainda são próximas de 3, sendo as medidas do clusters 3 e 5 as que mais se afastam desse valor (são 2 grupos bem espalhados pelo mapa). Novamente, as medidas de assimetria, no geral, se concentram entre -0,5 e 0,5.

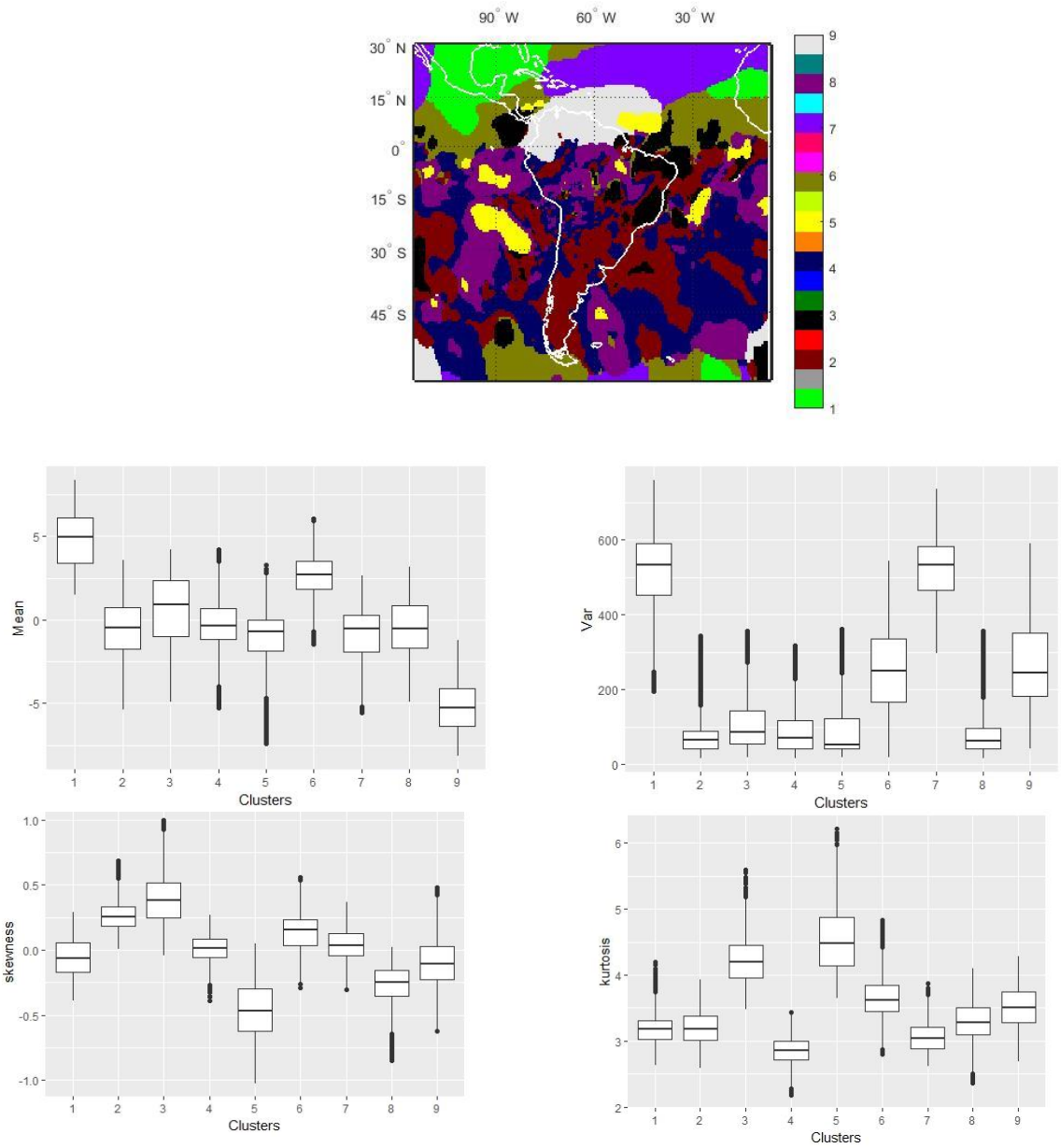

Figura 4: Agrupamento pelo K-means para a variável vento zonal a 300 hPa no ano de 2008; Boxplot das medidas de médias, variâncias, assimetrias e curtoses das séries temporais em cada cluster.

5.3 Altura Geopotencial

Na figura 5 é possível notar que o aumento do BSS já começa a ficar pequeno a partir de k=5 para k=6 mas o valor k=7 foi escolhido por estar acima de 70% (neste caso, como k=6 estava bem próximo de 70% também poderia ter sido escolhido).

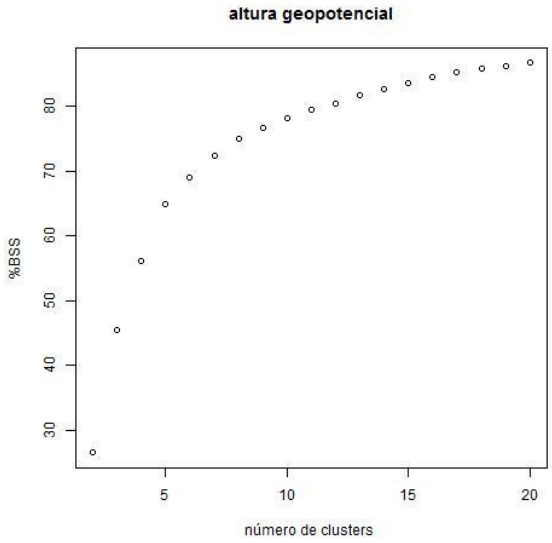

Figura 5: Escolhendo um número adequado para k (Altura Geopotencial).

Na figura 6 é apresentado o mapa dividido em regiões de acordo com as informações de média, variância, assimetria e curtose do erro entre a variável prognóstica de altura geopotencial produzida pelo Eta em relação ao observado registrado pelo NCEP ao nível de pressão de 300hPa. De novo, as variâncias causam o mesmo efeito que nos 2 casos anteriores. Os clusters 1 e 6 (que possuem as maiores variâncias) possuem as menores e maiores médias, respectivamente. Assim como no caso anterior, há uma certa tendência das medidas de curtose serem maiores que 3 mas somente nos clusters 5 e 7 há uma quantidade significativa de medidas notadamente superiores a este valor (são clusters bem fragmentados pelo mapa). Novamente as medidas de assimetria são, no geral, próximas de 0. Porém, no cluster 5 alguns valores são maiores que 1, o que aponta uma forte assimetria positiva, enquanto que no cluster 7 alguns outliers possuem forte assimetria negativa.

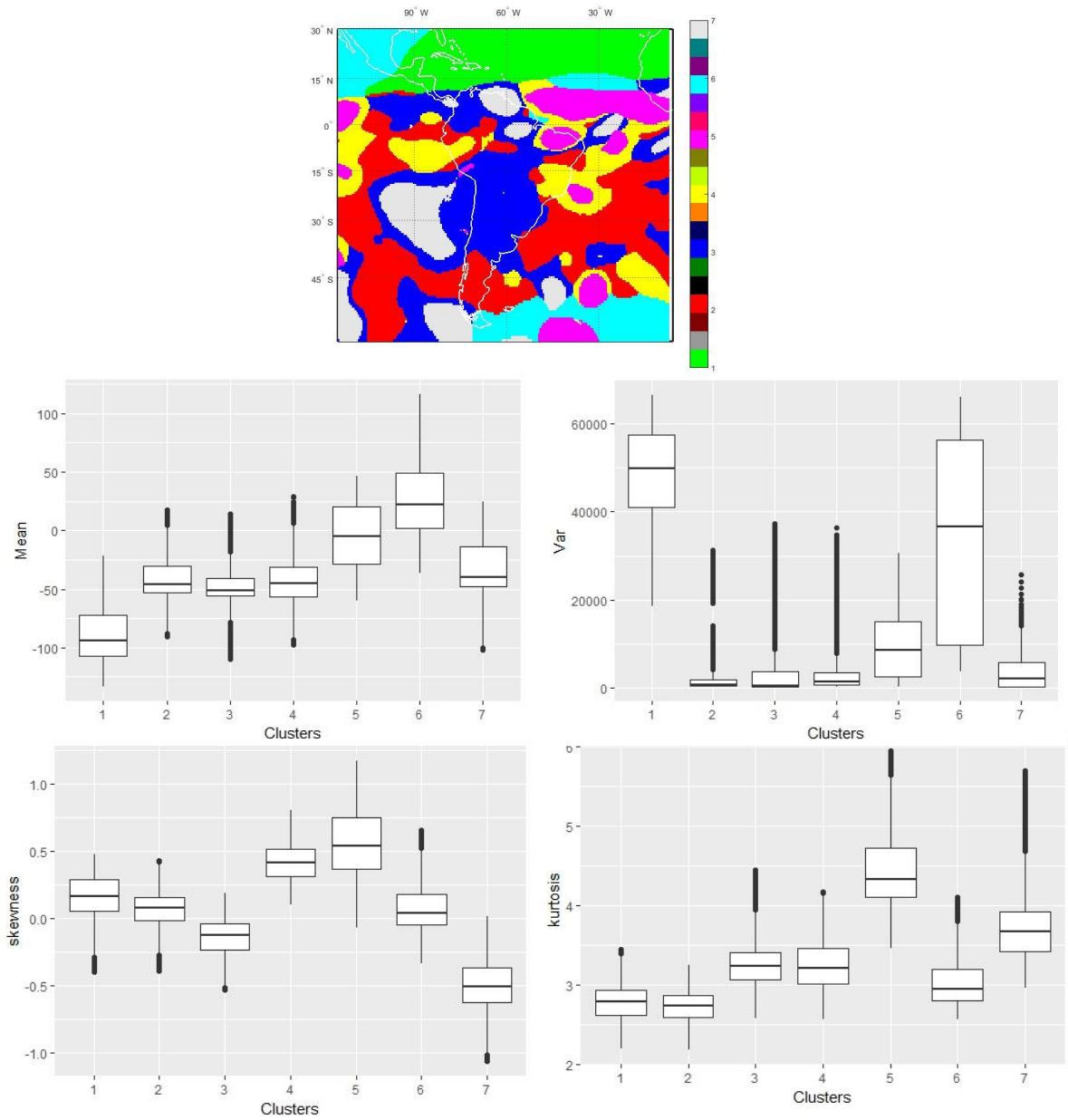

Figura 6: Agrupamento pelo K-means para a variável altura geopotencial a 300 hPa no ano de 2008; Boxplot das medidas de médias, variâncias, assimetrias e curtoses das séries temporais em cada cluster.

5.4 Umidade Específica

Na figura 7 é possível notar que o aumento do BSS já começa a ficar pequeno a partir de k=5 para k=6 mas somente quando k=8 é que o BSS assume um valor aceitável.

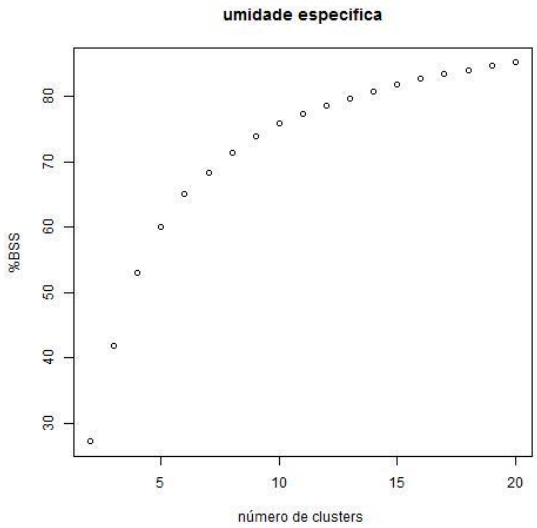

Figura 7: Escolhendo um número adequado para k (Umidade Específica).

Na figura 8, é apresentado o mapa dividido em regiões (8 clusters) de acordo com as informações de média, variância, assimetria e curtose do erro entre a variável prognóstica de umidade específica produzida pelo Eta em relação ao observado registrado pelo NCEP ao nível de pressão de 300hPa. Aqui o fenômeno das variâncias dos outros 3 casos não se repete, talvez pelo fato desta variável ter, de longe, a menor variância entre as 4 variáveis. O destaque foi o cluster 5 que acabou ficando com as maiores médias de séries temporais e acabou sendo o menor e mais compacto grupo desta separação. O cluster 3 também se destacou pelas altas medidas de curtose (tornando fácil de encontrar nesses grupos valores que não se aproximam da média a vários múltiplos do desvio padrão) e, talvez por isso, tenha ficado numa posição de destaque ao norte do mapa e com alguns poucos pedaços bem pequenos espalhados pelo resto do mapa. Este cluster também se destacou por ter as mais diversas medidas de assimetria entre todos os clusters. A maior delas pertence a um outlier e é próxima de 0,5 e ainda possui medidas de assimetria inferiores a -1 (apontando assimetria negativa forte), as menores entre todos os grupos.

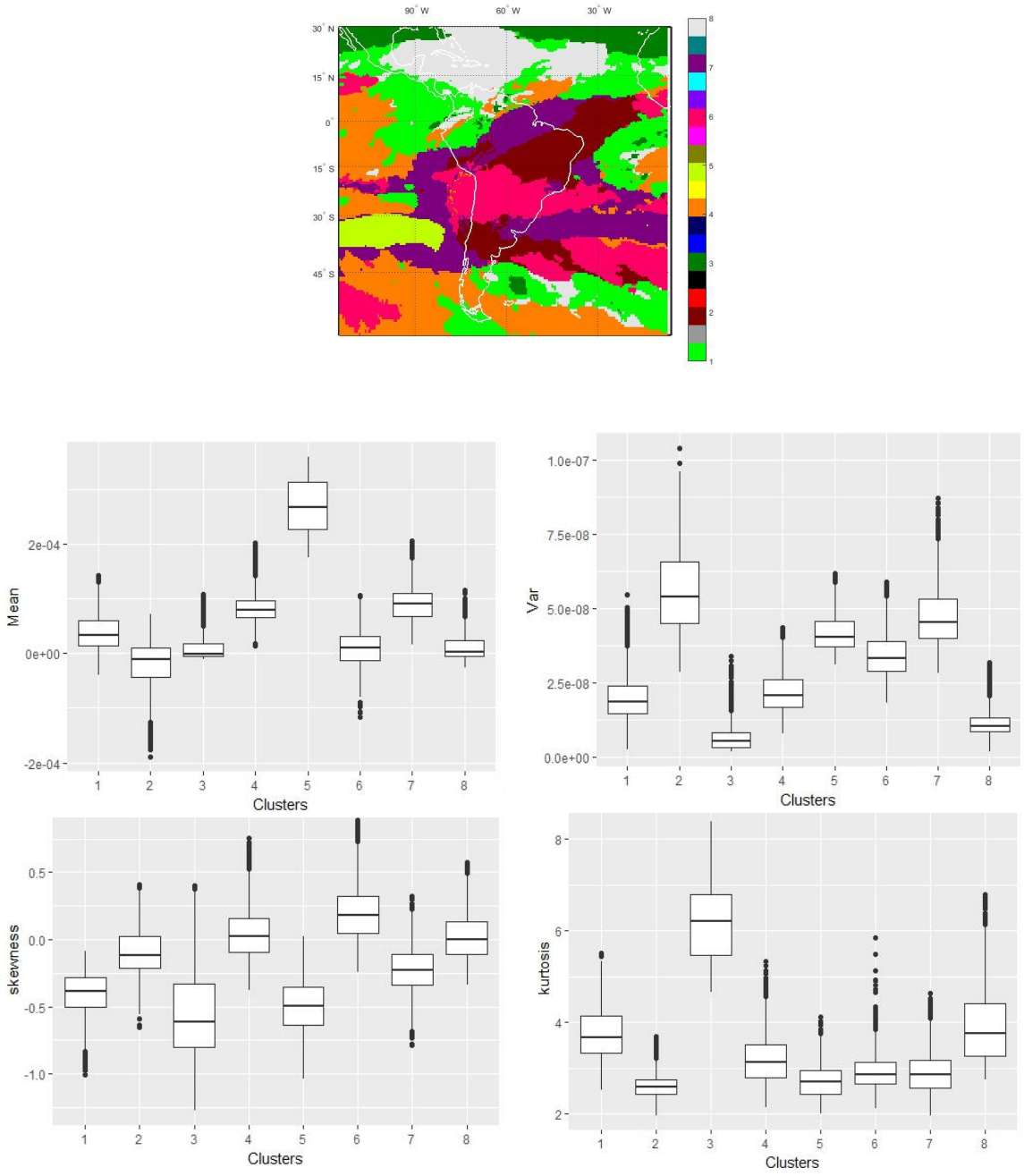

Figura 8: Agrupamento pelo K-means para a variável umidade específica a 300 hPa no ano de 2008; Boxplot das medidas de médias, variâncias, assimetrias e curtoses das séries temporais em cada cluster.

Em todos os casos, é possível notar que a parte central da área analisada concentra uma separação maior de grupos que no sul e, principalmente, no norte (é possível perceber, em alguns casos, que os clusters predominantes na área de 15°N a 30°N tendem a voltar, em uma proporção menor, na área de 45°S a 60°S). Com as 3 primeiras variáveis, esta fragmentação no centro resultou em clusters com as variâncias das series temporais menores, no geral, que as series dos outros clusters. Já com a umidade específica, o mesmo não ocorreu. O método K-means também foi aplicado para outros níveis de pressão (250, 350, 500 e 800hPa) para as 4 variáveis e o que se notou foi que o número de clusters foi poucas vezes alterado (e quando foi, apenas um a mais que os valores apresentados ao nível de 300hPa) mas as regiões se agruparam de forma diferente, principalmente quando o nível de pressão era bem maior (também não se deve menosprezar a influência de fatores aleatórios). Também foi testado a aplicação do K-means em mais de uma variável ao mesmo tempo, mas esta tentativa não se mostrou eficiente pois resultou em clusters com medidas não tão diferentes de cada variável, o que não ajudaria na futura aplicação do filtro adaptativo.

6 Conclusão

Pelo fato desta clusterização estar levando em conta as 4 medidas que caracterizam uma distribuição, é esperado uma eficiência suficiente para agrupar regiões que, por serem parecidas do ponto de vista estatístico, façam o desempenho do filtro adaptativo melhorar significativamente. A intensa separação que o algoritmo K-means fez na parte central da área analisada é um indício de que o filtro adaptativo tende a ser menos eficiente neste bloco se for aplicado na área toda (sem agrupamentos).

Referências bibliográficas

1. Bustamante, J. ; Chou, S.C. ; Rozante, J.R. ; Gomes, J.L., 2005. Uma Avaliação da Previsibilidade de Tempo do Modelo Eta para a América do Sul. Revista Brasileira de Meteorologia, Brasil, v. 20, n. 1, p. 59-70.
2. Cataldi, M. ; Osorio, C. ; Guilhon, L.G. ; Chou, S.C. ; Gomes, J.L. ; Bustamante, J., 2007. Análise das previsões de precipitação obtidas com a utilização do modelo Eta como insumo para modelos de previsão semanal de vazão natural. Revista Brasileira de Recursos Hídricos.
3. Chou, S.C., Marengo, J.A., Dereczynski, C.P., Waldheim, P., Manzi, A.O., 2007. Comparison of CPTEC GCM and Eta Model results with observational data from the Rondonia LBA reference site, Brazil. *Journal of the Meteorological Society of Japan*, vol. 85A, pp25-42.
4. Chou, S.C. , 1996. Modelo Regional Eta. Climanálise (São José dos Campos), Cachoeira Paulista, SP, v. 1, n. ED ESPECIAL, 1996.
5. Chou, S. C. ; Bustamante, J. ; Gomes, J. L., 2005. Evaluation of Eta Model seasonal precipitation forecasts over South America. Nonlinear Processes in Geophysics, Alemanha, v. 12, p. 537-555.
6. Dereczynski, C.P.; Pristo, MVJ; Chou, SC; Cavalcanti, IFA; Rozante, JR., 2010. Avaliação das Previsões do Modelo Eta na região da Serra do Mar (Estado de São Paulo), Brasil. Anuário do Instituto de Geociências, UFRJ, RJ, ISSN 0101-9759 e-ISSN 1982-3908 - Vol. 33 - 2.
7. Mesinger F, Chou SC, Gomes JL, Jovic D, Bastos P, Bustamante JF, Lazic L, Lyra AA, Morelli S, Ristic I, Veljovic K., 2012. An upgraded version of the Eta model. Meteorology and Atmospheric Physics. 116 (3), 63-79.
8. Pilotto ID, Chou SC, Nobre P., 2012. Seasonal climate hindcasts with Eta model nested in CPTEC coupled ocean-atmosphere general circulation model. Theoretical and Applied Climatology.
9. Laprise et al 2008. Challenging some tenets of Regional Climate Modelling. Meteorol Atmos Phys 100, 3–22.
10. Castro CL, Pielke RA Sr, Leoncini G, 2005. Dynamical downscaling: an assessment of value added using a regional climate model. J Geophys Res (Atmos) 110: D05108; DOI: 10.1029/2004JD004721
11. Saha et al., 2010: The NCEP Climate Forecast System Reanalysis. Bulletin of American Meteorological Society. 1015-1057.
12. Han, J., Kamber, M. & Pei, J. Data mining concepts and techniques, third edition Morgan Kaufmann Publishers, 2012