Preenchimento de Falhas em Dados Espaciais Binários de Precipitação Utilizando Máquinas de Vetor de Suporte (S*upport* V*ector* M*achines*)

Carlos Henrique Ribeiro Lima
Departamento de Estatística
Universidade de Brasília
chrlima@unb.br

Resumo

Falhas em dados observacionais é um problema frequente em estatística, aparecendo na análise de dados de diversas áreas do conhecimento e exigindo muitas vezes modelos complexos para preenchimento dessas falhas. No campo de hidro-climatologia, é comum observar falhas (ausências) em dados históricos de precipitação obtidos de estações pluviométricas. Com a crescente demanda pela água e sinais de esgotamento global desse recurso, torna-se necessário o preenchimento dessas falhas para um melhor entendimento dos padrões espaço-temporais de oferta hídrica e para que se possa prever com melhor confiabilidade e menor incerteza o comportamento futuro desse recurso natural. Dessa forma, é apresentado neste trabalho um modelo estatístico baseado em máquinas de vetor de suporte (SVM) para o preenchimento de falhas em dados de chuva diária de diversas estações pluviométricas. Os dados utilizados são binários, sendo que 0 representa um dia sem chuva ou estado seco, e 1 representa um dia chuvoso, ou estado úmido. A título de comparação, utilizou-se também o método dos vizinhos (knn) e regressão logística para o problema analisado. Os modelos foram testados a partir dos dados de chuva diária de 504 estações pluviométricas localizadas no Nordeste Brasileiro, que é uma região caracterizada por um complexo padrão espaço-temporal de chuva. As taxas de acerto obtidas a partir de validação cruzada mostram uma melhor performance do método SVM para preenchimento de falhas em dados binários de chuva. Assim, identificam-se novos caminhos para a disseminação e uso de técnicas inovadoras como SVM na análise de dados hidro-climatológicos.

Abstract

Missing data in observational studies is a common problem in statistics, arising in data analysis from several areas and requiring complex models to fill in those gaps. In the hydro-climatological field, it is common to find gaps in historical rainfall data from rainfall gauges. With the growing water demand and evidences of water depletion across the world, it is required to fill in those gaps in order to better understand the spatio-temporal patterns of water availability and to predict with better reliability and less uncertainty the future states of this natural resource. Hence, one shows here a statistical model based on support vector machines (SVM) to fill in missing rainfall data from several rainfall gauges. The data used are binary, being 0 representative of a dry state and 1 of a wet state. A comparison is made with the k-nearest neighbor (knn) and logistic regression based models. The models are tested using daily rainfall data from 504 rainfall gauges across Northeast Brazil, which is a region marked by a complex spatio-temporal pattern of rainfall. The cross-validated hit rates obtained show a better performance for the SVM model. Therefore, one identifies new paths to disseminate and use novel techniques as SVM in the analysis of hydro-climatological data.

1. Introdução

O preenchimento de falhas em dados observacionais é uma tarefa árdua que emerge em diversas áreas científicas, tendo cada problema sua relevância particular e desafio de modelagem. Livros inteiros tem sido dedicados exclusivamente para este tipo de problema (por exemplo, veja [12, 3]).

Um caso especial de falhas aleatórias aparece em dados de estações pluviométricas (isto é, estações que coletam dados de chuva). Devido a limitações físicas e econômicas, torna-se necessário que a precipitação sobre uma determinada área seja estimada através da medição de chuva em um número finito de pontos irregularmente distribuídos no espaço, usualmente chamados de estações pluviométricas ou estações de chuva. Na imensa maioria dessas estações o processo de medição diária de chuva é feito manualmente em horários específicos, normalmente às 7 horas da manhã e às 7 horas da noite. Quando essas medições não são realizadas, ocorre uma falha nos dados, normalmente registradas como -1 nas séries temporais de chuva. A introduções de satélites nos anos 50 e mais recentemente o uso de estações automáticas de chuva tem melhorado significamente a estimação espacial dessa variável, porém nenhum dos dois métodos é capaz de preencher falhas em dados de chuva em períodos mais antigos e nem substituir o uso de estações pluviométricas clássicas.

Um conjunto completo de dados de precipitação no espaço e no tempo é de fundamental importância para avaliação das variações temporais na disponibilidade hídrica de uma dada região e consequentemente para uma gestão eficiente dos recursos hídricos e uma alocação equalitária da água. Modelos estatísticos e matemáticos desenvolvidos para estudos de processos hidro-climatológicos usualmente necessitam da distribuição histórica de chuva sobre uma determinada área sem que haja a presença de

falha nos dados. Muitas das vezes é necessário apenas a classificação do dia de acordo com a ocorrência ou não de chuva (neste caso dados bináriaos: seco=0, chuva=1) para análise e modelagem (por exemplo, veja [9]). Métodos clássicos utilizados na literatura tem focado no preenchimento de falhas em dados de chuva em escalas temporais mensais e anuais, que são representações suaves do processo real de chuva. No conjunto dessses métodos, destacam-se a regressão linear, análise de séries temporais, métodos Bayesianos e métodos de interpolação espacial (krigagem, co-krigagem, inverso da distância, etc). Teegavarapua e Chandramouli [17] apresentam uma revisão dos principais métodos utilizados em hidro-climatologia.

Numa escala diária, entretanto, os padrões espaço-temporais do processo de chuva são complexos, não-lineares e altamente variáveis no tempo, o que dificulta a aplicação de métodos clássicos. Por outro lado, a análise e entendimento desses padrões torna-se um problema interessante para a comunidade estatística e de mineração de dados, que ao longo dos últimos 20 anos tem desenvolvido um grande número de métodos específicos para a complexidade do problema em análise. Destacam-se neste contexto os métodos de redes neurais artificiais (ANN) e de máquinas de vetor de suporte (SVM - *Support Vector Machines*) para mineração de dados, reconhecimento de padrões, agrupamentos e inferência. Além do sucesso numa variedade de áreas científicas, muitos dos trabalhos em hidro-climatologia tem utilizado parte desses métodos nas análises e interpretações de dados, tendo obtido uma significativa melhora sobre as metodologias clássicas (por exemplo, veja [17, 6, 16, 11, 10]. Além das complexidades mencionadas acima, vale também ressaltar que dados de chuva diária são multi-dimensionais e de grande tamanho, dificultando e restringindo muitas vezes a aplicação de métodos clássicos.

A diversidade dos mecanismos responsáveis pelo processo de chuva (por exemplo, convecção local, frentes, tornados) é responsável por definir processos e padrões espaço-temporais numa região que dificilmente são observados em dados passados e muito menos numa região diferente. Dessa forma, como observado na literatura, não existe um método universalmente aceito para preenchimento de falhas em dados diários de chuva que funcione bem em todo tipo de aplicação e local. Neste artigo são empregados alguns métodos já utilizados na literatura ([11, 10]) para o preenchimento de falhas em dados diários de chuva, porém inova-se ao aplicá-lo à dados de chuva ao longo do Nordeste brasileiro, região caracterizada por um grande número de estações pluviométricas (504) e por uma grande variabilidade espaço-temporal nos dados de chuva. O método SVM é então utilizado para o preenchimento de falhas em dados binários (0=seco e 1 = chuva) de chuva. A atividade em questão é reduzida a um problema de classificação bidimensional, onde, em um dado dia do ano, algumas estações pluviométricas pertecem à classe seca e outras a classe chuvosa. A localização espacial das estações é então utilizada como covariável e deseja-se saber a classe da estação (ou estações) que não possui dados de chuva. Os resultados obtidos com o modelo SVM são comporados com dois métodos clássicos na literatura, a saber método dos vizinhos (knn) e regressão logística.

2. Dados

São utilizados aqui dados de chuva diária para 504 estações pluviométricas ao longo do Nordeste Brasileiro. A localização espacial das estações é mostrada na figura 1. O período em que os dados foram coletados compreende 01 de janeiro de 1940 a 01 de janeiro de 2001 (22282 dias). A disponibilidade de dados nesse período é mostrada na figura 2. É interessante notar a queda no número de estações com dados de chuva disponíveis a partir da década de 1980 (Fig. 2a). Note também que a maior parte da estações apresentam em média menos de 30% dos dados com falhas (Fig. 2b).

A região coberta por essas estações é marcada por um período chuvoso cuja variabilidade espaço-temporal define basicamente três regiões distintas, com períodos chuvosos que vão de dezembro a julho do ano seguinte e caracterizados por diferentes processos de precipitação. Maiores detalhes sobre os processos climáticos na região do Nordeste podem ser obtidos em [8, 7, 9, 14] e nas referências citadas. Como em termos de planejamento e gerenciamento dos recursos hídricos o período chuvoso é o de maior importância, visto ocorrer pouca ou nenhuma chuva no período seco, são analisados somente os dados de chuva diária referentes a esse período (Janeiro-Junho).

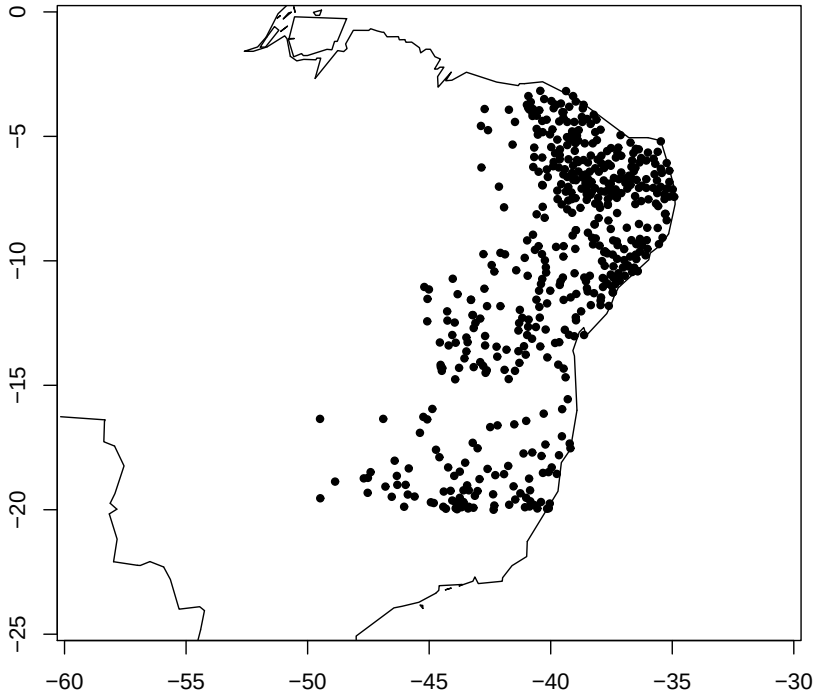

Figura 1: Localização das 504 estações de chuva no Nordeste Brasileiro.

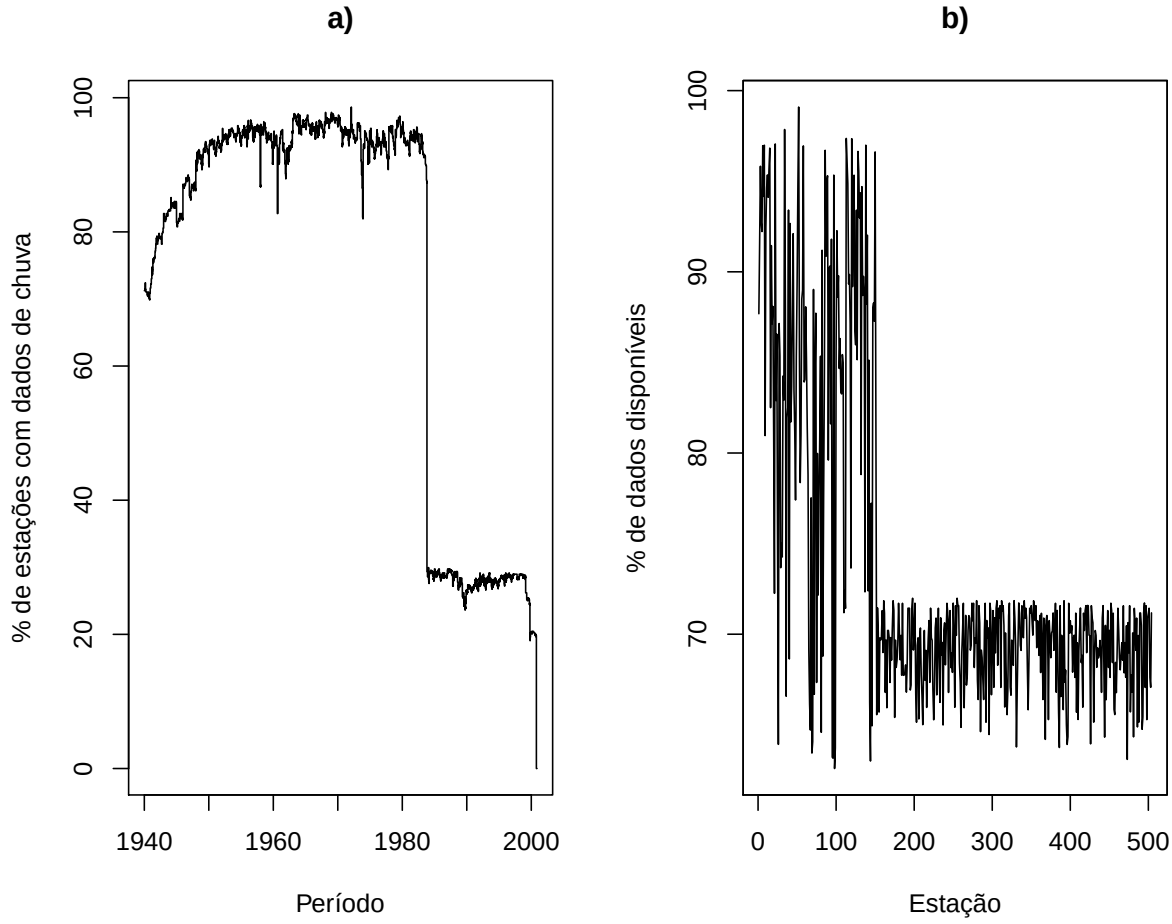

Figura 2: Disponibilidade de dados para o Nordeste: a) Percentagem de estações com dados disponíveis de acordo com o ano de ocorrência e b) para cada estações utilizada, percentagem de dados disponíveis.

3. Metodologia

O problema a ser explorado neste trabalho é relativamente simples: em cada dia do período dos dados, existem estações pluviométricas com dados disponíveis e uma ou mais estações sem dados (com falha), como mostrado na figura 3. O objetivo é preencher os dados nas estações com falhas. Entretanto, a estrutura espacial observada em cada dia pode ser bastante complexa, requerendo modelos robustos para levar em consideração os seguintes aspectos:

- Podem existir um grande número de estações com ausência de dados;
- As estações pluviométricas podem estar distribuídas esparsamente, não apresentando nenhuma correlação espacial significativa;

- Padrões espaciais similares podem não existir nos dados históricos;
- Os dados podem ser não-estacionários no espaço, assim o semi-variograma não é uma função somente da distância mas também do espaço;
- Os dados podem ser anisotrópicos, assim a correlação espacial não depende somente da distância mas também da orientação espacial;
- O grande número de zeros nos dados pode afetar as estimativas dos modelos.

Seguindo os objetivos do trabalho mencionados na seção anterior, é utilizado o método SVM para classificar a estação pluviométrica em um dos dois estados: 1 = chuvoso (representando a ocorrência de chuva no dia) e 0 = seco (representando a falta de chuva no dia). Na seção seguinte é apresentada a definição de SVM e como o mesmo é aplicada no problema proposto. São também definidos os métodos dos vizinhos e regressão logística para comparação com o método SVM. É importante observar que a natureza dos dados (dados diários) dificultam a construção de semi-variogramas e a estimação dos referidos parâmetros para cada dia analisado, portanto inviabilizam a utilização de krigagem no problema proposto.

3.1 Classificação

De um total de 504 estações pluviométricas, são selecionadas, aleatoriamente, 100 estações para serem utilizadas na fase de ajuste (calibração) e teste (validação) dos modelos desenvolvidos. Para cada uma das 100 estações testadas, são selecionados aleatoriamente 365 dias que não possuam nenhuma falha nos dados para que se possar comparar os dados observados com os dados previstos pelo modelo. Assim, utiliza-se um universo de $365 * 100 = 36500$ dias para validação dos modelos, o que acredita-se ser uma amostra suficiente para o problema em questão.

O problema básico foi ilustrado na figura 3: Dado um certo número de estações com dados binários de ocorrência de chuva (0 ou 1), deseja-se identificar a fronteira que separa eficientemente estações (e regiões) que apresentam ocorrência de chuva (indicadas por + na Fig. 3) daquelas não apresentam ocorrência (ou estado seco, indicadas por o na Fig. 3). Dado as observações, deseja-se desenvolver um contorno que delimite essa fronteira. SVM completa essa tarefa pelo uso combinado de funções *kernel* definidas sobre o espaço bidimensional (x, y) (respectivamente, longitude e latitude na Fig. 3). A formulação geral do método SVM é apresentada a seguir.

3.1.1 Máquinas de Vetor de Suporte - SVM

Máquinas de Vetor de Suporte (em inglês, *Support Vector Machines - SVM)* foram inicialmente desenvolvidas como classificadores lineares através da minimização do risco estrutural ([18]). Usualmente

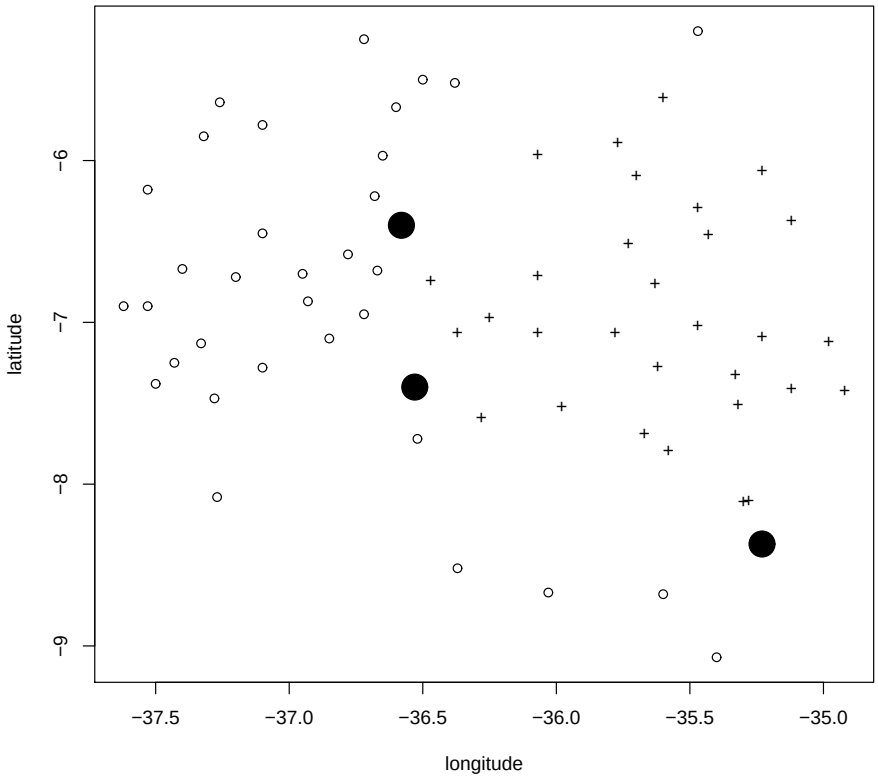

Figura 3: Exemplo do problema analisado neste trabalho. Para um dado dia do ano, as estações indicadas por $+$ (dia chuvoso) e o (dia seco) apresentam dados de chuva, enquanto as estações representadas pelos círculos maiores não apresentam dados. O objetivo é estimar o estado chuvoso (ou simplesmente preencher os dados) dessas estações.

estima-se os parâmetros do modelo por meio da minimização do risco empírico (ou função de perda), sendo este obtido na maior das vezes através de uma função de perda quadrática. Entretanto, devido ao número finito de dados, não existe garantia que com este método os estimadores corresponderam ao mínimo do risco verdadeiro, como mostrado na Figura 4. As estimativas ineficientes dos parâmetros do modelo são frequentemente observadas quando o desempenho do modelo em dados de testes (isto é, dados que não foram incluídos no processo de estimação) é bastante inferior ao desempenho do modelo em dados de treinamento (isto é, dados incluídos no processo de estimação dos parâmetros).

A idéia então é definir uma confidência $C(\theta)$:

$$J(\theta) = R_{emp}(\theta) + C(\theta) \tag{1}$$

tal que, se $R(\theta) \leq J(\theta)$, tem-se um limite para o risco verdadeiro e $J(\theta)$ é o risco garantido, como mostrado na Figura 5. θ é o parâmetro a ser estimado.

Durante os anos de 1970, Vapnik ([18]) desenvolveu a teoria da dimensão VC, na qual, com probabilidade $1-\eta$, obtém-se a seguinte fronteira para $R(\theta)$:

$$R(\theta) \leq J(\theta) = R_{emp}(\theta) + \sqrt{\frac{h\left(\log\left(\frac{2N}{h}+1\right) - \log\left(\frac{\eta}{4}\right)\right)}{N}} \tag{2}$$

onde N é o número de dados e h é a dimensão Vapnik-Chervonenkis (VC), que representa a capacidade do classificador $f(\cdot;\theta)$.

A dimensão VC não representa apenas o número de parâmetros do classificador, mas mede o número de diferentes conjuntos de dados que o classificador pode classificar corretamente. A estimativa dos parâmetros agora consiste em minimizar a fronteira do risco, ou minimização do risco estrutural. Maiores detalhes sobre a dimensão VC podem ser obtidos em [4, 15, 1].

Assuma então que os dados disponíveis consistem de N pares $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, onde $\mathbf{x}_i$ representa as coordenadas geográficas bidimensionais da localização das estações pluviométricas e $y_i \in [-1, 1]$ representa as observações de ocorrência ou não de chuva (neste caso utiliza-se a notação -1 ao invés de 0 para o estado seco) ou qualquer outro processo nessas estações. Deseja-se então definir a fronteira no plano $\mathbf{x}$ como na Figura 6 que separa eficientemente os dados baseados nos valores do processo y. Como ilustrado, existem muitas fronteiras plausíveis que poderiam separar os dados. Como poderia-se então definir a melhor fronteira, tal que futuras observações or localizações seriam corretamente identificadas?

Em SVM, a fronteira é escolhida como sendo aquela que maximiza a *margem*, isto é, a menor distância entre a fronteira de decisão e qualquer uma das amostras, para minimizar assim o risco garantido. A fronteira de decisão ou hiperplano poderia ser definida como:

$$x : f(x) = x^T\beta + \beta_0 = 0, \tag{3}$$

onde β é um vetor unitário: $\|\beta\| = 1$.

O classificador obtido de f é uma função linear:

$$G(x) = \text{sign}[x^T\beta + \beta_0]. \tag{4}$$

Como mencionado anteriormente, para um conjunto finitos de dados de tamanho N, existem várias soluções para os parâmetros β e β_0 que levam ao erro empírico zero na classificação. Deseja-se porém aquela que maximiza a margem, que é única. Assim, um modelo de otimização pode ser formulado para escolha dos parâmetros do modelo ([4, 5]):

$$\begin{aligned} &\max_{\beta,\beta_0,\|\beta\|=1} B \\ &\text{sujeito a } \; y_i(x_i^T\beta + \beta_0) \geq B, i = 1, \ldots, N, \end{aligned} \tag{5}$$

onde B define a largura da margem e é dado por:

$$B = \frac{1}{\|\beta\|}. \tag{6}$$

Essa formulação pode ser reescrita como uma problema quadrático, que apresenta como vantagens a convexidade e a solução ótima global:

$$\begin{array}{l} \min\limits_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \\ \text{sujeito a } y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \ldots, N, \end{array} \tag{7}$$

Convertendo (7) para um problema Lagrangiano (veja [4]), obtém-se a seguinte solução:

$$L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1]. \tag{8}$$

Fazendo com que as derivadas de β e β_0 sejam iguais a zero, obtém-se:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{9}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i. \tag{10}$$

Substituindo (9) em (8), obtém-se a função objetivo dual Lagrangiana:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j x_i^T x_j \tag{11}$$

que pode ser minimizada com respeito aos α's e sujeita às restrições em (10).

A solução para β tem a forma:

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \tag{12}$$

onde α_i são os multiplicadores Lagrangianos associados com cada restrição de cada ponto.

O classificador correspondente é dado por:

$$\hat{G}(x) = \text{sign}[x^T\hat{\beta} + \hat{\beta}_0] = \text{sign}\left[\sum_{i=1}^{N} \hat{\alpha}_i y_i x^T x_i + \hat{\beta}_0\right]. \quad (13)$$

O classificador linear descrito acima pode ser estendido para um contexto não-linear pelo uso de funções *kernel* (por exemplo, Gaussiana, polinomial) definidas no espaço $\mathbf{x}$ e então pela utilização do chamado *kernel trick* ([4, 5]), que mapeia $\mathbf{x}$ para um espaço dimensional maior representado pelo número de funções de base utilizadas:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x) \quad (14)$$

onde ϕ é uma função de base (1, x, x^2, $\sin(x)$) definida sobre $\Re^2$, que mapeia o espaço bidimensional para a dimensão apropriada.

O classificador então torna-se:

$$\hat{G}(x) = \text{sign}\left[\sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0\right]. \quad (15)$$

O kernel utilizado neste trabalho foi o Gaussiano:

$$K(x, x_i) = e^{-\frac{||x-x_i||^2}{c}}, \quad (16)$$

com o valor do parâmetro c padrão: $c = 1$. A figura 7 mostra um exemplo de como as fronteiras ótimas são definidas pelo SVM para o problema proposto neste trabalho utilizando o kernel Gaussiano.

A performance do modelo SVM neste problema de classificação foi avaliada a partir da taxa de acerto H (*hit rate*, veja [19]), isto é, a porcentagem de dias que são classificados corretamente como chuvoso ou seco num experimento de validação cruzada aleatório como descrito anteriormente. Os resultados desse experimento são comparados com os resultados obtidos pelo método dos vizinhos (knn) e por regressão logística, como explicado a seguir.

3.1.2 Método dos vizinhos (knn) e regressão logística

Para melhor avaliar o desempenho do modelo SVM, são utilizados os métodos dos vizinhos (knn, veja [4]) e a regressão logística ([13]) no mesmo contexto experimental explicado anteriormente. O método knn consiste primeiro em achar os k vizinhos mais próximos geograficamente da estação pluviométrica na qual deseja-se preencher as falhas. A distância Euclidiana (ou norma l^2) é utilizada aqui para calcular a distância d entre duas estações pluviométricas com coordenadas geográficas $\mathbf{x}$ e $\mathbf{x_0}$:

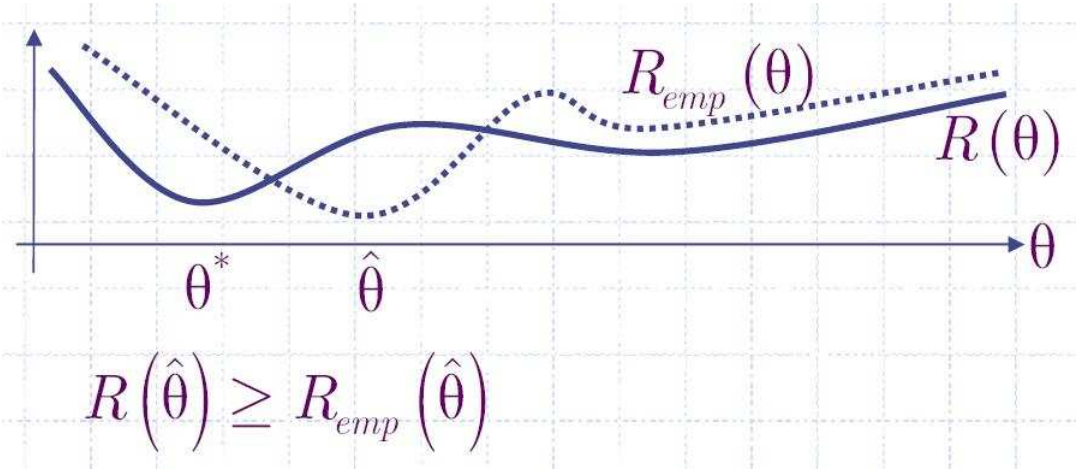

Figura 4: O risco verdadeiro ($R(\theta)$ - linha sólida) é maior que o risco empírico ($R_{emp}(\theta)$ - linha pontilhada) para os parâmetros estimados $\hat{\theta}$. Adaptado de [5].

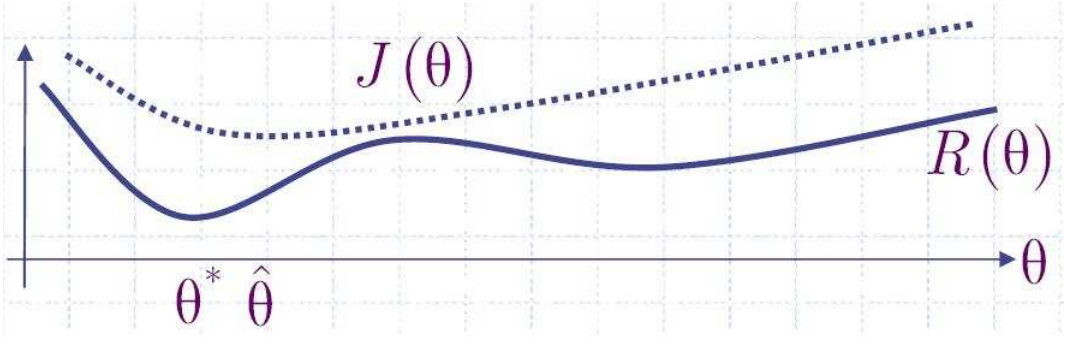

Figura 5: O risco verdadeiro e o risco garantido $J(\theta)$. Adaptado de [5].

$$d(\mathbf{x}, \mathbf{x_0}) = |\mathbf{x} - \mathbf{x_0}| = \sqrt{(x^{lat} - x_0^{lat})^2 + (x^{long} - x_0^{long})^2}, \tag{17}$$

onde os superescritos lat e $long$ representam latitude e longitude, respectivamente.

As distâncias d entre todas as estações são então ordenadas crescentemente e os k vizinhos são selecionados. Para um dado dia dos dados de validação, a classificação da estação em análise é dada simplesmente pela moda do conjunto de valores $(-1, 1)$ das k estações vizinhas.

A regressão logística (maiores detalhes da formulação matemática podem ser vistos em [13]) foi utilizada tendo como variável resposta os valores históricos binários (0 e 1) da estação em análise e como co-variáveis os valores de precipitação real das k estações vizinhas. Observe que os dados de validação não são utilizados na estimativa dos parâmetros. De acordo com a probabilidade p de chuva

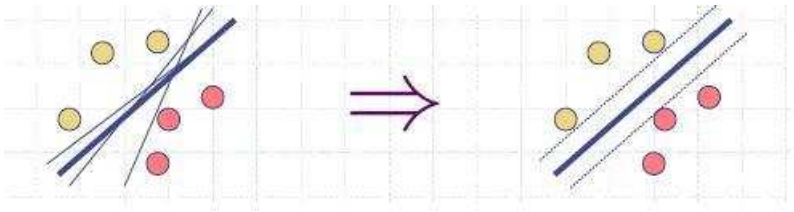

Figura 6: Seleção da fronteira ótima pelo SVM. Adaptado de [5].

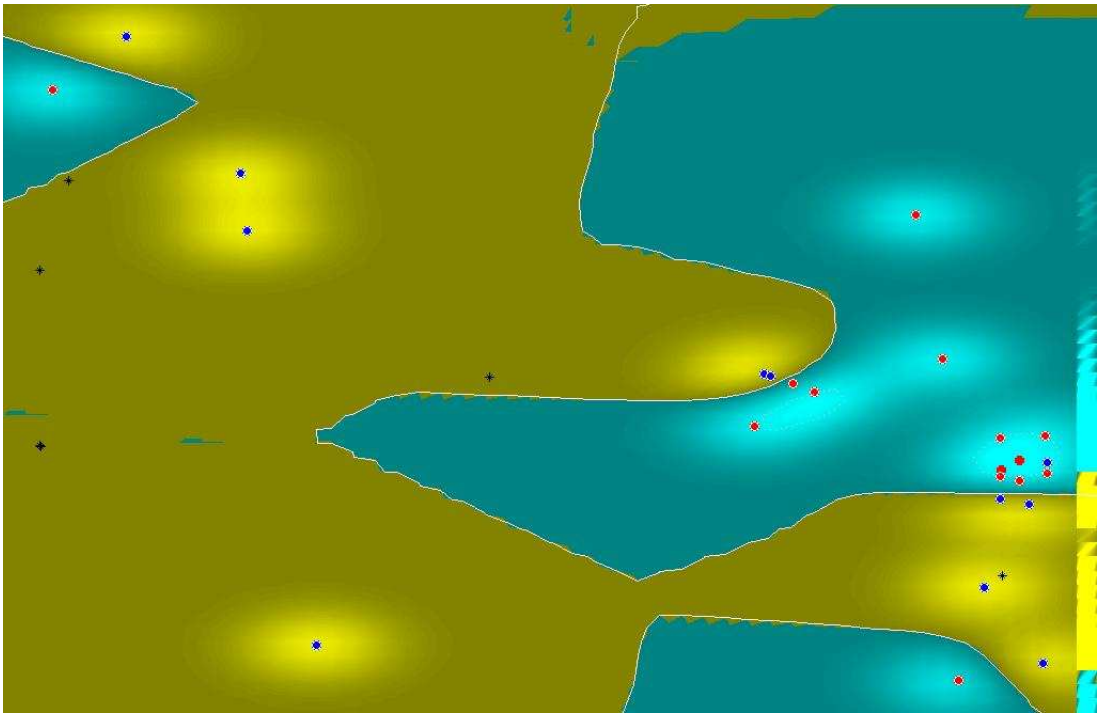

Figura 7: Exemplo de fronteira ótima determinada pelo SVM para as classes chuvosa (pontos azul) e seca (pontos vermelho). Os pontos pretos indicam estações com ausência (falha) de dados.

prevista para a estação em análise, preencheu a falha na estação em análise. Para $p > 0.5$, considerou-se estado chuvoso (1) e para $p \leq 0.5$ admitiu-se estado seco (-1) para a estação em questão. Observe que a regressão logística tem como base dados históricos para estimativa dos parâmetros, enquanto os métodos knn e SVM não dependem de nenhuma informação histórica.

Como no método knn, utilizou-se as k estações vizinhas para ajuste do modelo SVM. Esse procedimento é similar ao proposto por [2], onde primeiramente o método dos vizinhos é utilizado para selecionar um sub-conjunto dos dados e em seguida o método SVM é aplicada a este sub-conjunto.

4. Resultados

Como o desempenho de todos os três métodos é uma função não-linear do número k de vizinhos escolhidos (veja discussão em [4]), avaliou-se o desempenho para um conjunto de valores de k, a saber: 10, 20, 30, 40, 50, 60 e 70. A Figura 8 apresenta os valores médios (para cada estação calcula-se a taxa de acerto H para os 365 dias de validação e a partir daí calcula-se a média de H para as 100 estações analisadas) de H para os três modelos testados em função do número de vizinhos k. É interessante notar que todos os três modelos apresentam uma curva côncova, com um número ótimo de vizinhos k na qual a taxa de acerto é maximizada: 40 para SVM e o método dos vizinhos e 30 para a regressão logística. O melhor desempenho é obtido com o modelo SVM com $k = 40$, onde a taxa de acerto (ou previsões corretas) é de aproximadamente $82,1\%$.

Para avaliar a variabilidade da taxa de acerto H em torno dos valores médios apresentados na Fig. 8, é apresentado na figura 9 o diagrama de caixas para essa variável. Observe que a maior variabilidade é encontrada no modelo knn seguido do modelo SVM, que ainda apresenta o menor valor de mediana (ou percentil de 50%) porém os maiores valores no intervalo de 50% (percentis de 25% e 75%). O modelo baseado na regressão logística apresenta ainda os maiores valores na cauda inferior e os menores valores entre os três modelos na cauda superior da distribuição.

Como mencionado anteriormente, todos os resultados apresentados anteriormente foram obtidos para o período chuvoso (JAN-JUL), que operacionalmente é o mais importante para a gestão dos recursos hídricos. Entretanto, como forma de complementar a análise dos modelos, foram realizadas também análises para o período seco, onde, em geral, são obtidas melhores previsões ([10, 11]) dada a homogeneidade do estado seco nas estações analisadas. Os resultados obtidos (não mostrados aqui) apresentaram taxas de acerto H em torno de $0,90$ para $k = 30$, sendo que nenhuma diferença significativa foi observada para os três modelos testados.

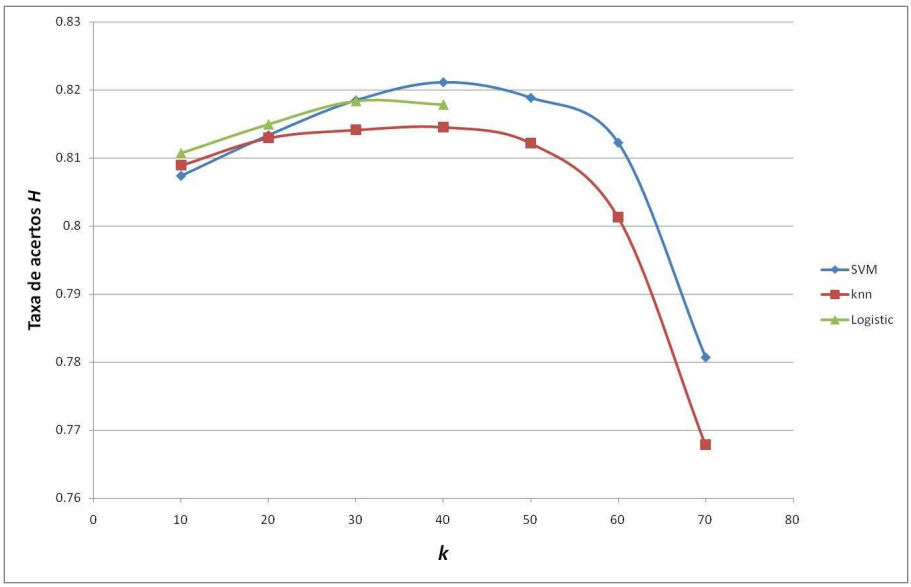

Figura 8: Taxa de acerto H para os três modelos testados em função do número de vizinhos k escolhido. Para $k \geq 50$, não é nem sempre possível encontrar períodos históricos nos quais todos os 50 ou mais vizinhos apresentam dados disponíveis, inviabilizando assim a estimativa dos parâmetros para o modelo de regressão logística.

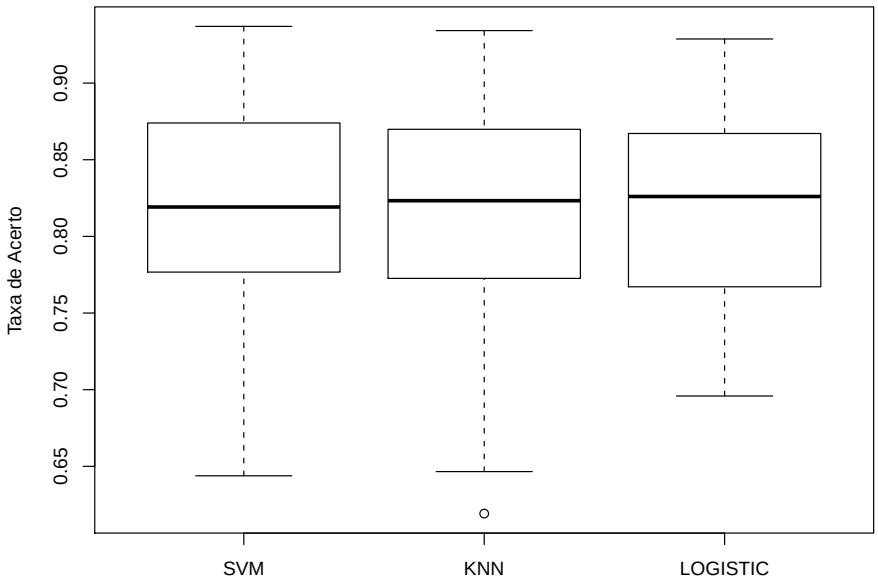

Figura 9: Diagrama de caixas para a taxa de acerto H para os três modelos testados ($k = 40$).

5. Discussão e Conclusões

Neste trabalho foi apresentado o modelo de máquinas de vetor de suporte (SVM) para uso em preenchimento de falhas em dados espaciais de ocorrência de chuva, tarefa fundamental para melhoria da qualidade dos dados históricos de precipitação e desenvolvimento de modelos matemáticos e estatísticos para suporte da gestão dos recursos hídricos. Aplicou-se o modelo para dados de chuva diária de 504 estações pluviométricas localizadas no Nordeste brasileiro, região em que o regime espacial de precipitação e a escala temporal utilizada no trabalho dificultam ou até mesmo inviabilizam o uso de metodologias clássicas, como regressão linear e krigagem. A performance do modelo SVM foi comparada com a performance obtida por meio do método dos vizinhos (knn) e regressão logística.

Para um número de vizinhos k maiores que 30, os resultados obtidos pelo modelo SVM para o período chuvoso, a julgar pela taxa de acerto H média, são superiores aos resultados obtidos pelos outros dois modelos. A maior taxa de acerto ($H = 0,821$) foi obtida pelo modelo SVM para $k = 40$. A distribuição da taxa de acerto H mostra que os percentis de 25% e 75% são também maiores para o modelo SVM, sendo as caudas inferior e superior do modelo logístico a maior e menor, respectivamente, entre os três modelos. Para o período seco foram obtidas taxas de acerto da ordem de $0,90$, porém nenhuma diferença significativa foi observada entre os modelos, o que era de certa forma esperado para este período dado a grande homogeneidade espacial da chuva, pois a grande maioria das estações pluviométricas apresentam-se no estado seco (sem chuva) durante esse período.

Assim, mesmo com a escolha de valores de parâmetros padrões ($c = 1$), considera-se que o modelo SVM obteve a melhor performance entre os três modelos testados para o problema em estudo. Acredita-se ainda que os valores obtidos para H podem ser melhorados com o uso de outros kernels e o desenvolvimento de métodos adaptativos para escolha dos parâmetros do modelo SVM e do número k de vizinhos. Apesar de não ter sido foco deste trabalho, o modelo SVM desenvolvido pode também ser utilizado no preenchimento de múltiplas falhas binárias de chuva, pois o contexto de desenvolvimento do modelo SVM considera implicitamente a estrutura de correlação espacial existente entre as várias estações pluviométricas. Será também foco de trabalho futuro a extensão do uso do modelo SVM para preenchimento de falhas em dados reais (quantitativos e não binários) de precipitação.

Referências

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] C. Domeniconi and D. Gunopulos. Adaptive Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[3] A. Gelman and Xiao-Li Meng. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley, 2004.

[4] T.H. Hastie, R.Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[5] T. Jebara. Advanced Machine Learning. Notas de Aula, Columbia University. 2006.

[6] Tae-Woong Kim and H. Ahn. Estimating missing daily rainfalls using neural network-based pattern classifiers. *Hydrological Process*, 2007.

[7] V. E. Kousky. Frontal Influences on Northeast Brazil. *Mon. Weather Rev.*, 107:1140–1153, 1979.

[8] V. E. Kousky. Diurnal rainfall variations in Northeast Brazil. *Mon. Weather Rev.*, 108:488–498, 1980.

[9] C. H. R. Lima and U. Lall. Hierarchical Bayesian Modeling of Multisite Daily Rainfall Occurrence: rainy season onset, peak and end. *Water Resources Research*, 45, 2009.

[10] C. H. R. Lima, U. Lall, T. Landot, and C. Pathak. Missing-Data Estimation for Daily Rainfall in Everglades Florida Using Machine Learning Methods. In *AGU 2008 Joint Assembly*, 2008.

[11] C.H.R. Lima, U. Lall, and C. Pathak. Machine Learning methods for missing-data estimation for daily rainfall. In *World Environmental and Water Resources Congress*, 2008.

[12] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2002.

[13] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman e Hall/CRC, 1989.

[14] A. D. Moura and L. Shukla. On the dynamics of droughts in northeast Brazil: observations, theory and numerical experiments with a general circulation model. *J. Atmos. Sci.*, 38:2653–2675, 1981.

[15] B. Schlkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.

[16] R.S.V. Teegavarapu. Use of universal function approximation in variance-dependent surface interpolation method: An application in hydrology. *Journal of Hydrology*, 332:16–29, 2007.

[17] R.S.V. Teegavarapua and V. Chandramouli. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312:191–206, 2005.

[18] V. N. Vapnik. *The nature of statistical learning theory*. Springer, 1995.

[19] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences An Introduction*. Academic Press, 1995.