UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS
CURSO DE ESTATÍSTICA

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

Marco Aurelio Valles Leal

Maringá
2022

MARCO AURELIO VALLES LEAL

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de bacharel em Estatística pela Universidade Estadual de Maringá.

Orientador: Profº Drº George Lucas Moraes Pezzott

Maringá
2022

AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA PREDIÇÃO EM MODELOS ADITIVOS

MARCO AURELIO VALLES LEAL

Trabalho de conclusão de curso apresentado como requisito parcial para a obtenção do título de bacharel em Estatística pela Universidade Estadual de Maringá.

Aprovado em: _____/_____/_____.

BANCA EXAMINADORA

_____

Orientador
Profº Drº George Lucas Moraes Pezzott
Universidade Estadual de Maringá

_____

Membro da banca
Profº Drº Brian Alvarez Ribeiro de Melo
Universidade Estadual de Maringá

_____

Membro da banca
Profº Drº Willian Luís de Oliveira
Universidade Estadual de Maringá

RESUMO

É comum, nas mais diversas áreas, investigar e modelar a relação entre variáveis. A análise de regressão assume que a média da variável resposta é modelada como uma função linear das variáveis explicativas, supondo erros aleatórios com média zero, variância constante e não correlacionados. Entretanto, nem sempre a relação existente é perfeitamente linear. Neste contexto, é possível flexibilizar o modelo de regressão, modelando a dependência da variável resposta com cada uma das variáveis explicativas em um cenário não paramétrico. Esta nova classe de modelos é dita modelos aditivos e mantém a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos. Portanto, este projeto visa apresentar os modelos aditivos, comparando técnicas de suavização utilizadas para ajustar modelos no contexto não paramétrico. Empregar-se-á um método de validação cruzada para se obter um parâmetro de suavização considerado ótimo, obtendo o melhor ajuste. Além do mais, métricas foram abordadas no intuito de verificar e comparar o poder preditivo e qualidade do ajuste dos suavizadores. Todavia, para validar a metodologia, realizou-se um estudo de simulação tomando vários cenários em 1000 amostras. Os resultados obtidos entre os métodos de suavização foram comparados por meio das métricas introduzidas. Verificou-se qual suavizador exibiu o melhor poder preditivo e qual se ajustou melhor aos dados simulados. Por fim, os procedimentos adotados foram aplicados em dados reais, nos quais os suavizadores foram mais uma vez comparados e validados em relação ao poder preditivo e qualidade de ajuste.

Palavras-chave : Regressão. Modelo aditivo. Suavizadores.

Sumário

1 Introdução

A análise de regressão é uma técnica amplamente utilizada na estatística que visa explorar e modelar a relação entre variáveis (MONTGOMERY ET AL.,2012). Resumidamente, a análise de regressão tem como objetivo descrever a relação entre uma variável de interesse, chamada de variável resposta ou dependente (y), e um conjunto de p variáveis preditoras ou independentes ($x = (x_1, x_2, \dots, x_p)$), chamadas comumente de covariáveis. Em particular, o modelo de regressão linear múltipla é aplicado nas mais diversas áreas do conhecimento como, por exemplo, nas ciências biológicas, físicas e químicas, na economia e na engenharia. Tal modelo parte do pressuposto que a variável resposta está relacionada com as covariáveis pela seguinte relação linear:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

onde os parâmetros $\beta_0, \beta_1, \dots, \beta_p$ são os coeficientes de regressão e o termo ε é um erro aleatório que deve satisfazer os pressupostos de ter média zero, variância σ^2 constante e serem não correlacionados. O parâmetro β_j representa a mudança esperada na resposta y a cada mudança unitária de x_j quando todas as outras variáveis regressoras são constantes e, por isso, são frequentemente chamadas de coeficientes parciais de regressão. Em geral, os coeficientes de regressão são parâmetros desconhecidos e podem ser estimados utilizando, dentre outras técnicas, o método dos mínimos quadrados, por exemplo.

Os modelos de regressão linear múltipla são vistos como modelos empíricos ou funções aproximadas no sentido que a verdadeira função que descreve o relacionamento entre y e $x_1, x_2, \dots, x_p$ é desconhecida, mas em certos intervalos das variáveis regressoras, o modelo de regressão linear é uma aproximação adequada para a verdadeira função desconhecida. Porém, em muitos casos, a relação existente entre a variável resposta e cada uma das covariáveis pode não ser linear. Uma solução seria acrescentar uma transformação nas variáveis regressoras adotando, exemplificando, o método de transformação de Box-Cox. Contudo, determinar uma transformação que represente a correta relação existente nem sempre é uma tarefa fácil. Outra possibilidade é flexibilizar o modelo de regressão linear, modelando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico (EUBANK, 1999). Esta nova classe de modelos é dita modelos aditivos (HASTIE E TIBSHIRANI,1990).

Nos modelos aditivos, considera-se que cada uma das covariáveis está relacionada com a variável resposta y por meio de uma função univariada desconhecida (função suave) não especificada de uma forma paramétrica, ou seja, o componente sistemático é formado por uma soma de funções suaves não especificadas das covariáveis, além de um termo aleatório. Esta nova classe de modelos é dita modelos aditivos e mantém a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos. Os modelos aditivos são um caso particular de uma classe mais geral denominada modelos aditivos generalizados (HASTIE & TIBSHIRANI,

1990), definido por,

$$y = \alpha + \sum_{j=1}^{p} f_j(x_j) + \varepsilon,$$

onde os erros ε são independentes com média zero e variância constante σ^2. Cada $f_j(x_j)$ é uma função univariada arbitrária. Observa-se que um modelo de regressão linear múltiplo é obtido adotando $f_j(x_j) = \beta_j x_j$ na equação acima.

Os modelos aditivos mantêm muitas das boas propriedades dos modelos lineares. A título de exemplo, uma das vantagens de modelos lineares é sua simplicidade na interpretação: caso o interesse seja saber como a previsão muda conforme mudanças em x_j, é necessário saber apenas o valor de β_j. Nota-se que a função de resposta parcial f_j desempenha esse mesmo papel em um modelo aditivo, mas não precisa ter necessariamente um comportamento linear. De certa forma, os modelos aditivos podem ser vistos como uma flexibilização do modelo de regressão linear.

As funções f_j devem ser estimadas por meio de suavizadores, que estimam uma tendênica menos variável e descreve sua dependência em relação à variável resposta. Algumas das técnicas mais conhecidas para obter as estimativas suavizadas são: suavizador *bin*, média móvel, linha móvel, suavizador, *Loess* ou *Lowess*, de *Kernel* e *Splines* de regressão (HASTIE E TIBSHIRANI,1990). De fato, a "suavização" de cada técnica supracitada depende da escolha (estimativa) do seu respectivo "parâmetro de suavização". Em algumas técnicas, por exemplo, dependendo do valor do parâmetro de suavização, pode-se ter uma reta estimada ou até uma interpolação dos dados. Em termos gerais, pode-se dizer que a variância e o viés na estimação/predição do modelo suavizado dependem do parâmetro suavizador. Este problema é, de certa forma, análogo à questão de quantas variáveis preditoras colocar em uma equação de regressão. Logo, torna-se de suma importância, além do estudo comparativo das técnicas de suavização no ajuste em modelos aditivos no contexto não paramétrico, também a estimação do parâmetro suavizador de cada técnica. Adicionalmente, o estudo do desempenho na predição dessas técnicas também é pertinente partindo da hipótese prévia que nem sempre o melhor ajuste resulta em melhor poder preditivo. Dito isto, descreve-se a seguir os objetivos de estudo deste trabalho nesta direção.

1.1 Objetivo Geral

O objetivo deste trabalho é estudar algumas das principais técnicas de estimação do modelo aditivo no contexto não paramétrico. Em particular, serão abordados os suavizadores *Kernel*, *Loess* e *Splines* de regressão considerando apenas uma covariável, e seus desempenhos serão comparados com foco principal na predição.

1.2 Objetivos Específicos

- Apresentar algumas das principais técnicas de suavização da literatura para estimar funções não paramétricas presentes nos modelos aditivos, identificando suas principais características;
- Introduzir uma métrica para estimar o parâmetro de suavização de cada técnica;
- Adotar uma métrica para estimar avaliar a qualidade de estimação e predição;
- Realizar um estudo de simulação para comparar a qualidade do ajuste e predição dos modelos em alguns cenários, tendo em vista os modelos aditivos;
- Aplicar a metodologia estudada a um conjunto de dados reais, comparando modelos e técnicas de estimação e predição.

2 Referencial Teórico

Existem distintas abordagens para obter estimativas das funções em métodos de regressão não-paramétricos. Estas estimativas dependem dos próprios dados e de suas observações vizinhas em torno de um dado ponto. Um dos primeiros e mais utilizados métodos de regressão não-paramétrica foi apresentado por Nadaraya-Watson (1964), denominados estimadores tipo núcleo (*Kernel*), os quais foram aperfeiçoados com os métodos de regressão polinomial local, conhecidos como *loess* (CLEVELAND, 1979). Em um modelo com apenas uma covariável, a estimação de $f(x)$ consiste em ajustes locais, realizando vários ajustes paramétricos por meio de regressão polinomial com pesos (*Lowess*), considerando os dados mais próximos do ponto onde deve ser feita a estimação da função (DELICADO, 2008). No entanto, deve-se escolher de forma apropriada os parâmetros da largura da banda (parâmetro de suavização) e os graus de ajuste polinomiais a fim de se obter o melhor ajuste da regressão.

Além destes, tem-se que o método *splines* (vide, por exemplo, Reinsch 1967 e Eubank 1999) corresponde em encontrar um estimador para $f(x)$ que minimiza a soma de quadrados dos resíduos (GREEN E SILVERMAN, 1994). Por exemplo, tem-se os splines cúbicos, que são generalizações de polinômios cúbicos adotados na regressão paramétrica, sendo amplamentes utilizados na literatura por apresentar em geral um bom ajuste. As técnicas *Kernel*, *Lowess* e *Splines* de regressão (linear e cúbico) são discutidas em mais detalhes na seção a seguir.

3 Metodologia

Neste trabalho, considerar-se-á o estudo da relação entre uma variável resposta (y) e uma variável explicativa (x) a partir de um modelo aditivo (HASTIE & TIBSHIRANI, 1990) da seguinte forma,

$$y = \alpha + f(x) + \varepsilon,$$

onde os erros ε são independentes com média zero e variância constante σ^2 e $f(x)$ é uma função univariada arbitrária.

3.1 Suavizadores e Técnicas de suavização

A função $f(x)$ do componente sistemático pode ser estimada a partir de um suavizador (*smoother*). Um suavizador pode ser definido como uma ferramenta para resumo da tendência das medidas y como função de uma (ou mais medidas) x. É importante destacar que as estimativas das tendências terão menos variabilidade que as variáveis respostas observadas, o que explica o nome de suavizador para a técnica aplicada (HASTIE & TIBSHIRANI, 1990). Chama-se a estimativa produzida por um suavizador de "*smooth*". No caso de uma variável preditora é chamado de suavizador em diagrama de dispersão.

Os suavizadores possuem dois usos principais, sendo o primeiro uso a descrição. Um suavizador em diagrama de dispersão pode ser usado para melhorar a aparência visual do gráfico de x *versus* y para encontrar uma tendência nos dados. O segundo uso é de estimar a dependência da esperança de y.

O suavizador mais simples é o caso dos preditores categóricos, como sexo (masculino, feminino), por exemplo. Para suavizar y podemos simplesmente realizar a médias dos valores de y para cada categoria. Este processo captura a tendência de y em x. Pode não parecer que simplesmente realizar as médias seja um processo de suavização, mas este conceito é a base para a configuração mais geral, já que a maioria dos suavizadores tenta "imitar" a média da categoria por meio da média local, ou seja, realizar a média dos valores de y, tendo os valores preditores próximos dos valores alvo. Esta média é feita nas vizinhanças em torno do valor alvo. Nesse caso, têm-se duas decisões a serem tomadas:

- O quão grande a vizinhança deve ser;
- Como realizar a média dos valores da resposta y em cada vizinhança.

O tamanho da vizinhança a ser tomada é, normalmente, expressa em forma de um parâmetro (parâmetro suavizador). Intuitivamente, grandes vizinhanças produzirão estimativas com

variância pequena mas potencialmente com um grande viés e inversamente quando adotado vizinhanças pequenas. Portanto, temos uma troca fundamental entre variância e viés estipulada pelo parâmetro suavizador.

A questão de como realizar a média em uma vizinhança é a questão de qual tipo de suavizador utilizar, pois os suavizadores diferem principalmente pelo jeito de realizar as médias. Algumas das técnicas mais conhecidas para obter as estimativas suavizadas são: suavizador *bin*, média móvel, linha móvel, suavizador *Loess* ou *Lowess*, suavizadores de *Kernel* e *Splines* de regressão (HASTIE E TIBSHIRANI,1990), discutidas a seguir.

O suavizador *bin*, também conhecido como regressograma, imita um suavizador categórico, particionando os valores preditores em regiões disjuntas e, então, realizando a média da resposta em cada região. A estimativa final não tem uma forma bem suavizada, pois é possível ver um salto em cada ponto de corte.

A média móvel (*running mean*) é outra técnica que leva em conta o cálculo da média. É muito comum utilizar uma vizinhança/região de $(2k+1)$ observações, k para a esquerda e k para a direita de cada observação, no qual o valor de k tem um comportamento de troca entre suavidade e qualidade do ajuste.

Um problema comum encontrado na média móvel é o viés. Uma saída é usar pesos para dar mais importância às vizinhanças mais próximas. Uma solução ainda melhor é utilizar a técnica de linha móvel (*running line*), na qual novamente são definidas as vizinhanças para cada ponto, tipicamente os k pontos mais próximos de cada lado. Nesse caso é mais interessante considerar a proporção de pontos em cada vizinhança, ou seja, $w = \frac{(2k+1)}{n}$, denominado *span*. Então, ajusta-se uma linha de regressão aos pontos de cada região, que é usada para encontrar o valor predito suavizado para o ponto de interesse.

3.1.1 Loess

Também chamado de *Lowess*, essa técnica pode ser vista como uma linha móvel com pesos locais (*locally weighted running line*). Um suavizador desse tipo, seja denominado $s(x_0)$, adotando k vizinhos mais próximos, pode ser computada por meio dos seguintes passos:

Procedimento 1

- *Os k vizinhos próximos de x_0 são identificados e denotados por N_{x_0};*
- *É computada a distância do vizinho-próximo mais distante de x_0:*

$$\Delta(x_0) = max_{x \in N_{x_0}} |x_0 - x|$$

- *Os pesos w_i são designados para cada ponto em N_{x_0}, usando a função de peso tri-*

cúbica:

$$w_i = W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right)$$

onde

$$W(u) = \begin{cases} (1-u^3)^3, & 0 \leq u \leq 1 \\ 0, & \textit{caso contrário} \end{cases}$$

- *Basicamente, para cada observação x_i, reserva-se k vizinhos próximos de x_i denotados N_{x_i}. Ou seja, N_{x_i} irá conter todos os vizinhos mais próximos de x_i onde sua quantidade é controlada pelo parâmetro suavizador, usualmente denotado por span. Considerando a i-ésima observação, ajusta-se uma regressão linear, levando em conta a função de peso tri-cúbica, que deve ser incorporada no modelo. A estimativa para $s(x_i)$, para i-ésima observação, será o valor predito de x_i em relação a seu i-ésimo modelo de regressão linear.*

As hipóteses em relação ao modelo *Loess* são menos restritivas se comparadas às do modelo de regressão linear, já que se assume que ao redor de cada ponto x_i o modelo deve ser aproximadamente uma função local linear.

Destaca-se que nessa técnica deve-se ter atenção à escolha do valor do *span*. Um valor muito pequeno faz com que a curva seja muito irregular e tenha variância alta. Por outro lado, um valor muito grande fará com que a curva seja sobre-suavizada, podendo não se ajustar bem aos dados e resultando em perda de informações e viés alto. Nos passos mostrados anteriormente o valor do *span* foi escolhido mediante o método de vizinhos mais próximos.

3.1.2 Kernel

Um suavizador *kernel* usa pesos que decrescem suavemente enquanto se distância do ponto de interesse x_0. Vários métodos podem ser chamados de suavizadores *kernel* por intermédio dessa definição. Porém, na prática, o suavizador *kernel* representa a sequência de pesos descrevendo a forma da função peso com o auxílio de uma função densidade com um parâmetro de escala que ajusta o tamanho e a forma dos pesos perto de x_0. Um suavizador *Kernel* pode ser definido da forma

$$\hat{y}_i = \frac{\sum_{j=1}^{n} y_i K\left(\frac{x_i - x_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right)}$$

onde b é o tamanho da vizinhança (parâmetro – de escala – suavizador), e K uma função kernel, ou seja, uma função densidade. Existem diferentes escolhas para K, geralmente usa-se a densidade de uma Normal, tendo-se, assim, um *kernel* Gaussiano.

3.1.3 Splines de regressão

Um *Spline* pode ser visto como uma função definida por um polinômio por partes. Pontos distintos são escolhidos no intervalo das observações (nós) e um polinômio é definido para cada intervalo, dessa forma, é possível modelar com polinômios mais simples as curvas mais complexas. Os *splines* dependem, principalmente, do grau do polinômio e do número e localização dos nós. Essa técnica é interessante, pois, tem uma maior flexibilidade para o ajuste dos modelos em comparação com o modelo de regressão polinomial ou linear e, após a determinação da localização e quantidade de nós, o modelo é de fácil ajuste. Além disso, o *spline* permite modelar um comportamento atípico dos dados, o que não seria possível com apenas uma função.

Existem várias diferentes configurações para um *spline*, mas uma escolha popular é o *spline* cúbico, contínuo e contendo primeira e segunda derivadas contínuas nos nós. As *splines* cúbicas são as de menor ordem nas quais a descontinuidade nos nós são suficientemente suaves para não serem vistas a olho nu, então, a não ser que seja necessário mais derivadas suavizadas, existe pouca justificativa para utilizar *splines* de maior ordem.

Para qualquer grupo de nós, o *spline* de regressão é ajustado a partir de mínimos quadrados em um grupo apropriado de vetores base. Esses vetores são as funções base representando a família do pedaço do polinômio cúbico, com valor dado a partir dos valores observados de x.

Uma variação do *spline* cúbico é o *spline* cúbico natural, que contém a restrição adicional de que a função é linear além dos nós dos limites. Para impor essa condição, é necessário que, nas regiões dos limites: $f''' = f'' = 0$, o que reduz a dimensão do espaço de $K + 4$ para K, se há K nós. Então, com K nós no interior e dois nos limites, a dimensão do espaço do ajuste é de $K + 2$.

Quando trabalha-se com *splines*, existe uma dificuldade em escolher a localização e quantidade ideal dos nós, sendo mais importante o número de nós do que sua localização. Salienta-se que incluir mais nós que o necessário pode resultar em uma piora do ajuste do modelo. Existem algumas maneiras para fazer essas escolhas, como, por exemplo, colocar os nós nos quantis das variável preditora (três nós interiores nos três quartis).

Outro problema é a escolha de funções base para representar o *spline* para dados nós. Suponha-se que os nós interiores são denotados por $\xi_1 < ... < \xi_k$ e os nós dos limites são ξ_0 e ξ_{k+1}. Uma escolha simples de funções base para um *spline* cúbico é conhecida como base de séries de potência truncada, que deriva de:

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{k} \theta_j (x - \xi_j)_+^3$$

onde $(x - \xi_j)_+ = \max(0, x - \xi_j)$. A função s tem as propriedades necessárias: é um polinômio cúbico em qualquer subintervalo $[\xi_j, \xi_{j+1})$, possui duas derivadas contínuas e possui uma terceira derivada.

Observa-se que *splines* de regressão podem ser atrativos devido a sua facilidade computacional, quando os nós são dados. Porém, a dificuldade em escolher o número e localização dos nós pode ser uma grande desvantagem da técnica.

3.2 Seleção dos parâmetros de suavização

Em regressão, a fim de elaborar boas funções de predição, cria-se um critério para mensurar o desempenho que determinada função predição $g : \mathbb{R}^d \Rightarrow \mathbb{R}$, valendo-se, por exemplo, do do risco quadrático (IZBICKI E SANTOS, 2020).

$$R_{pred}(g) = E\left[(Y - g(X))^2\right].$$

Constata-se que (X, Y) é uma observação nova não utilizada ao se estimar g. Sendo assim, melhor será a função de predição g, quanto menor for o risco. Outras funções de perda pode ser empregadas, porém, a função $L(g(X); Y) = (Y - g(X))^2$ (denominada função de perda quadrática), será utilizada neste trabalho.

Para se medir a performance de um estimador, baseando-se em seu risco quadrático, criar uma boa função de predição equivale a encontrar um bom estimador para a função de regressão, sendo que a melhor função de predição para y é a função de regressão. Normalmente, é habitual ajustar distintos modelos para a função de regressão e encontrar qual deles apresenta um melhor poder preditivo, ou seja, aquele que possui o menor risco. Um modelo pode interpolar os dados e, mesmo assim, possuir um baixo poder preditivo (IZBICKI E SANTOS, 2020).

O método de seleção de modelos pretende selecionar uma boa função g. Nesse sentido, usa-se o critério do risco quadrático para averiguar a qualidade da função. Assim, escolhe-se uma função g em uma classe de candidatos G que tenha um bom poder preditivo (baixo risco quadrático). Dessa maneira, visa-se evitar modelos que tenham sub ou super-ajuste.

O risco observado, conhecido como erro quadrático médio em relação aos dados, e determinado por,

$$EQM(g) = \frac{1}{n} E \sum_{i=1}^{n} \left[(Y_i - g(X_i))^2\right],$$

onde g é escolhida, a fim de minimizar o EQM acima, sendo $\hat{Y}_i = g(X_i)$ o valor predito de Y_i por g. Nota-se que a predição é feita para cada observação, após o ajuste do modelo utilizando todos os dados disponíveis. Contudo, este estimador, se empregado para realizar a seleção de modelos pode levar um super-ajuste (ajuste perfeito aos dados).

Usualmente, é comum dividir os dados em dois conjuntos, um de treinamento e validação. Utiliza-se os dados de treinamento para estimar a regressão e se avalia o erro quadrático médio por meio do conjunto de validação. Este procedimento de divisão é chamado de *data splitting* (IZBICKI E SANTOS, 2020).

Algumas variações podem ser realizadas como o processo *k-fold cross validation* (IZBICKI E SANTOS, 2020), que consiste em dividir a base dados em K conjuntos disjuntos, realizando uma varredura nos conjuntos. Treina-se o modelo com C = K-1 conjuntos e se valida com o conjunto que ficou de fora. Deve-se realizar o rodízio dos K conjuntos até que todos os dados sejam vistos como dados de treino e validação. Alternativamente, pode-se adotar o *leave-one-out cross validation* (LOOCV) (IZBICKI E SANTOS, 2020), no qual o modelo é ajustado utilizando todas as observações com exceção da i-ésima delas, sendo um caso particular da técnica anterior *k-fold* no caso de $K = n - 1$.

O processo para seleção do melhor parâmetro de suavização será aquele que prover o menor erro quadrático médio e pode ser obtido por meio do procedimento abaixo:

Procedimento 2

1. *Supondo o parâmetro suavizador, denotado por p (tamanho do span ou número de nós), para cada valor possível de seu domínio faça:*

   (a) *Considerando um conjunto de dados de tamanho n, para cada observação contidas no conjunto de dados faça:*

      i. *Divida o conjunto de dados, em dados de treino e validação. Considere para os dados de treino todas as observações, exceto i-ésima delas, consequentemente, ter-se-á apenas uma observação compondo os dados de validação*

      ii. *Construa (ajuste) o modelo utilizando apenas os dados de treino.*

      iii. *Utilize o modelo para predizer o valor da resposta ($\widehat{y}_i = g(x_i)$), considerando a observação que compoe os dados de validação e calcule a distância $(y_i - \widehat{y}_i)^2$.*

   (b) *Repita o processo (a) n vezes até que todas as observações sejam "vistas" como dados de teste. Ao final, um vetor de tamanho n das diferenças quadráticas. O erro quadrático médio, para o respectivo parâmetro suavizador, será a média:*

$$EQM(p) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

2. *Repita a etapa (1), para todo o domínio do parâmetro suavizador. O melhor parâmetro suavizador p será aquele que gerou o menor EQM, dentre todos os cadidatos possíveis em seu domínio.*

3.3 Seleção das técnicas de suavização

Após selecionado o melhor parâmetro de suavização, ter-se-á um parâmetro considerado ótimo, que fornecerá o melhor ajuste para um determinado suavizador. Em seguida, consideraremos duas métricas para selecionar a melhor técnica de suavização.

(i) EQM_c: Erro quadrático médio "completo": Após a escolha do parâmetro de suaviazação e tendo em vista todos as observações do conjuntos dados, ajusta-se o modelo. Para cada observação, considere a predição $\hat{y}_i = g(x_i)$, para todo $i = 1, 2, \dots, n$. Em seguida, calcula-se o EQM:

$$EQM_c = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Destacar-se-á a melhor técnica de suavização aquela obter o menor EQM_c dentre as técnicas comparadas.

(ii) EQM_{loocv}: Erro quadrático médio "LOOCV": Essa métrica será exatamente o EQM escolhido no **Procedimento 2** durante a escolha do parâmetro suavizador. Novamente, a melhor técnica de suavização será aquela que obter o menor EQM_{loocv} dentre as técnicas comparadas.

De fato, o menor EQM_c apresentará a técnica com melhor ajuste aos dados e o menor EQM_{loocv} escolherá aquela de maior poder preditivo.

4 Resultados e Discussão

4.1 Estudo de simulação

Nesta seção, serão utilizadas simulações de dados para gerar situações nas quais possam ser aplicadas as técnicas estudadas, analisando suas respectivas performances. Para os resultados obtidos, quatro técnicas de suavização serão empregadas, sendo elas: o suavizador de *kernel*, *Loess*, *splines* de regressão de linear e cúbico. Para as técnicas *kernel* e *loess*, o parâmetro suavizador é indicado, usualmente, pela largura da banda ou *span*. Sendo assim, um valor que indica a proporção de observações a serem utilizadas próximas, ou ainda, nos arredores do ponto de estimação de interesse. Para os *splines* de regressão, o parâmetro de suavização indicado, normalmente, pela quantidade de nós utilizados para os ajustes. No presente trabalho, a quantidade de nós fora considerado, sendo que a posição e localização deste nós fora mantidas fixas, equidistantes espaçadas, em K percentis distintos.

Realizar-se-ão ajustes para o primeiro cenário, considerando distintos parâmetros de suavização, avaliando, visualmente, os comportamentos das curvas em diagramas de dispersão. Em seguida, adotando o método de *data splitting*, *leave one out cross-validation*, encontrar-se-á um parâmetro de suavização que forneça a ocorrência do menor erro quadrático médio (EQM_{loocv}) possível, desta forma, evitando um super-ajuste do modelo. Ademais, ajustes serão executados tendo em consideração tais parâmetros e, em seguida, calcular-se-á os erros quadráticos médios (EQM_c) para cada técnica suavizadora, entre os valores observados e estimativas do modelo.

Posteriormente, este procedimento será repetido para cada cenário em mil amostras, contabilizando a quantidade de vezes em que cada técnica apresenta o menor erro quadrático médio. Por exemplo, para o primeiro cenário, será gerado mil amostras aleatórias de tamanho n. Para cada amostra será empregado o procedimento acima, salvando seus repectivos erros quadráticos médio. Ao final da simulação, será contabilizado se a ocorrência do erro quadrático médio em cada técnica foi mínima e, por fim, comparar e verificar qual técnica obtém o melhor resultado em uma simulação de mil amostras. Primeiramente, será realizada escolha do melhor parâmetro de suavização selecionados por meio da métrica EQM_{loocv}. Considerar-se-a a técnica com o melhor desempenho de predição a que obtiver o menor EQM_{loocv}. Em seguida, comparar-se-á dentro dos mesmos cenários, avaliando por meio da métrica EQM_c, para, assim, concluir qual dos suavizadores são mais aderentes aos dados.

Vale ressaltar, que serão empregados dois comportamentos, um proveniente de uma função senoidal e outra de uma função Gamma: Cenário 1 e Cenário 2. Não obstante, serão gerados nove sub-cenários, valendo-se da combinação de três tamanhos amostrais (150, 250 e 350), em três valores de desvio padrão distintos.

4.1.1 Cenário 1

Para este cenário, será considerado X uma sequência de 0 a 50 e Y, definido pela função

$$y = 10 + 5sen\pi\frac{x}{24} + \varepsilon,$$

onde ε é um termo aleatório, provenientes de uma distribuição normal, com média zero e variância constante. Os tamanhos amostrais utilizados serão iguais a $150, 250$ e 350 e valores de desvio padrão $0, 5, 1$ e 2. Na Figura 1, tem-se o comportamento dos dados para cada desvio padrão, tendo em conta 350 observações com a curva : $10 + 5sen\pi\frac{x}{24}$.

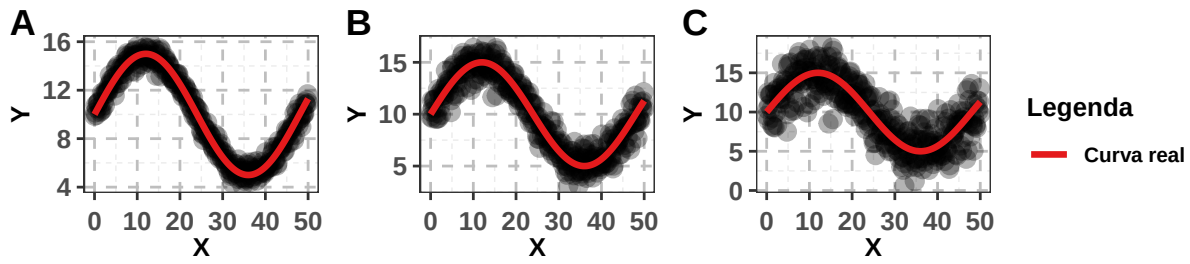

Figura 1: Gráfico de dispersão dos dados simulados e curva real, para o Cenário 1. (A) DP = 0,5, (B) DP = 1 e (C) DP = 2.

Levando em conta a configuração do gráfico C (Figura 1), curvas distintas para cada método de suavização foram ajustadas, adotando parâmetros de suavização arbitrários e são apresentados na Figura 2.

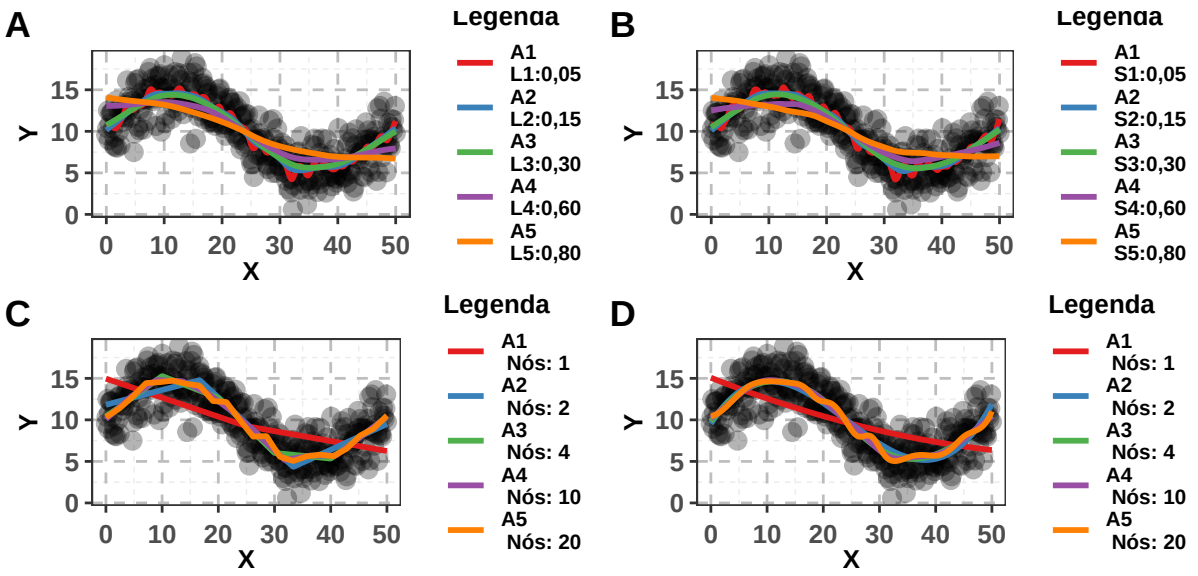

Figura 2: Comparação entre diferentes ajustes (Cenário 1), com parâmetros de suavização distintos, considerando os suavizadores, (A) Kernel, (B) Loess, (C) Splines de Regressão Linear e (D) Splines de Regressão Cúbico.

Ao avaliar os ajustes dos gráficos A e B, verifica-se que, conforme o parâmetro de suavização aumenta, estes tendem a ser muito suaves. Ou seja, na medida em que o parâmetro se torna suficientemente grande, a curva tenderá a ser uma reta. Porém, quando este parâmetro tende a ser muito pequeno, o ajuste realizado interpolará os dados. Para os gráficos C e D, na proporção em que o parâmetro de suavização aumenta, a curva ajustada apresenta rugosidade em sua forma. Conforme este parâmetro se torna pequeno, a curva tenderá a ser uma reta.

Os suavizadores em diagramas de dispersão remetem a uma ideia visual de como o ajuste está se comportando em relação aos dados. Na Figura 3, são apresentados os gráficos contendo os resultados do erro quadrático médio mínimo obtidos, realizando o *leave one out cross-validation* (EQM_{loocv}). Neste gráfico, verifica-se o comportamento dos erros quadrático médio em relação a seus repectivos parâmetros de suavização. Entretanto, nota-se para qual parâmetro de suavização haverá o melhor ajuste (evitando super-ajuste), levando em consideração o menor erro quadrático possível dentre todos os candidatos. Em outras palavras, ao realizar ajustes controlando o valor do parâmetro de suavização, obter-se-á um ajuste no qual o Erro Quadrático Médio (EQM_{loocv}) será o menor de todos, logo, este será o melhor candidato para representar os dados. Observa-se, então, os parâmetros que, supostamente, induzirão um melhor ajuste para cada suavizador.

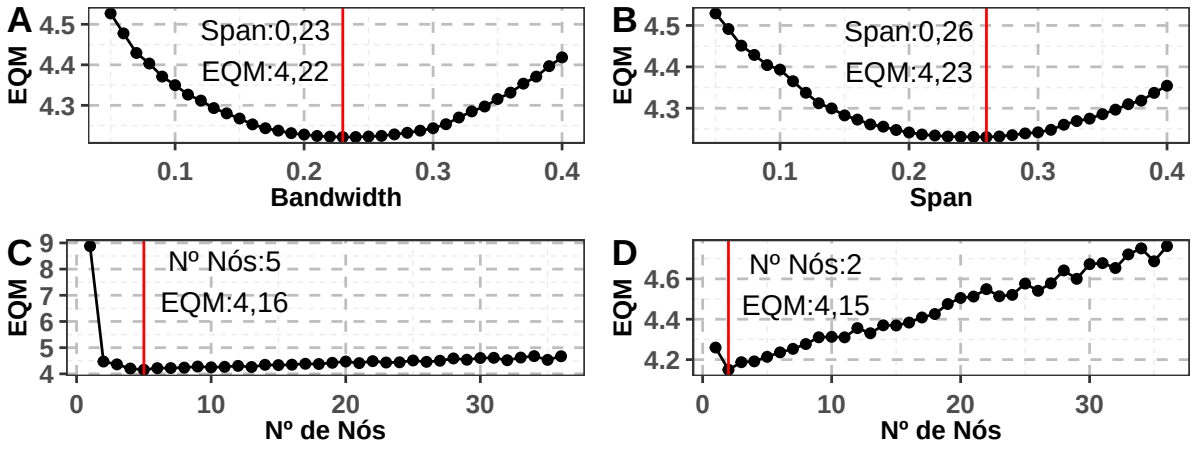

Figura 3: Erro quadrático médio (EQM_{loocv}) versus parâmetro de suavização (Cenário 1) pós aplicação do Leave One Out Cross-Validation. (A) Kernel, (B) Loess, (C) Splines de Regressão Linear e (D) Splines de Regressão Cúbico.

Com o auxilio da Tabela 1, concluir-se-á que o melhor método de suavização para predição, levando em consideração o Cenário 1 com apenas uma amostra, é o *splines* de regressão cúbico por obter o menor erro quadrático possível. Porém, observando e analisando os resultados obtidos, apenas em uma amostra pode levar à decisões equivocadas.

Tabela 1: Erro Quadrático Médio (EQM_{loocv}), para os suavizadores Kernel, Loess e Splines de Regressão Linear e Cúbico.

Suavizador	Parâm. Suavizador	EQM
Kernel	0,23	4,22
Loess	0,26	4,23
Sp. Reg. Linear	5	4,16
Sp. Reg. Cúbico	2	4,15

Na Figura 4, são demonstrados os ajustes, levando em consideração os melhores parâmetros obtidos por meio do processo de validação cruzada.

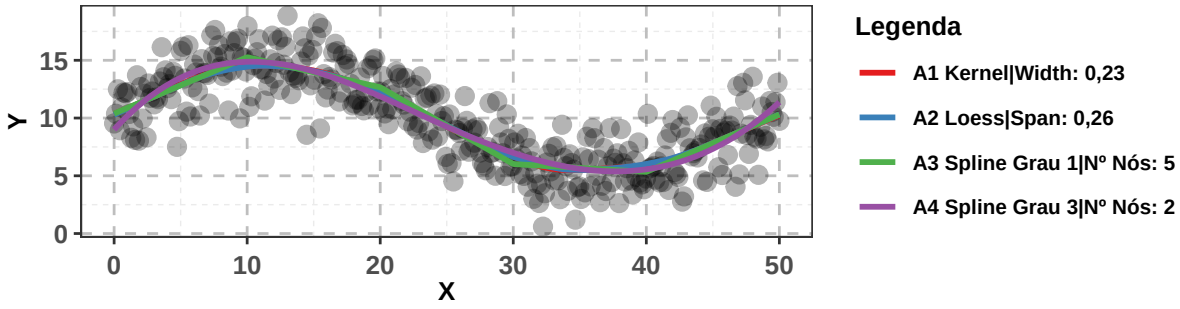

Figura 4: Comparação entre os ajustes, considerando parâmetros de suavização obtidos por meio da validação cruzada.

Na Tabela 2, é mostrado os EQM's provenientes dos ajustes apresentados na Figura 4. Conclui-se que a téncnica suavizadora que melhor se ajusta aos dados simulados (Cenário 1) é o *Loess*.

Tabela 2: Erro Quadrático Médio (EQM_c), para os suavizadores Kernel, Loess e Splines de Regressão Linear e Cúbico.

Suavizador	Parâm. Suavizador	EQM
Kernel	0,23	0,2652
Loess	0,26	0,2634
Sp. Reg. Linear	5	0,2674
Sp. Reg. Cúbico	2	0,2723

Neste momento, será reanalisado este procedimento em um processo de geração de amostras aleatórias. Serão considerados os cenários apresentados no começo deste capítulo: três tamanhos amostrais, para três variabilidades, totalizando nove cenários distintos. Para cada cenário serão geradas mil amostras aleatórias, aplicando o procedimento *leave one out cross-validation* em cada método de suavização. Em seguida, contabilizar-se-á a quantidade de vezes que cada

técnica obteve erro quadrático médio (EQM_{loocv}) foi mínimo. Também, será avaliado qual suavizador apresenta o melhor ajuste para representar os dados simulados, por meio do cálculo dos erros quadrático médio completo (EQM_c), levando em conta o valores observados e ajustados de cada técnica. Por fim, avaliar-se-á qual técnica obteve o melhor desempenho de predição e ajuste.

Na Tabela 3, estão esquematizados os resultados obtidos, por meio do *leave one out cross-validation* para cada cenário, tendo em mente mil amostras simuladas para os suavizadores *Kernel*, *Loess*, *Splines* de regressão Linear e Cúbico.

Tabela 3: Percentual do Erro quadrático médio EQM_{loocv} mínimo, obtidos por meio de aplicação do Procedimento 1, para seleção do melhor parâmetro de suavização, considerando cada suavizador em 1000 amostras.

Sub-Cenário	Tamanho	Desvio Padrão	Kernel	Loess	Sp. Reg. Linear	Sp. Reg. Cúbico
1	150	0,5	0,9%	5,3%	19,7%	74,1%
2	150	1,0	1,4%	8,4%	31,1%	59,1%
3	150	2,0	1,4%	10,8%	48,0%	39,8%
4	250	0,5	0,7%	3,8%	16,9%	78,6%
5	250	1,0	1,1%	7,0%	23,5%	68,4%
6	250	2,0	1,7%	13,1%	35,8%	49,4%
7	350	0,5	0,8%	3,7%	15,8%	79,7%
8	350	1,0	1,3%	7,0%	20,5%	71,2%
9	350	2,0	1,8%	11,7%	33,5%	53,0%

Destaca-se o suavizador *splines* cúbico por obter o melhor desempenho em quase todos os cenários, exceto o terceiro (Sub-Cenário 3), no qual o *splines* de linear obteve erro quadrático médio mínimo em cerca de 48% das amostras simuladas. Ainda, quando fixamos o valor do tamanho amostral, constata-se que, conforme a variabilidade dos dados aumenta, há indícios de que os resultados obtidos para a técnica *splines* cúbico estejam se dispersando para as demais técnicas. Observa-se que conforme o tamanho amostral aumenta, os percentuais estejam convergindo e estabilizando, envidenciando que os *splines* cúbico tendem a obter um desempenho melhor em relação aos demais métodos. Ademais, ressalta-se que os percentuais para o suavizador de *kernel* foram os piores, obtendo EQM's mínimos de até no máximo 1,8% das amostras em relação aos cenários simulados.

Na Figura 5, são apresentados os comportamentos dos erros quadráticos médios obtidos do processo de simulação das amostras. De forma sucinta, verifica-se uma tendência decrescente conforme o tamanho amostral aumenta, assim, como a sua aplitude tende a ficar menor. Outrossim, visualmente, não é observada diferença significativa entre as técnicas *Kernel* e *Loess*. Aliás, percebe-se que os *splines* (linear e cúbico) apresentam um comportamento mediano relativamente inferior quando comparado aos demais.

Portanto, levando em consideração a Tabela 3 e a Figura 5 pode-se concluir que, para o Cenário 1, o suavizador *splines* de regressão cúbico tende demonstrar um melhor desempenho

de predição em relação às demais técnicas.

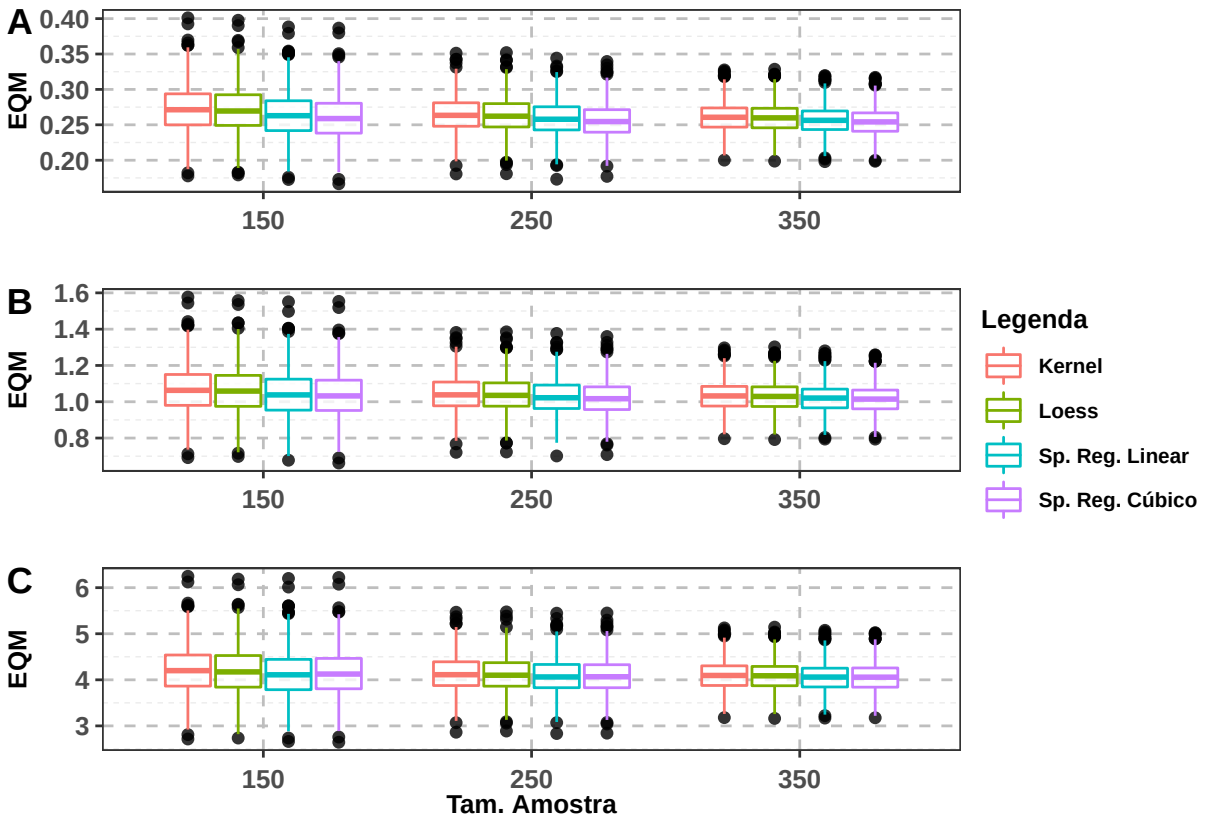

Figura 5: Comparação do erro quadrático para as 1000 amostras para cada suavizador. (A) DP = 0.5, (B) DP = 1 e (C) DP = 2

Na Tabela 4 são apresentados os percentuais, resultantes da contabilização do menor erro quadrático médio (EQM_c) para os melhores ajustes obtidos por meio da seleção do melhor parâmetro suavizador.

Tabela 4: Percentual do Erro quadrático médio EQM_{loocv} mínimo, considerando os ajustes provenientes dos melhores parâmetros obtidos por meio de aplicação do Procedimento 1, para cada suavizador em 1000 amostras.

Sub-Cenário	Tamanho	Desvio Padrão	Kernel	Loess	Sp. Reg. Linear	Sp. Reg. Cúbico
1	150	0,5	65,3%	15,4%	11,7%	7,6%
2	150	1,0	61,3%	11,8%	14,6%	12,3%
3	150	2,0	63,2%	11,2%	11,3%	14,3%
4	250	0,5	67,2%	17,1%	11,0%	4,7%
5	250	1,0	66,1%	14,0%	12,7%	7,2%
6	250	2,0	56,2%	12,5%	13,9%	17,4%
7	350	0,5	79,9%	8,8%	8,7%	2,6%
8	350	1,0	63,3%	16,8%	13,1%	6,8%
9	350	2,0	56,2%	12,4%	13,8%	17,6%

Verifica-se que o suavizador com *kernel* gaussiano obteve o melhor desempenho de

ajuste em todos os cenários propostos, sendo considerada a melhor técnica em no mínimo 56,2% (Sub-Cenário 6 e 9) das amostras simuladas. Em seguida, o suavizador *Loess* apresenta um melhor desempenho em quatro sub-cenários, em relação as técnicas restantes, seguido do *splines* de regressão linear e, por fim, o *splines* o cúbico. A Figura 6 contém o comportamento para os EQM_c. Quando comparadas as técnicas *kernel* e *Loess*, em relação ao seus comportamentos medianos, visualmente não aparentam uma diferença significativa. Porém, em alguns dos cenários, a mediana dos $EQM_c's$ para suavizador *kernel* aparenta ser relativamente inferior. Ao comparar os $EQM_c's$ medianos das técnicas, *splines* de regressão linear e cúbico estes mostram-se ser relativamente maiores em relação ao *kernel* e *loess*, e não possuem diferença significativa. Ressalta-se o ajuste paramétrico, regressão cúbico, apresentando o pior comportamento para os $EQM_c's$, indicando um comportamento mediano significativo superior quando comparada às demais técnicas.

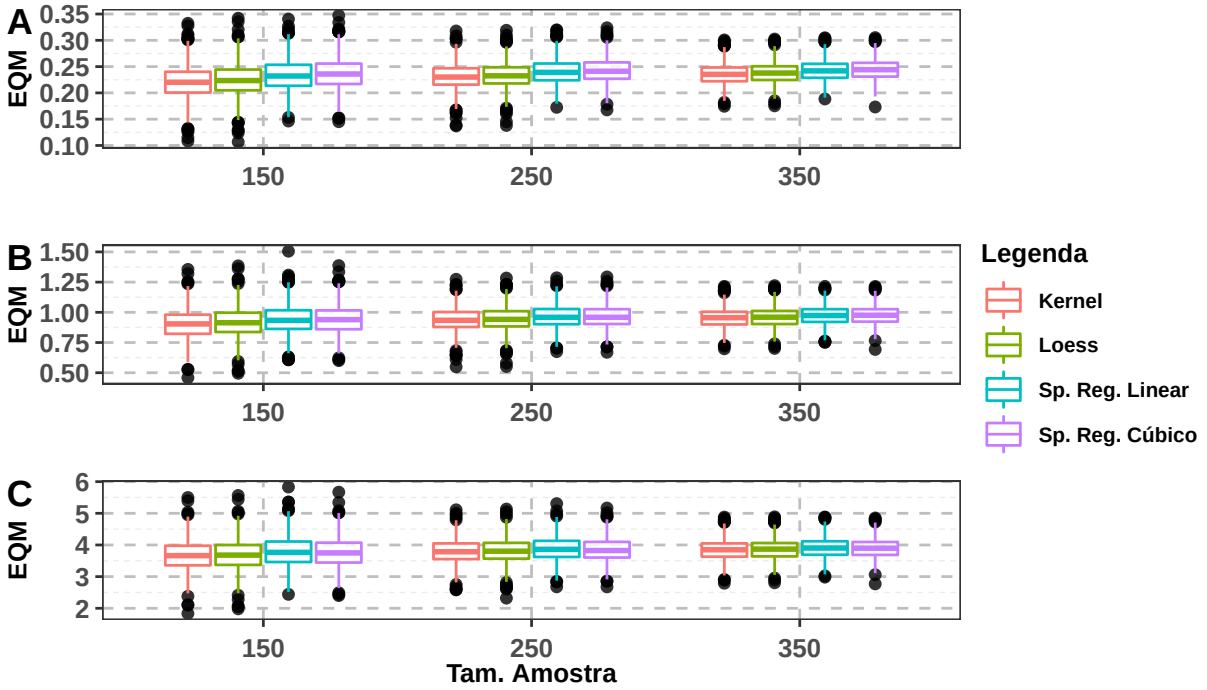

Figura 6: Comparação do erro quadrático para as 1000 amostras para cada suavizador. (A) DP = 0.5, (B) DP = 1 e (C) DP = 2

4.1.2 Cenário 2

Para este cenário, os valores de x serão uma sequência de 0.1 à 2. Ainda, temos que $y = f(x) + \varepsilon$, com $f(x) \sim Gamma(6, 10)$ e $\varepsilon \sim N(0, \sigma^2)$. Serão considerados tamanhos amostrais iguais a $150, 250$ e 350 e valores de desvio padrão $0.05, 0.1$ e 0.15. Na Figura 7, tem-se o comportamento dos dados para cada desvio padrão, presumindo 350 observações.

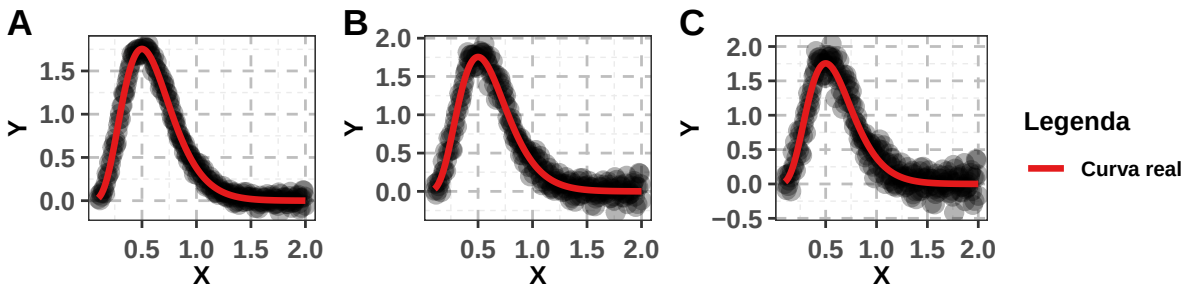

Figura 7: Gráfico de dispersão dos dados simulados e curva real, para o Cenário 2. (A) DP = 0,05, (B) DP = 0,10 e (C) DP = 0,15.

Quando analisado os resultados obtidos do procedimento de simulação de amostras, que se encontra sumarizado na Tabela 5, observa-se que o suavizador *splines* de regressão cúbico demonstrou o melhor desempenho em todos cenários, obtendo o EQM_{LOOCV} mínimo, de aproximadamente 70% à 92% das amostras entre os cenários simulados. Outrossim, ressalta-se que, conforme o tamanho amostral aumenta, o percentual apresenta uma têndencia de aumento e indícios de estabilização. Quando fixado um tamanho amostral, o desempenho entre as amostras simuladas com variabilidade distinta apresenta uma tendência decrescente.

Tabela 5: Percentual do Erro quadrático médio EQM_{loocv} mínimo, considerando os ajustes provenientes dos melhores parâmetros obtidos por meio de aplicação do Procedimento 1, para cada suavizador em 1000 amostras.

Sub-Cenário	Tamanho	Desvio Padrão	Kernel	Loess	Sp. Reg. Linear	Sp. Reg. Cúbico
1	150	0,05	0,6%	2,6%	7,4%	89,4%
2	150	0,10	0,7%	6,3%	13,5%	79,5%
3	150	0,15	0,7%	9,1%	20,6%	69,6%
4	250	0,05	0,4%	1,7%	8,2%	89,7%
5	250	0,10	0,7%	3,4%	10,3%	85,6%
6	250	0,15	0,8%	6,5%	12,7%	80,0%
7	350	0,05	0,2%	2,0%	5,9%	91,9%
8	350	0,10	0,6%	2,5%	8,9%	88,0%
9	350	0,15	1,0%	4,7%	11,4%	82,9%

De forma análoga ao Cenário 1, o comportamento para os $EQM'_{LOOCV}s$ (vide Figura 8), apresenta uma tendência decrescente conforme o tamanho amostral aumenta, e, percebe-se que a amplitude tende a ficar menor. Ademais, percebe-se que não há diferença signficativa entre os suavizadores *kernel* e *loess*. Os *splines* apresentam ter um comportamento mediano inferior, com destaque ao *splines* cúbico que, visualmente, aparenta ser menor quando comparados com as demais técnicas. Sendo assim, baseado na Tabela 5 e na Figura 8, os dados evidenciam que o suavizador *splines* de regressão cúbico possui o melhor desempenho de predição.

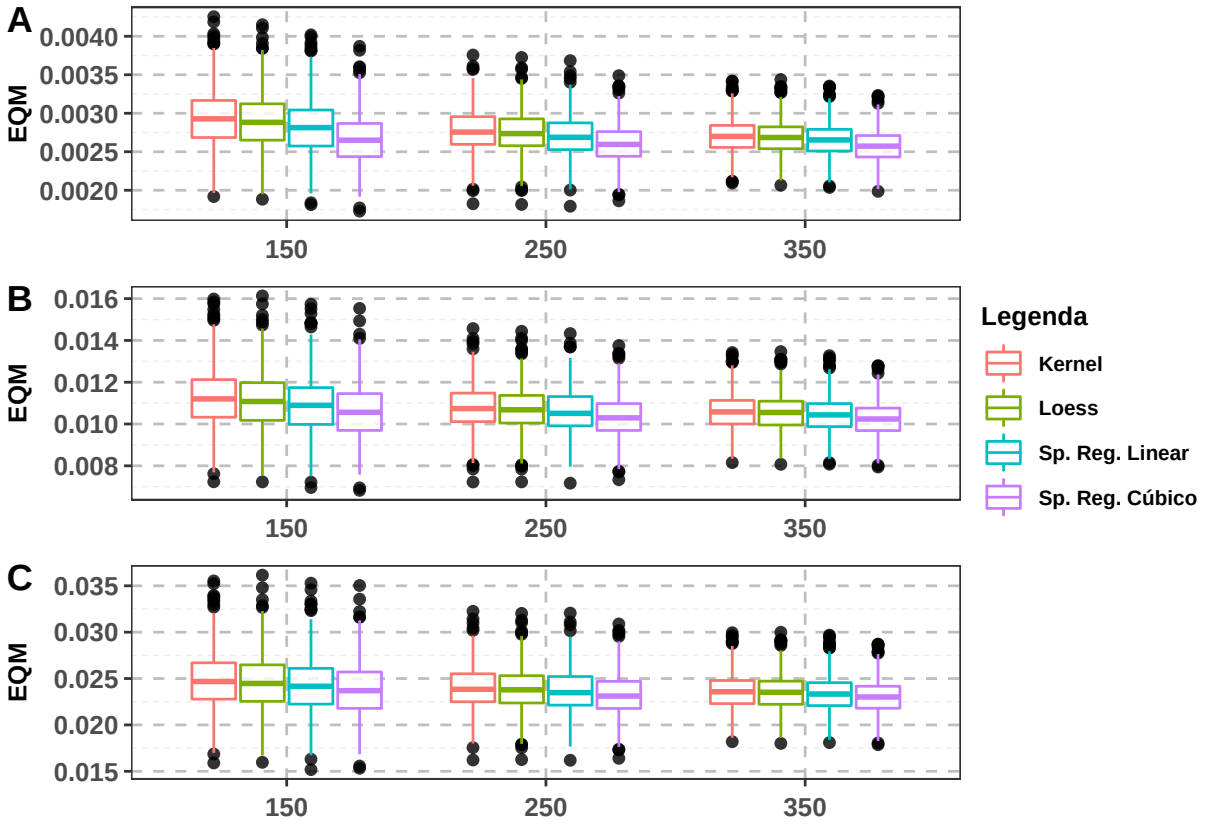

Figura 8: Comparação do erro quadrático médio (EQM_{loocv}) em 1000 amostras para cada suavizador. (A) DP = 0.05, (B) DP = 0.1 e (C) DP = 0.15

Em relação ao desempenho do melhor ajuste, este pode ser observado na Tabela 6. Novamente, destaca-se o suavizador *kernel*, por obter o melhor desempenho em todos os sub-cenários, de no mímino 49,1% das amostras simuladas. Em seguida, comparando as técnicas restantes, destaca-se o método de splines de regressão linear por se destacar em quatro sub-cenários, seguidos do *loess* e, por último, o *splines* de regressão cúbico.

Tabela 6: Percentual do Erro quadrático médio EQM_{loocv} mínimo, considerando os ajustes provenientes dos melhores parâmetros obtidos por meio de aplicação do Procedimento 1, para cada suavizador em 1000 amostras.

Sub-Cenário	Tamanho	Desvio Padrão	Kernel	Loess	Sp. Reg. Linear	Sp. Reg. Cúbico
1	150	0,5	90,1%	6,0%	3,8%	0,1%
2	150	1,0	71,1%	9,3%	16,3%	3,3%
3	150	2,0	49,1%	13,5%	25,6%	11,8%
4	250	0,5	91,9%	4,2%	3,5%	0,4%
5	250	1,0	79,9%	7,6%	10,7%	1,8%
6	250	2,0	61,9%	10,6%	20,3%	7,2%
7	350	0,5	91,9%	3,7%	3,7%	0,7%
8	350	1,0	82,0%	7,6%	8,7%	1,7%
9	350	2,0	65,4%	10,3%	20,3%	4,0%

Na Figura 9, observa-se que o suavizador *kernel* apresenta um comportaemnto mediano relativamente inferior aos demais suavizadores. Seguidos do *loess*, *splines* de regressão linear e, por último, o *splines* de regressão cúbico. Portanto, levando em conta a Tabela 6 e Figura 9, há indícios de que o suavizador com *kernel* gaussiano apresenta o melhor desempenho para ajustar os dados.

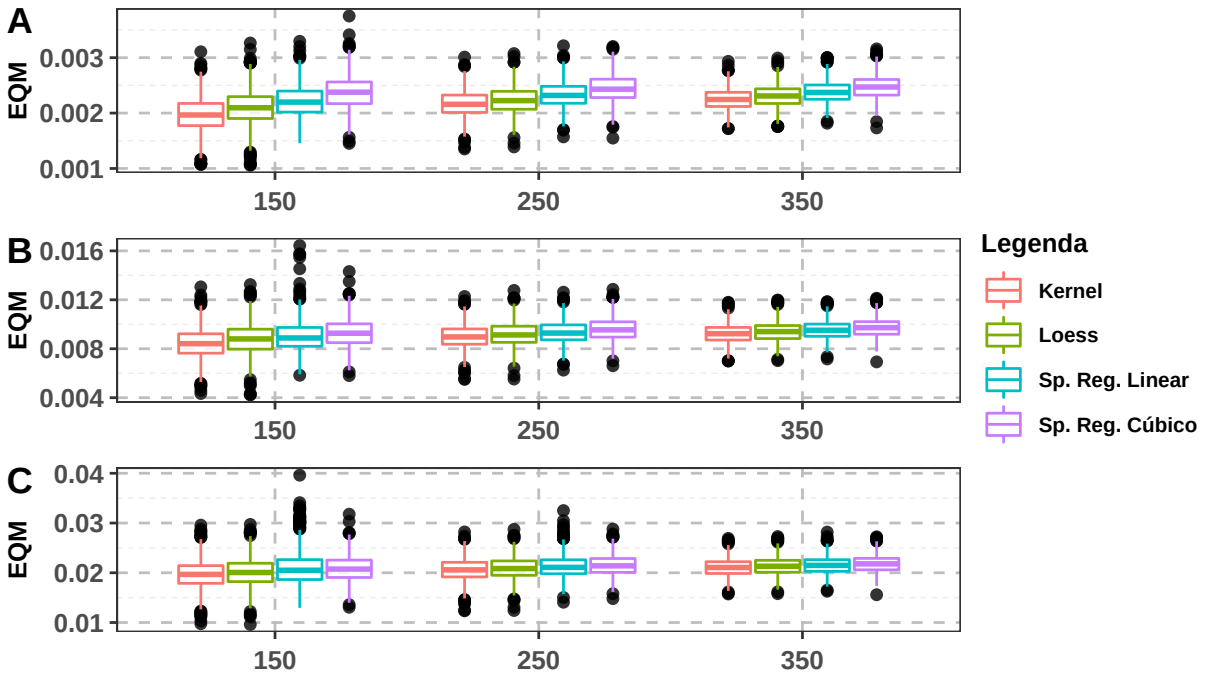

Figura 9: Comparação do erro quadrático médio (EQM_c), para as 1000 amostras para cada suavizador. (A) DP = 0.05, (B) DP = 0.1 e (C) DP = 0.15

4.1.3 Considerações finais do estudo de simulação

Pode-se comparar as técnicas de suavização apresentadas neste trabalho em diversos cenários mostrando quais delas apresentam o melhor poder preditivo e quais delas ajustam melhor os dados. Segundo IZBICKI (2020), nem sempre o modelo que se adequa melhor aos dados, necessariamente irá obter o melhor poder preditivo. Verificou-se dentro das simulações que o suavizador *splines* de regressão cúbico obteve o melhor desempenho de predição, avaliando-se por meio da métrica EQM_{LOOCV}, em ambos os cenários propostos (Cenário 1 e Cenário 2). Em contrapartida, ao avaliar os cenários por meio da métrica EQM_c, pode-se verificar que o suavizador com *kernel* gaussiano obteve os melhores resultados.

5 Aplicações

5.1 Aplicação 1

Para esta aplicação serão empregadas as técnicas de suavização em dados reais. Os dados foram retirados do site NIST Standard Reference Database 140 [1] . É um estudo referente à expansão térmica de cobre. A variável resposta é o coeficiente de expansão térmica e a variável preditora é a temperatura em graus kelvin. Neste trabalho, será abordado um modelo com apenas uma covariável neste caso, sendo o modelo aditivo da seguinte forma

$$y = \alpha + f(x) + \varepsilon,$$

onde os erros ϵ são independentes, com $E(\varepsilon) = 0$ e $var(\varepsilon) = \sigma^2$. A $f(x)$ é uma função univariada arbitrária, que será suavizada pelos métodos vistos até o momento.

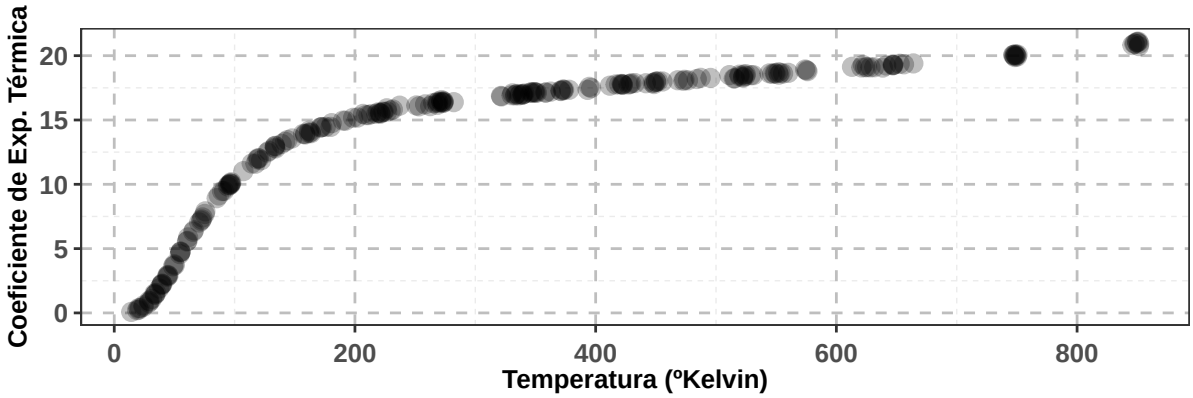

Figura 10: Gráfico de dispersão referente ao estudo de espansão térmica do cobre versus a temperatura em graus Kelvin.

Na Tabela 7, é apresentado o processo de seleção do melhor parâmetro, selecionados por meio da métrica EQM_{LOOCV}.

Tabela 7: Erro Quadrático Médio (EQM_{loocv}) para os suavizadores Loess, Kernel e Splines de Regressão Linear e Cúbico (Aplicação 1).

Suavizador	Parâm. Suavizador	EQM
Kernel	0,1	0,0091
Loess	0,1	0,0078
Sp. Reg. Linear	20,0	0,0078
Sp. Reg. Cúbico	14,0	0,0065

[1] https://www.itl.nist.gov/div898/strd/index.html

Ressalta-se que o *splines* de regressão cúbico apresenta o menor EQM_{LOOCV}. Na Figura 11, são demonstrados os ajustes, levando em consideração os melhores parâmetros obtidos por meio do processo de validação cruzada. Um ajuste parâmétrico obtido por meio de uma regressão polinomial cúbica fora ajustada (Figura 11, gráfico E), e será comparado com ajustes obtidos pelo métodos de suavização. Visualmente, observa-se que as técnicas suavizadoras apresentam um bom ajuste para captar a tendência de expansão térmica em relação a temperatura. Como pode-se observar o ajuste paramétrico não consegue descrever o comportamento de forma adequada para estes dados.

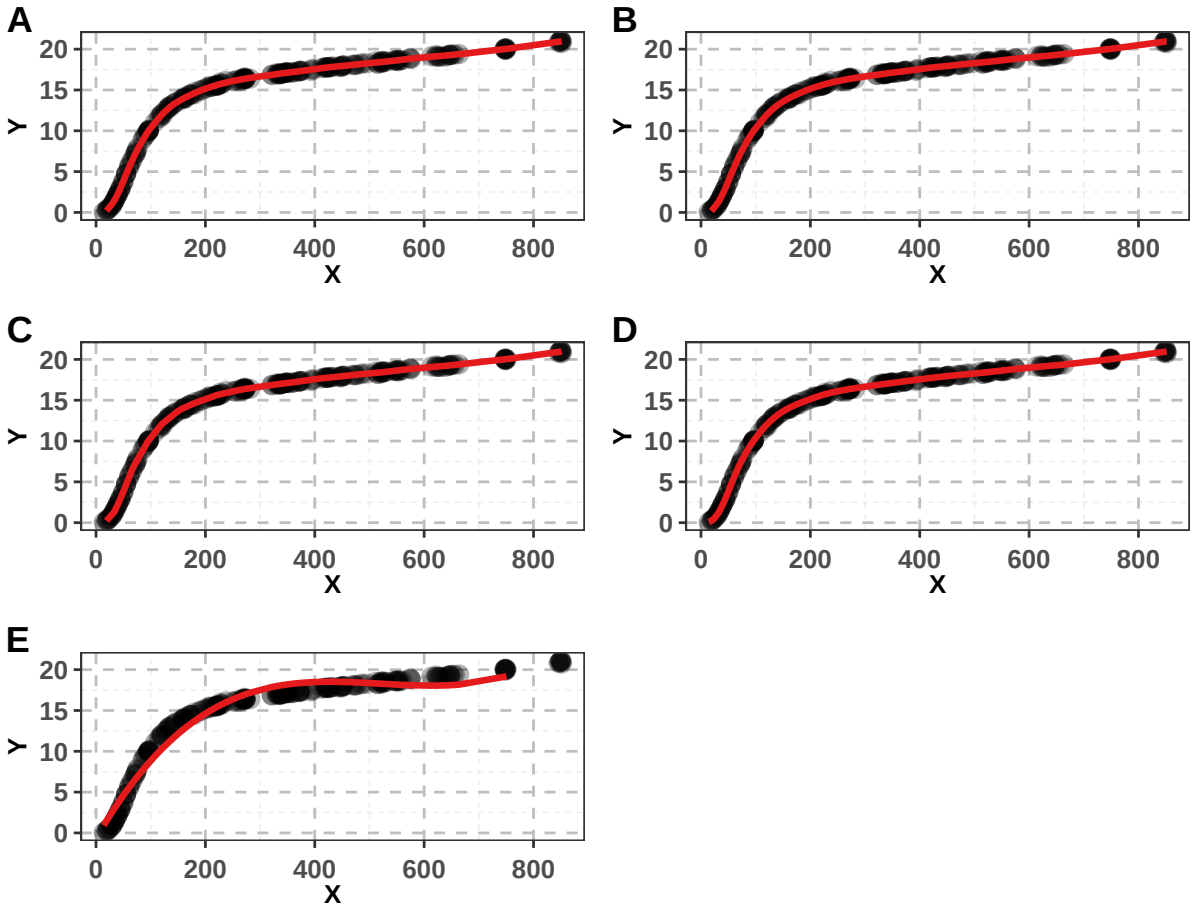

Figura 11: Comparação entre os ajustes, considerando parâmetros de suavização obtidos por meio da validação cruzada. (A) Kernel - Parâm. 0,1. (B) Loess - Parâm. 0,1. (C) Splines de Regressão Linear - Parâm. 20. (D) Splines de Regressão Cúbico - Parâm. 14. (E) Regressão Polinômial Cúbica.

Ao comparar os $EQM_c's$, na Tabela 8, ressalta-se que a regressão polinomial cúbica obteve o pior resultado para esta métrica. Para as técnicas *kernel*, *loess* e *splines* de regressão linear apresentaram valores bem próximos entre si. Destaca-se o *splines* de regressão por obter o menor erro quadrático médio (EQM_c).

Portanto, levando em consideração os resultados obtidos, o suavizador em *splines* de regressão cúbico, apresenta o melhor desempenho tanto em relação ao poder de predição quanto melhor técnica para ajustar, o comportamento de expansão térmica em relação a temperatura.

Tabela 8: Erro Quadrático Médio (EQM_c) para os suavizadores Loess, Kernel e Splines de Regressão Linear e Cúbico.

Suavizador	Parâm. Suavizador	EQM
Kernel	0.1	0,000200
Loess	0.1	0,000211
Sp. Reg. Linear	20	0,000211
Sp. Reg. Cúbico	14	0,000171
Polinômio Cúbico		0,031052

5.2 Aplicação 2

Para esta aplicação, a avaliação das técnicas de suavização serão aplicadas em dados referentes ao preço mediano de casas de Boston. Este conjunto de dados possui 506 observações e 14 variáveis. Será empregada considerado para a variável resposta y, a coluna *medv* que representa o valor mediano das casas ocupadas pelos proprietários em \$1000s. Para a variável preditora x, será considerado a coluna *lstat* que representa o percentual de status baixo da população. Tendo em conta o modelo aditivo da seguinte forma,

$$y = \alpha + f(x) + \varepsilon,$$

onde os erros ε são independentes, com $E(\varepsilon) = 0$ e $var(\varepsilon) = \sigma^2$. A $f(x)$ é uma função univariada arbitrária, que será suavizada pelos métodos vistos até o momento.

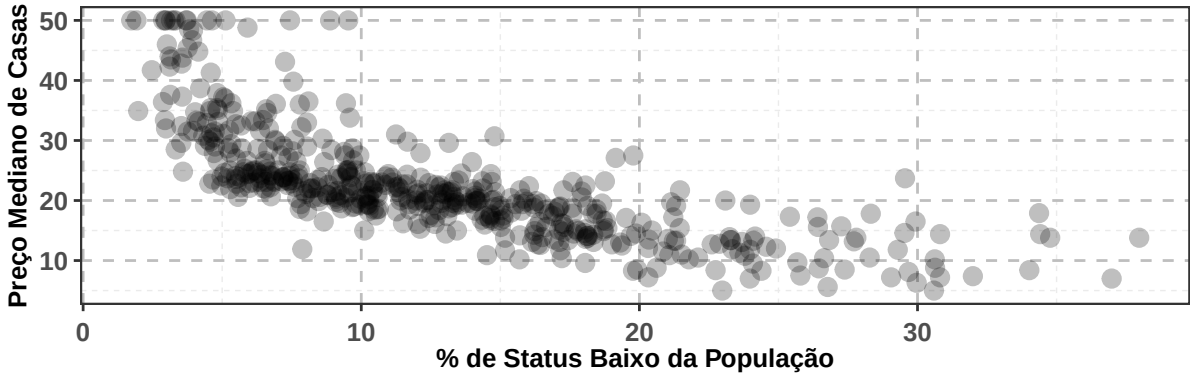

Figura 12: Gráfico de dispersão referente ao estudo de espansão térmica do cobre versus a temperatura em graus Kelvin.

Na Tabela 9, são apresentados os melhores parâmetros selecionados por meio da métrica EQM_{LOOCV}. Destaca-se que o suavizador *splines* de regressão linear apresenta o menor EQM_{loocv}.

Tabela 9: Erro Quadrático Médio (EQM_{loocv}) para os suavizadores Loess, Kernel e Splines de Regressão Linear e Cúbico (Aplicação 1).

Suavizador	Parâm. Suavizador	EQM
Kernel	0,21	27,1909
Loess	0,18	27,2788
Sp. Reg. Linear	5,00	27,1191
Sp. Reg. Cúbico	5,00	27,3552

De forma análoga a Aplicação 1, a Figura 13 apresenta os melhores ajustes para os suavizadores comparados com um ajuste polinomial cúbico. Visualmente, observa-se que todos os modelos conseguem captar e representar de forma satisfatória o comportamento do preço mediano de casas conforme o percentual de status baixo da população aumenta.

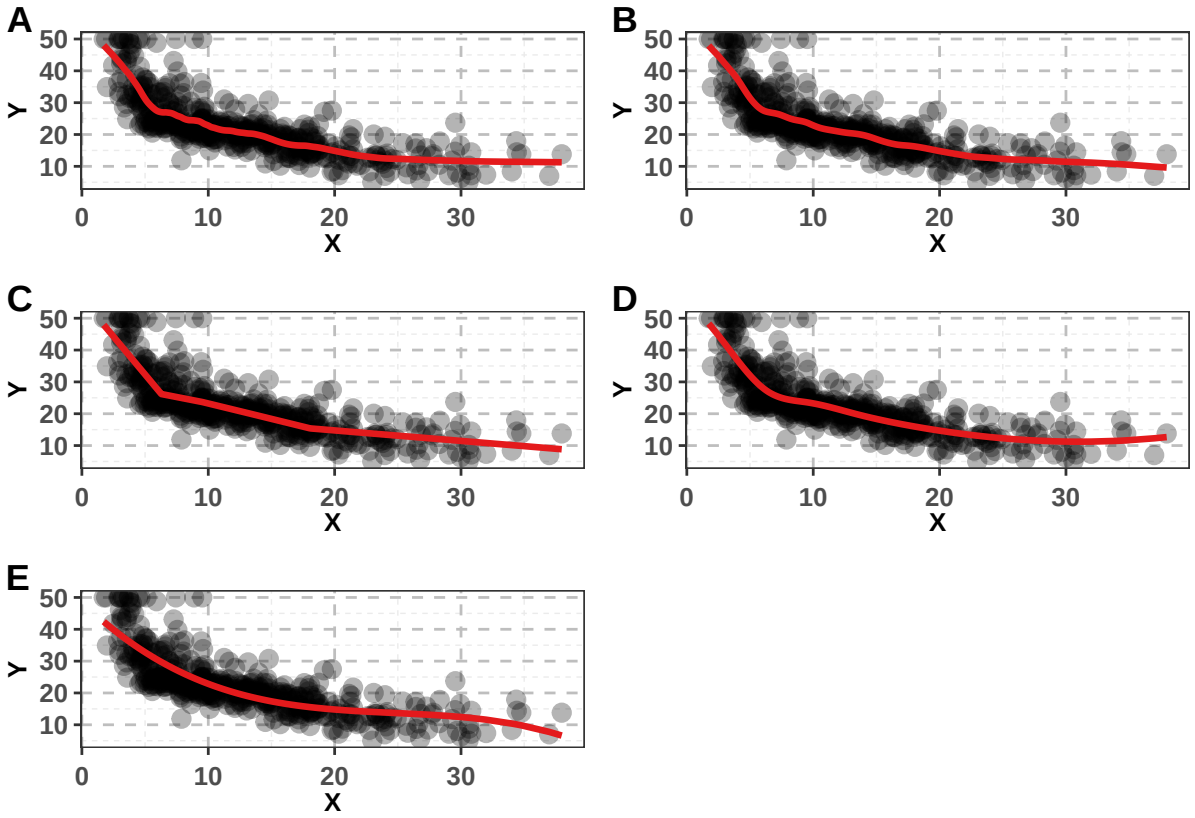

Figura 13: Comparação entre os ajustes, considerando parâmetros de suavização obtidos por meio da validação cruzada. (A) Kernel - Parâm. 0,21. (B) Loess - Parâm. 0,18. (C) Splines de Regressão Linear - Parâm. 5 (D) Splines de Regressão Cúbico - Parâm. 5 (E) Regressão Polinômial Cúbica.

Ao comparar os $EQM_c's$, na Tabela 10, ja era esperado que a regressão polinomial cúbica desempenhasse o pior resultado para esta métrica. Os *splines* de regressão linear e cúbico apresentam valores bem próximo entre si. Destaca-se que o método de suavização com *kernel* gaussiano obteve o melhor desempenho levando em consideração esta métrica.

Tabela 10: Erro Quadrático Médio (EQM_c) para os suavizadores Loess, Kernel e Splines de Regressão Linear e Cúbico.

Suavizador	Parâm. Suavizador	EQM
Kernel	0.21	0,307130
Loess	0.18	0,308202
Sp. Reg. Linear	5	0,315720
Sp. Reg. Cúbico	5	0,314639
Polinômio Cúbico		0,342152

Portanto, levando em consideração estes resultados, o suavizador em *splines* de regressão linear, obtém os melhores resultados de predição. Porém, varificou-se que a técnica que melhor se adequa para representar e captar a tendência do preço mediano de casas em relação ao percentual dos status baixo na população, é o suavizador *kernel* gaussiano.

6 Conclusão

Para investigar e modelar relações entre variáveis, o modelo de regressão linear pode ser utilizado, porém, quando essa relação não possui forma linear, uma alternativa é o uso de ferramentas que não impõem suposições paramétricas. Nesse contexto, existem técnicas de suavização que podem ser utilizadas, inclusive, na estimação das funções do componente sistemático dos modelos aditivos.

As técnicas de suavização *kernel*, *loess* e splines de regressão, em particular, os de grau um e grau três foram apresentadas, que são utilizadas para estimar as funções presentes em modelos aditivos. Fora introduzido um método para obtenção no melhor parâmetro suavizador, a fim de evitar sub ou super-ajuste, realizado por meio de método de validação cruzada. Duas métricas foram abordadas: uma para verificar a qualidade de predição, erro quadrático médio obtidos por validação cruzada, *leave one out cross validation* (EQM_{loocv}) e a outra para apurar a qualidade dos ajustes, erro quadrático médio completo (EQM_c).

Para validar a metodologia estudada, foram realizadas análises em dados simulados e dados reais. Em dados simulados de diferentes cenários, foram observados os resultados em relação ao comportamento de técnicas de suavização, comparando os modelos obtidos, por meio das métricas introduzidas anteriormente. Ainda, verificou-se que o ajuste mais adequado para descrever o comportamento dos dados não obtém necessariamente o melhor poder preditivo.

Por meio dos resultados obtidos do estudo de simulação, pode-se concluir que, ao avaliar e comparar os ajustes considerando a métrica de qualidade de predição (EQM_{loocv}) e a métrica de qualidade de ajuste (EQM_c), tanto para o Cenário 1, quanto para o Cenário 2, o suavizador que obteve o melhor poder preditivo foi o *splines* de regressão cúbico. Levando em consideração o método com melhor qualidade de ajustes, os suavizadores com *kernel* se destacou em ambos os cenários.

Finalmente, os métodos discutidos foram aplicados em dados reais, Aplicação 1 e Aplicação 2, nos quais, mais uma vez, os suavizadores foram avaliados. Validou-se qual apresenta o melhor poder preditivo e qual representa de forma mais adequada os dados. Ressalta-se que para a Aplicação 1, a técnica que obteve o melhor poder preditvo e o melhor ajuste fora o *splines* de regressão cúbico. Para a Aplicação 2, o suavizador que se destacou por obter o melhor poder preditivo (menor EQM_{loocv} entre os suavizadores) foi *splines* de regressão linear. Em contrapartida, o que denotou melhor qualidade de ajuste (menor EQM_c dentre os suavizadores) foi o método *kernel* gaussiano.

Ademais, outras métricas para validação da qualidade de predição e adequabilidade dos modelos podem ser adotadas. Existem outras técnicas que podem ser adotadas para seleção dos parâmetros de suavização que não foram discutidas neste trabalho. Outrossim, especificamente para os suavizadores *splines* de regressão, além da seleção da quantidade de nós, a localização dos nós (que foram mantidas fixas e equidistantes nos k percentis possíveis) pode ser avaliada a

fim de obter um melhor ajuste. Aliás, todas discussões realizadas podem ser extendidas, quando mais de uma covariável está disponível para predizer a resposta. Frequentemente, utiliza-se o algoritmo de retroajuste (*backfitting*, HASTIE & TIBSHIRANI, 1990) para estimar cada função suave f_j em um cenário não paramétrico.

7 Referências

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). **Linear smoothers and additive models**. The Annals of Statistics, 17, 453-510.

CLEVELAND, W. S. (1979). **Robust locally weighted regression and smoothing scatterplots**. Journal of the American Statistical Association, 74, 829-836.

DELICADO, P., 2008 **Curso de Modelos no Paramétricos** p. 200.

EUBANK, R. L(1999) **Nonparametric Regression and Spline Smoothing**. Marcel Dekker, 2o edição. Citado na pág. 1, 2, 29

FAHRMEIR, L. & TUTZ, G. (2001) **Multivariate Statistical Modelling Based on Generalized Linear Models**. Springer, 2o edição. Citado na pág. 15

GREEN, P. J. & YANDELL, B. S. (1985) **Semi-parametric generalized linear models**. Lecture Notes in Statistics, 32:4455. Citado na pág. 15

GREEN P. J. & SILVERMAN B. W. (1994). **Nonparametric regression and generalized linear models: a roughness penalty approach**. Chapman & Hall, London.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). **Generalized additive models**, volume 43. Chapman and Hall, Ltd., London. ISBN 0-412-34390-8.

MONTGOMERY, D. C. & PECK, E. A. & VINING, G. G. **Introduction to Linear Regression Analysis**. 5th Edition. John Wiley & Sons, 2012.

IZBICK, R. & SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. ISBN 978-65-00-02410-4.

TEAM, R. CORE. R: **A language and environment for statistical computing**. (2013).