Análise de dados funcionais

Luben Miguel Cruz Cabezas
Vinicius Hideki Yamada Santiago

Universidade Federal de São Carlos

Introdução a dados funcionais

O que são dados funcionais?

- Consiste em dados que podem ser ordenados de acordo com alguma dimensão;
- A dimensão (contínua) pode ser o tempo, frequência, comprimento de onda, etc;
  - Geralmente representada por t
- Em vez de pensar nos dados como vetores, pensa-se neles como uma função (curva);
  - $x(t)$
- Daí o nome Dados funcionais $\implies$ Dados que são uma função;

Surgimento

- Decomposição de Karhunen-Loève: decomposição de processos estocásticos em componentes principais [Grenander, 1950].
- Análise de componentes principais funcionais [Kleffe, 1973];
- Ramsay and Silverman [2008] expandiram o termo Análise de Dados Funcionais (aplicações e teoria).

Figura 1: James O. Ramsay.

Exemplo de dados funcionais (Dados de CO_2)

Tabela 1: Variáveis do banco de dados e suas características.

	Plant	*conc*	*uptake*
1	Qn1	95.00	16.00
2	Qn1	175.00	30.40
3	Qn1	250.00	34.80
4	Qn1	350.00	37.20
5	Qn1	500.00	35.30
6	Qn1	675.00	39.20
7	Qn1	1000.00	39.70
8	Qn2	95.00	13.60
...			
41	Qc3	675.00	39.60
42	Qc3	1000.00	41.40
43	Mn1	95.00	10.60
44	Mn1	175.00	19.20
...			
84	Mc3	1000.00	19.90

Exemplo de dados funcionais (Dados de CO_2)

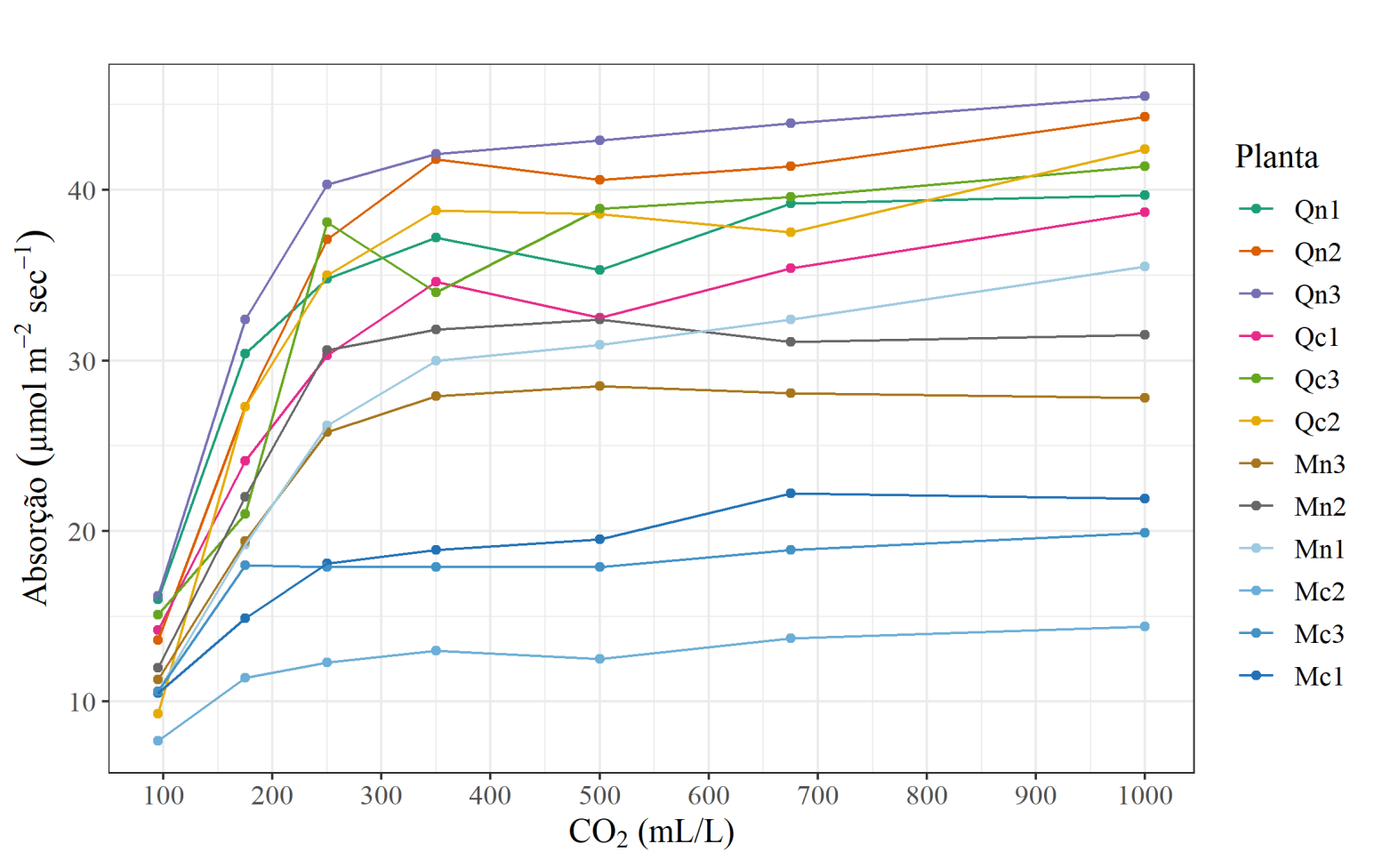

Figura 2: Absorção de CO_2 em relação em relação ao níveis de concentração de CO_2.

Vantagens de Dados Funcionais

- Além de trabalhar com estimações da curva, pode-se trabalhar com suas derivadas e integrais;
- O ajuste das curvas é não paramétrico, ou seja, não depende da especificação de alguma distribuição probabilística;
- Como o foco são simplesmente curvas, pode-se pensar em trabalhar com imagens, caracteres, curvas de nível, etc;

Exemplo de dados funcionais (Câncer de Mama)

Tabela 2: Variáveis da base sobre Câncer de Mama.

	Idade	Ano	TaxaCancer
1	47	1921	33.50
2	52	1921	59.10
3	57	1921	49.80
4	62	1921	55.80
5	67	1921	56.00
6	72	1921	50.00
7	77	1921	140.10
8	82	1921	116.50
9	87	1921	90.90
...			
41	67	1925	69.00
42	72	1925	90.00
43	77	1925	134.20
...			
729	87	2001	178.90

Exemplo de dados funcionais (Câncer de Mama)

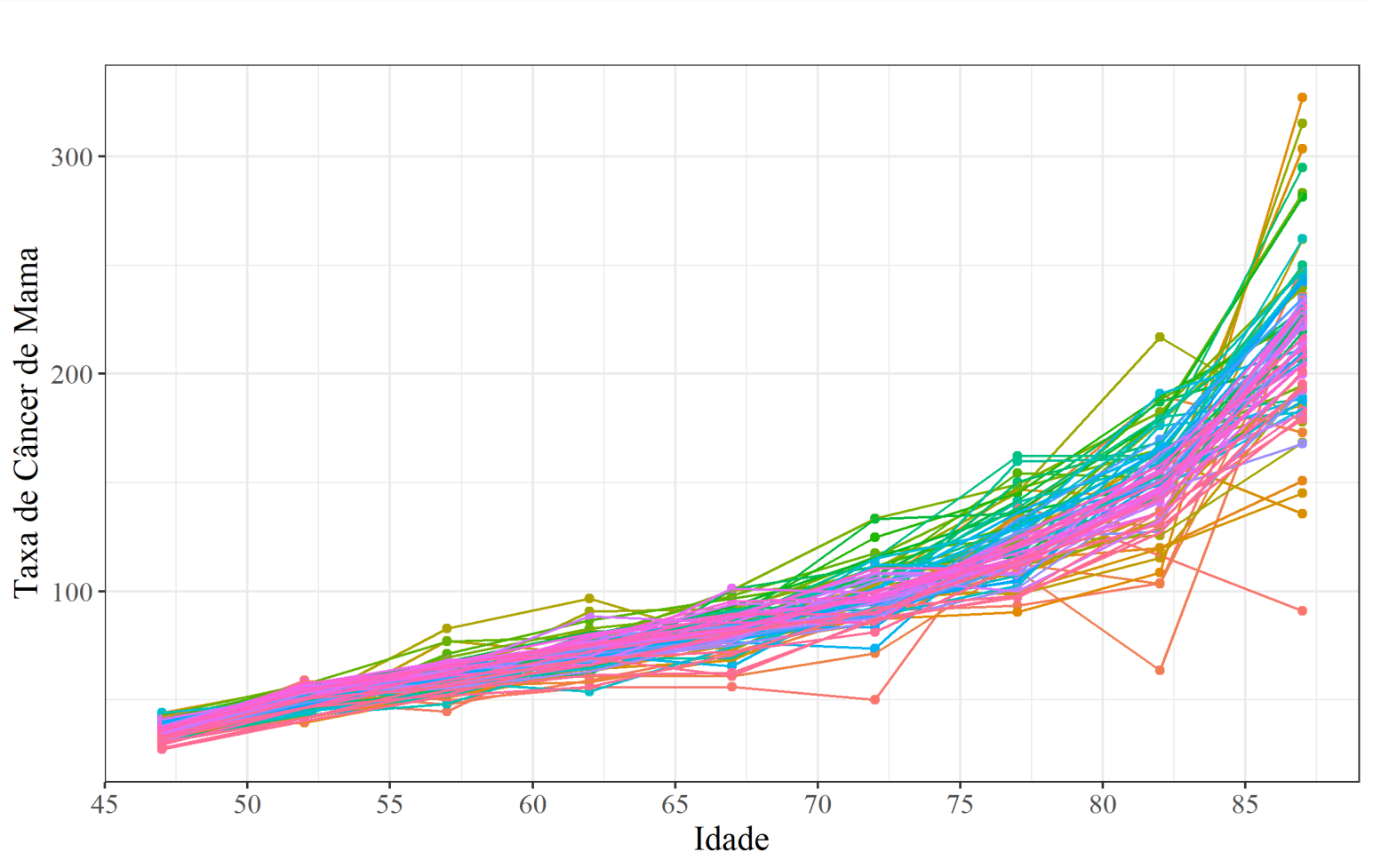

Figura 3: Taxa de Câncer de Mama em relação a Idade das mulheres.

Diferentes modelagens

Tabela 3: Modelagens.

Modelagem	Objeto	Suposição	Modelo	Predição
Séries temporais	$X_1, ..., X_n$	Observações igualmente espaçadas no tempo	AR(1):$X_t = _{t-1} + \omega_t$ ou ARMA, MA, ARIMA, SARIMA	Predizer X_t com X_{t-1}
Dados Longitudinais	$(\boldsymbol{Y}, \boldsymbol{X})$	Observações igualmente espaçadas ou não no tempo $Y_i \sim Distribuição$	$Y_{ij} = \boldsymbol{X}\beta + \boldsymbol{Z}b$	Predizer Y_{ij} com X_{ij}
Dados Funcionais	$(\boldsymbol{Y}, \boldsymbol{X(t)})$	Observações igualmente espaçadas ou não no tempo Y_i modelado de forma não paramétrica	$Y_{ij} = \sum_{j=1}^{K} c_j \phi_j(t)$	Predizer Y com $X(t)$

Visualizações

- $t \;\; VS \;\; x_1(t), x_2(t), ..., x_n(t)$;
- $t \;\; VS \;\; x_1'(t), x_2'(t), ..., x_n'(t)$;
- Phase Plane plot $(x'(t) \;\; VS \;\; x''(t))$;
- Outras visualizações [Ramsay and Silverman, 2008];

Phase Plane Plot (dado funcional)

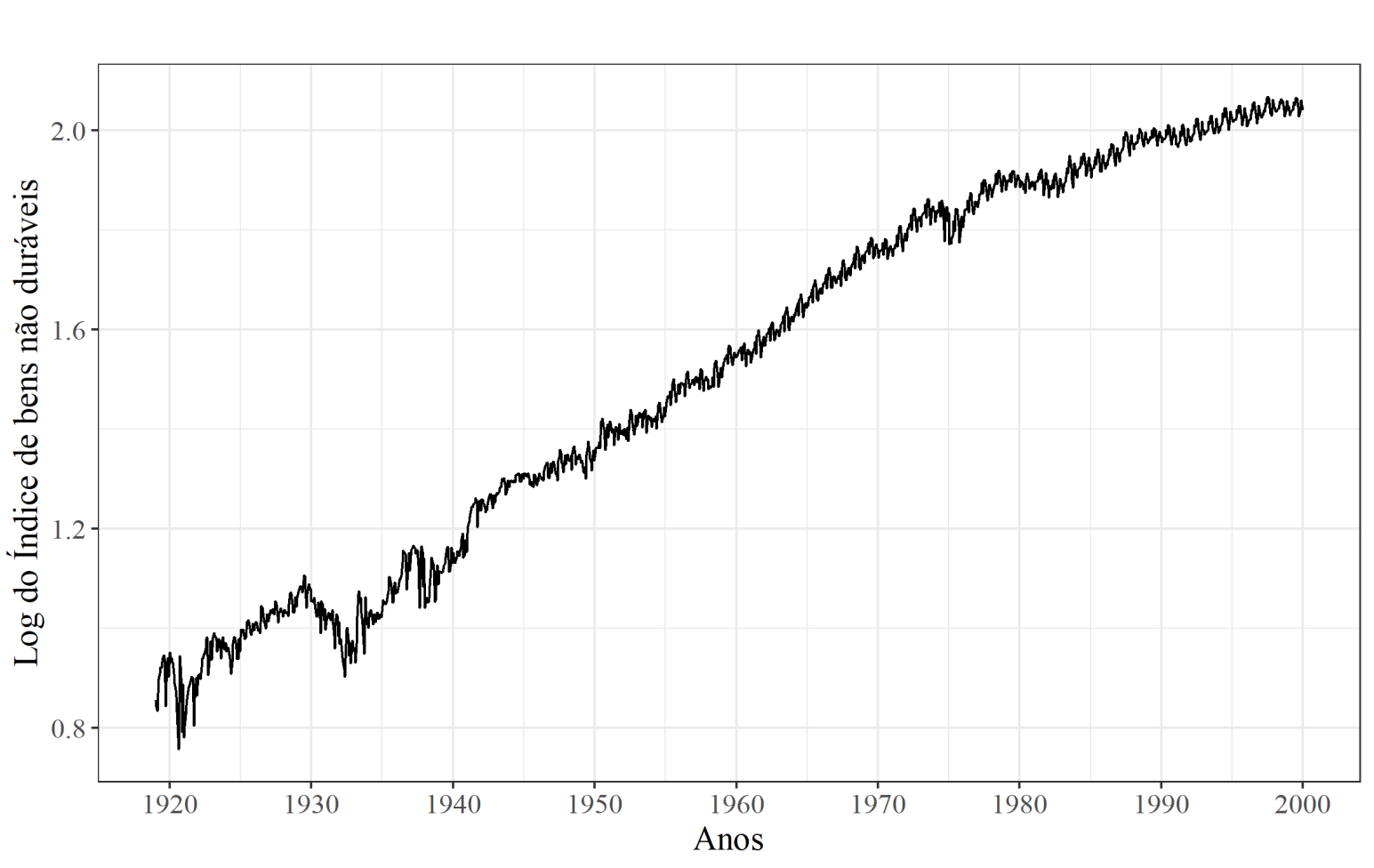

Figura 4: Log do Índice de produtos não duráveis dos EUA.

Phase Plane Plot (após o ajuste da função)

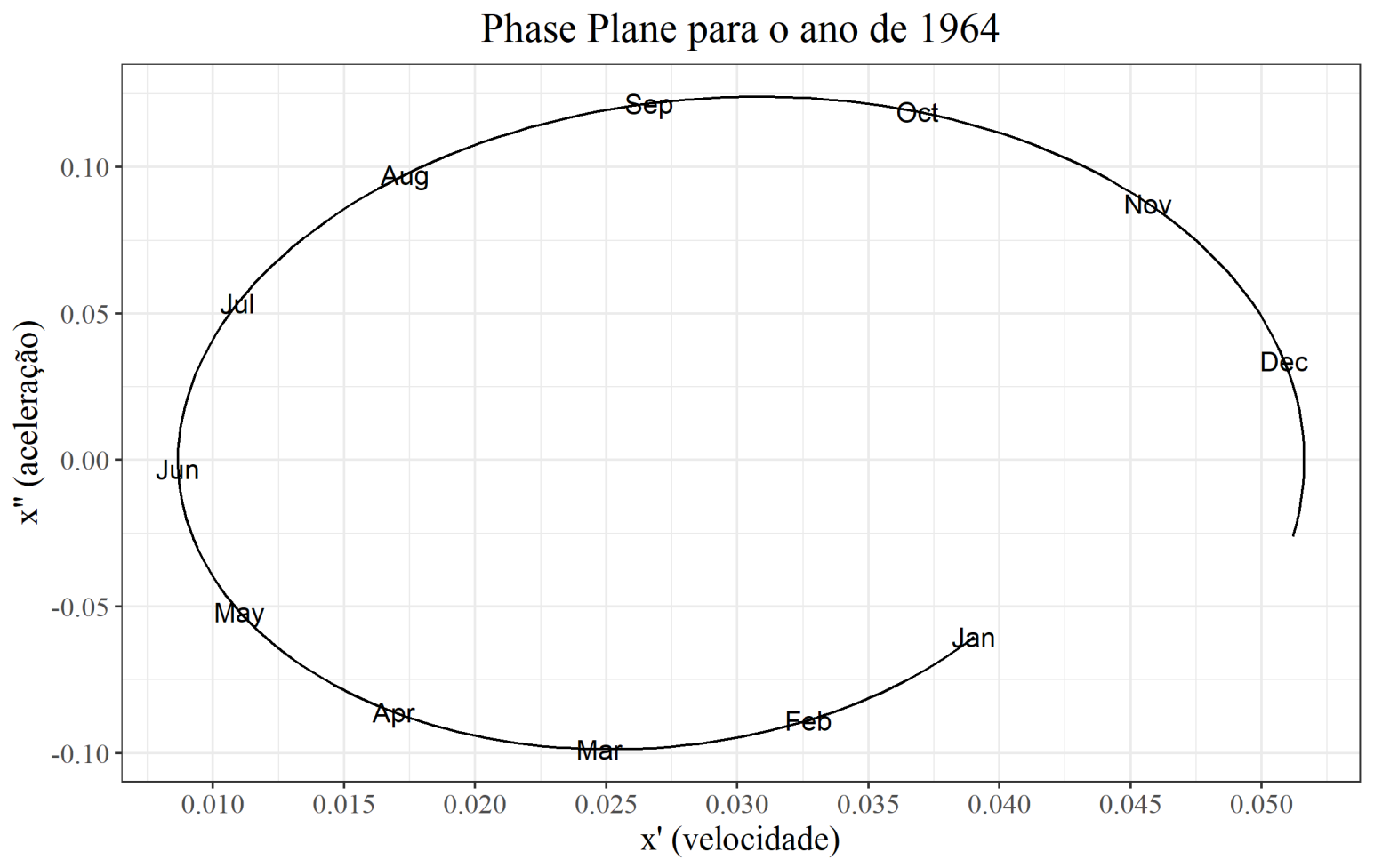

Figura 5: Phase Plane Plot para o ano de 1964 do Log do Índice de produtos não duráveis dos EUA (função ajustada por uma b-spline).

Phase Plane Plot (após o ajuste da função)

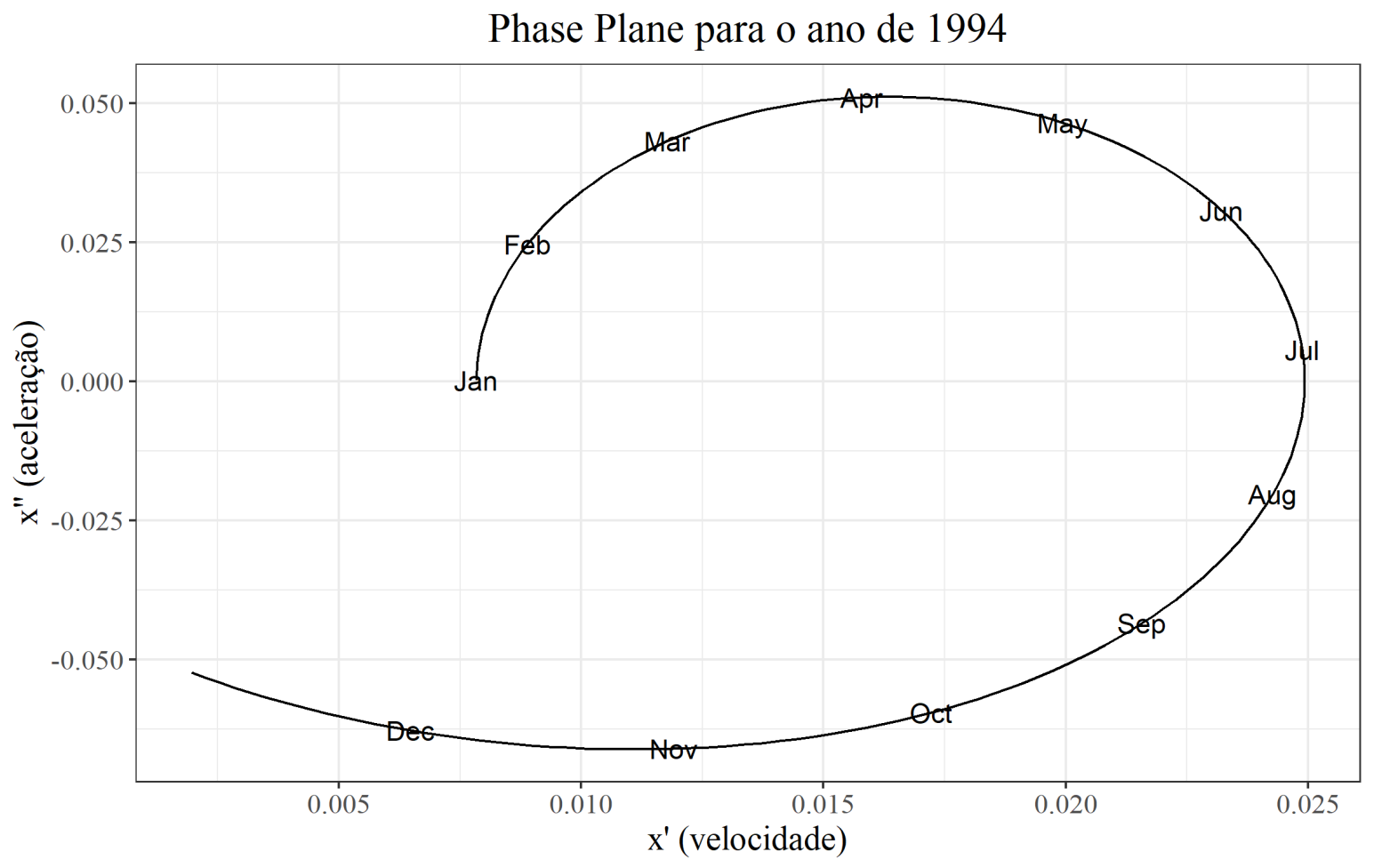

Figura 6: Phase Plane Plot para o ano de 1994 do Log do Índice de produtos não duráveis dos EUA (função ajustada por uma b-spline).

Análises descritivas em dados funcionais

Estatísticas básicas

Média em Dados Funcionais

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t).$$

Variância em Dados Funcionais

$$Var_x(t) = \frac{1}{n}\sum_{i=1}^{n}(x_i(t) - \bar{x}(t))^2.$$

Estatísticas básicas

Covariância em Dados Funcionais

$$Cov_x(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n}(x_i(t_1) - \bar{x}(t_1))(x_i(t_2) - \bar{x}(t_2)).$$

Correlação em Dados Funcionais

$$Corr_x(t_1, t_2) = \frac{Cov_x(t_1, t_2)}{\sqrt{Var_x(t_1)Var_x(t_2)}}$$

Estatísticas básicas

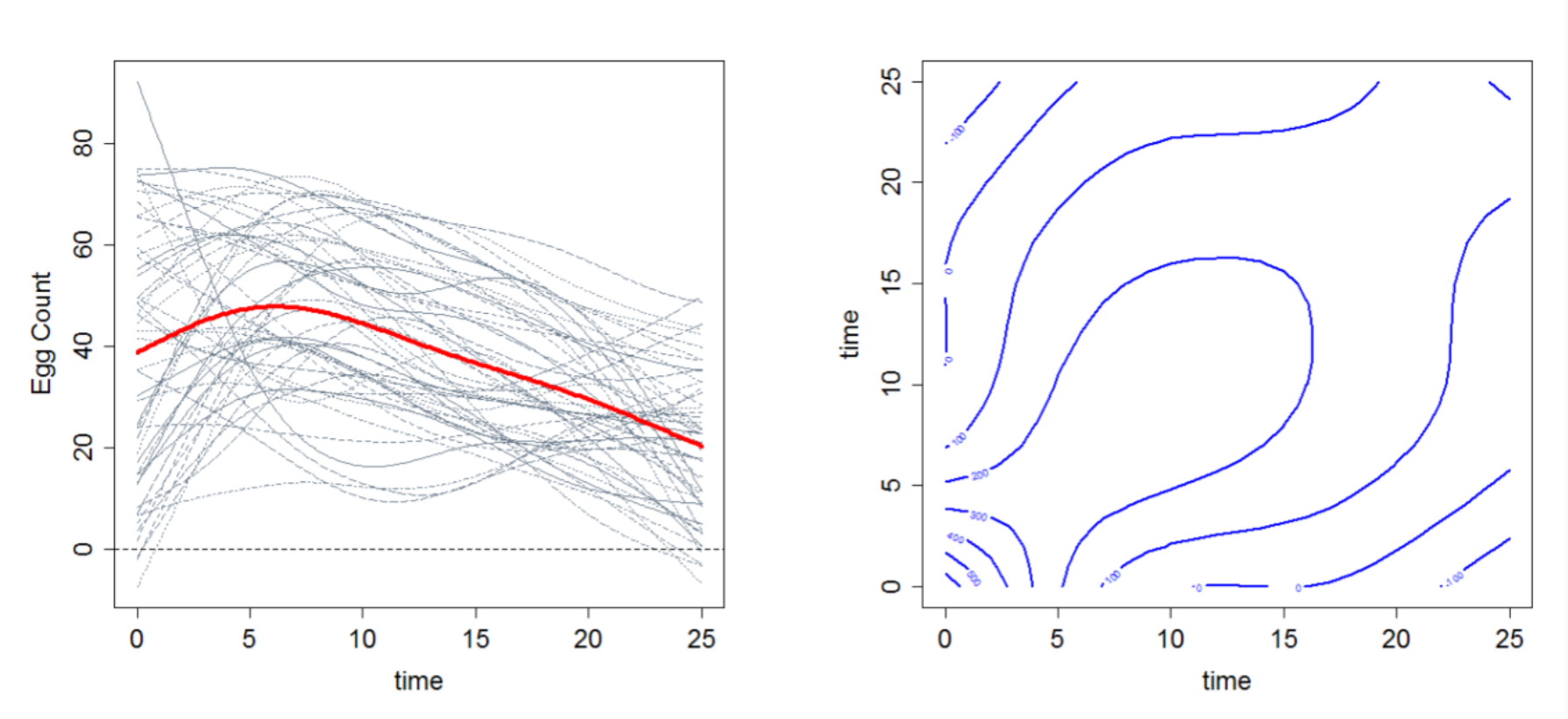

Figura 7: Média e Covariância de um dado funcional $x(.)$ [Ramsay et al., 2009].

Bases

Representação funcional dos dados

- Dados que têm uma estrutura funcional, vêm em um formato tabular discreto, com n pares (t_j, y_j);
- Interesse em obter a função suave $x(t)$ que gera esses dados tal que, $y_j = x(t_j) + \epsilon_j$;
- Considera-se geralmente que ϵ_j são independentes, com média 0 e variância σ^2 e $x(t_j)$ são fixos, tendo então $\mathbb{V}[\boldsymbol{Y}] = \Sigma_e = \sigma^2 \boldsymbol{I}$.

Representação de funções através de bases

- Para se obter uma função suave $x(t)$ representando nossos dados funcionais pode-se utilizar da combinação linear de **bases**;
- Isso vem do fato de que qualquer função $x(t)$ pode ser aproximada pela combinação linear de um conjunto com K funções linearmente independente entre si ϕ_k;
- $x(t) = \sum_{j=1}^{K} c_j \phi_j(t) = \boldsymbol{c}^t \boldsymbol{\phi}$;
- A representação exata de $x(t)$ é dada quando $K = n$;
- Problemas: Como escolher as bases ϕ? Como achar um bom K?
- Principais bases: **Fourier**, **Splines** (B-splines, splines naturais, etc.).

Base de fourier

- Expansão de Fourier:
  $\hat{x}(t) = c_0 + c_1 \text{sen}\omega t + c_2 \cos\omega t + c_3 \text{sen} 2\omega t + c_4 \cos 2\omega t + \ldots;$
- Base periódica: $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \text{sen} rwt$, $\phi_{2r}(t) = \cos rwt$;
- ω determina o período $2\pi/\omega$;
- Transformada rápida de fourier torna possível encontrar os coeficientes c_k de forma muito eficiente;
- Assim, a base é boa de se utilizar em dados periódicos ou séries de duração muito longa.
- Derivadas são facilmente calculaveis:
  $\boldsymbol{D}_x = (0, c_1, -\omega c_2, 2\omega c_3, -2\omega c_4, \ldots)$ e
  $\boldsymbol{D}_x^2 = (0, -\omega^2 c_1, -\omega^2 c_2, -4\omega^2 c_3, -4\omega^2 c_4, \ldots)$
- "Séries de Fourier são como margarina".

Base de fourier

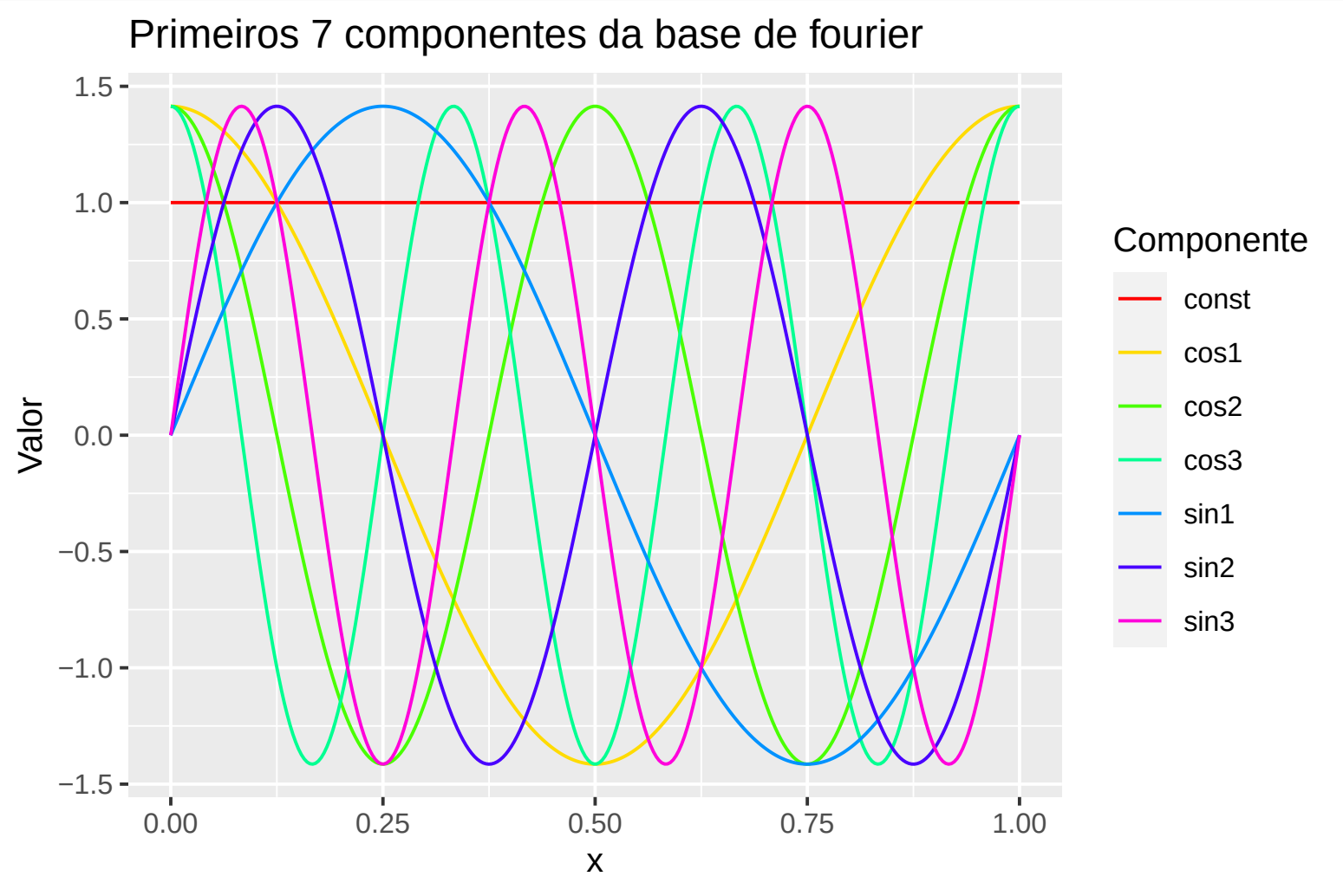

Figura 8: Base de Fourier de periodo 1 com $K = 7$

Splines

- Escolha mais comum para aproximar funções não periódicas
- Idéia principal: Dividir o intervalo de interesse em L subintervalos separado por nós $\tau_l, l = \{1, \ldots, L-1\}$ e adicionar polinômios distintos de ordem m para cada intervalo.
- Ou seja, a spline é um conjunto de polinômios de ordem m por partes
- Ressalta-se que os polinômios de cada intervalo se juntam de forma suave, tendo valores iguais para os nós que os separam.
- São usados $m + L - 1$ parâmetros para determinar uma spline, sendo m a ordem dos polinômios em partes e L o número de subintervalos separados por um sequência de nós τ.

Splines

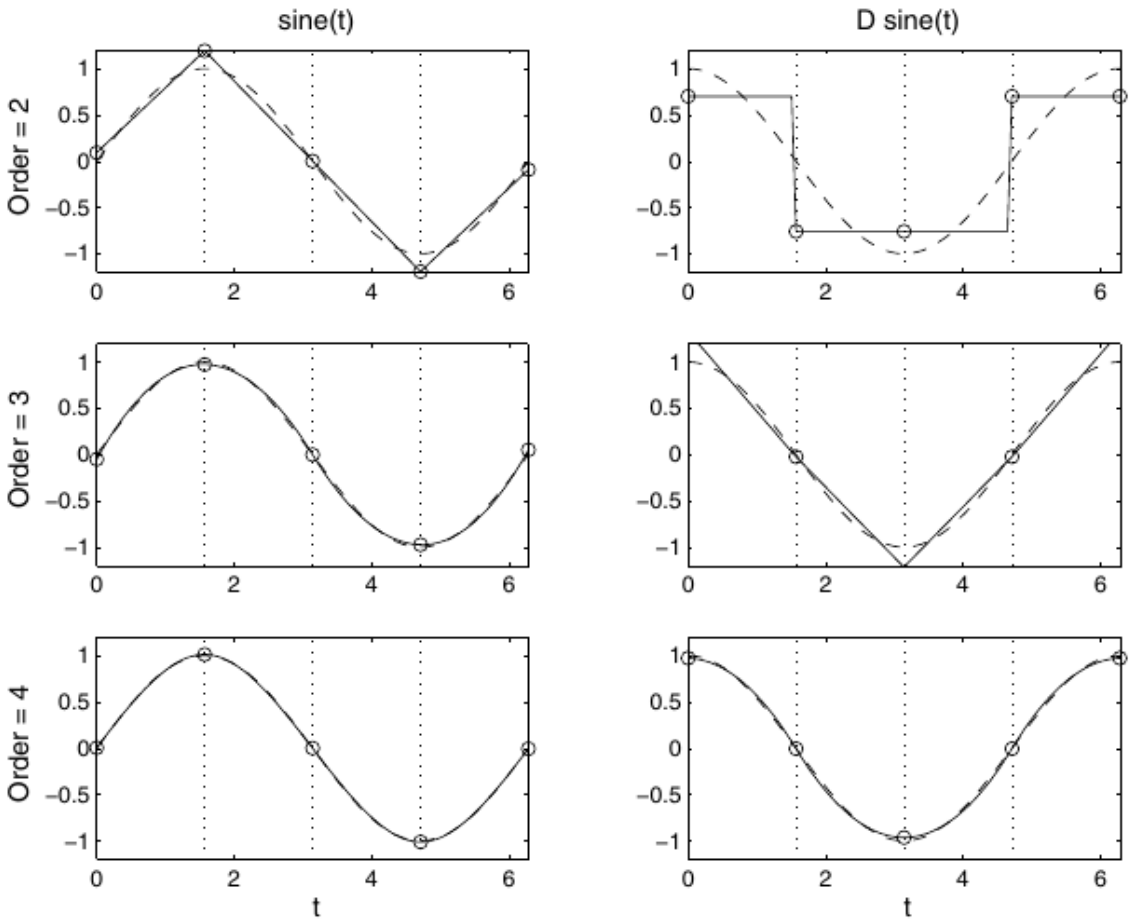

Figura 9: Exemplo de splines interpoladas à função seno (à esquerda) e sua derivada (à direita), retirado de Ramsay and Silverman [2008]

B-splines

- Uma maneira de especificar um conjunto de funções de base $\phi_k(t)$ que construam uma spline é a partir das seguintes propriedades:
  - Cada base $\phi_k(t)$ é uma spline de ordem m e sequência de nós τ;
  - Qualquer combinação linear dessas bases resulta em uma função spline;
  - Qualquer spline definida por m e τ pode ser expressada pela combinação linear dessas bases.
- Há varias formas de construir essa base, mas o mais popular são as B-splines, desenvolvidas por Boor [2001].
- Propriedade interessante: cada função B-spline de ordem m é positiva sobre apenas m intervalos adjacentes.
- Isso garante certa esparsidade no produto interno da matriz Φ, com as bases sendo parcialmente ortogonais.

B-splines

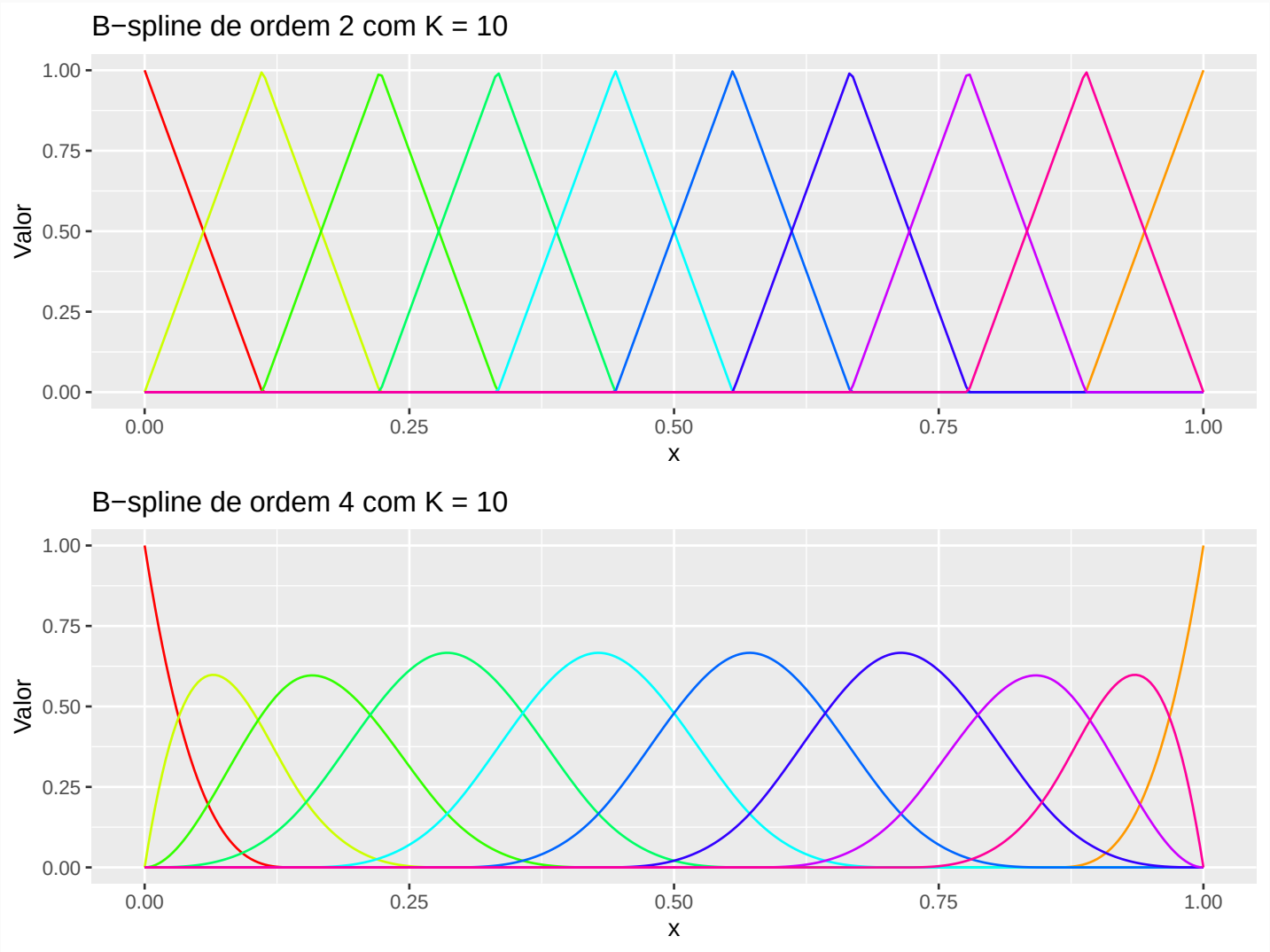

Figura 10: B-splines de ordem diferentes com $K = 10$.

Outras bases

- Há várias outras bases comumente utilizadas não só para aproximar $x(t)$ mas também para modelos de equação diferencias dos dados funcionais, e para análise de componentes principais em dados funcionais;
- Ondaletas: $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$
- Base exponencial: $e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_k t}, \ldots;$
- Base de monômios: $\phi_k(t) = (t - \omega)^k, k = 0, \ldots, K.$

Suavizando dados funcionais pelos mínimos quadrados

Mínimos quadrados ordinários

- Como visto anteriormente, se usarmos a expansão por bases, escrevemos $x(t) = \boldsymbol{c}^t \phi$
- Definimos assim a matriz de covariáveis $\boldsymbol{\Phi}$ com dimensão $n \times K$ com entradas $\phi_k(t_j)$
- Minimizar
  $SQT(\boldsymbol{y}|\boldsymbol{c}) = \sum_{i=1}^{n}[y_i - \sum_{j=1}^{K} c_j \phi_j(t_i)]^2 = (\boldsymbol{y} - \boldsymbol{\Phi c})^t(\boldsymbol{y} - \boldsymbol{\Phi c})$
- Estimador de mínimos quadrados: $\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^t\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^t\boldsymbol{y}$
- Mínimos quadrados ponderados:
  - Minimizar $SQT(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})^t\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\Phi c})$
  - Estimador de mínimos quadrados ponderados:
    $\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^t\boldsymbol{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^t\boldsymbol{W}\boldsymbol{y}$

Exemplo: Precipitação em Vancouver

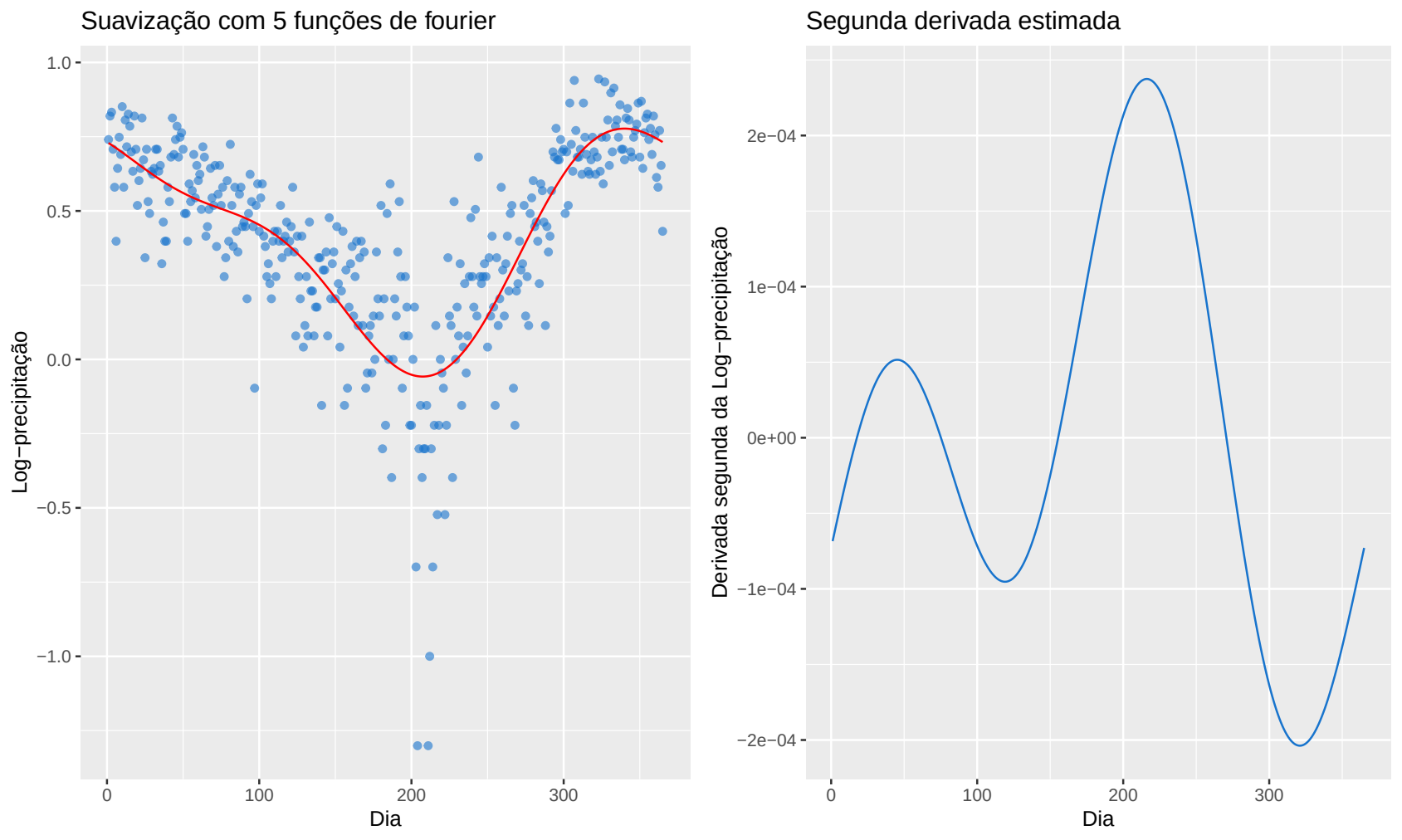

Figura 11: Suavização da precipitação em Vancouver usando mínimos quadrados com uma base de fourier.

Exemplo: Precipitação em Vancouver

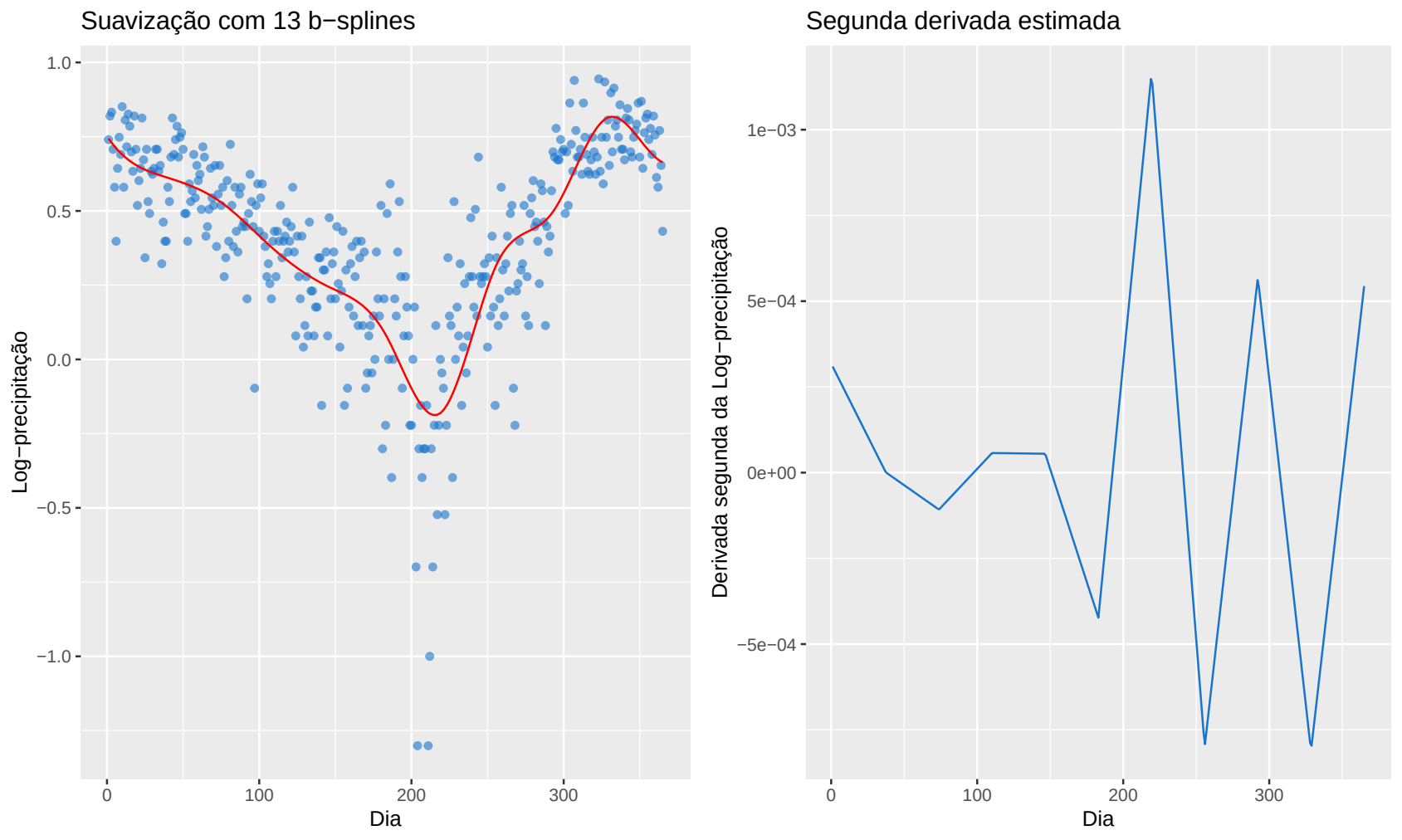

Figura 12: Suavização da precipitação em Vancouver usando mínimos quadrados com uma base de b-splines.

Exemplo: Precipitação em mais outras estações

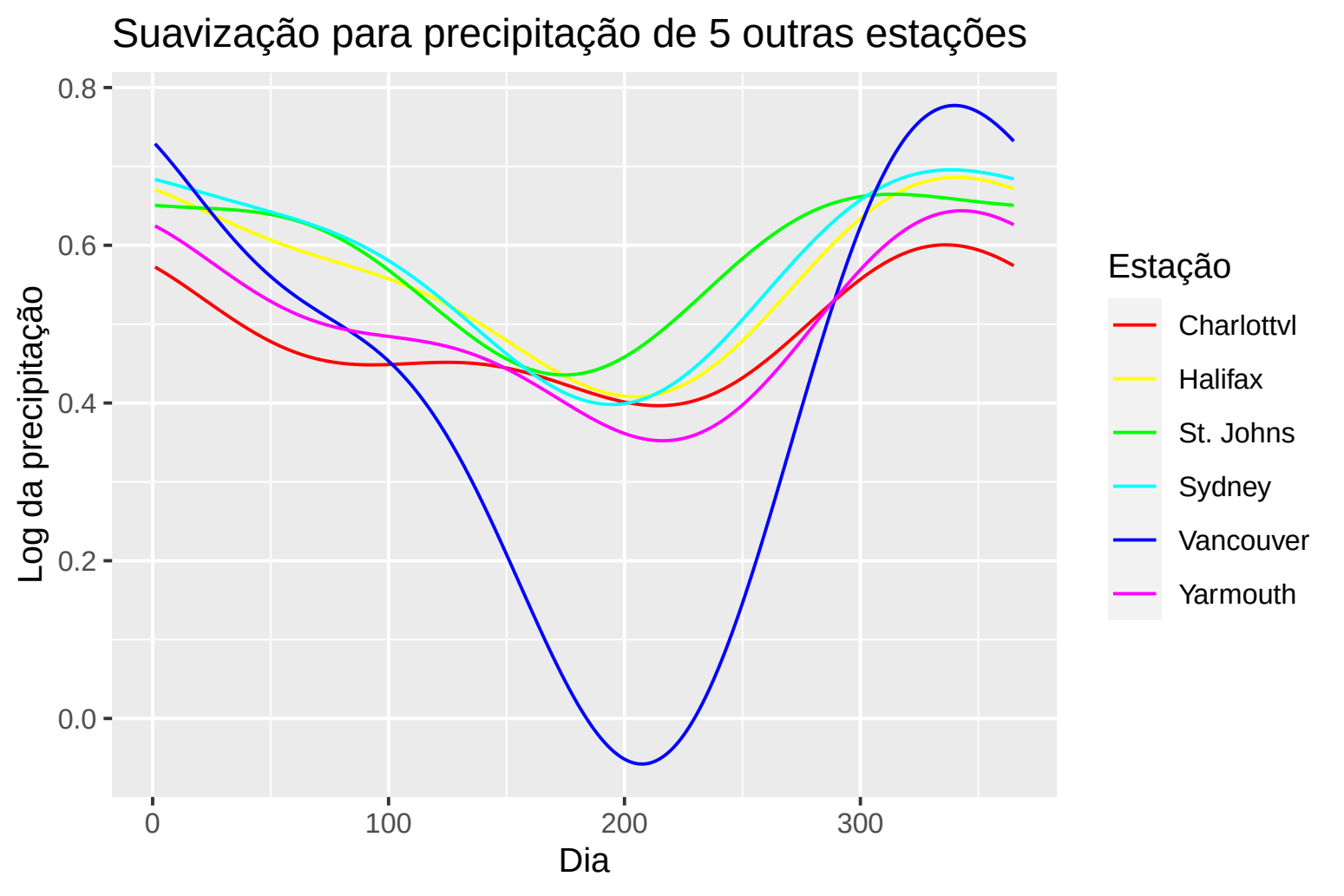

Figura 13: Curvas estimadas de precipitações para outras estações.

Escolher K para as funções de base

- Uma dúvida já levantada antes era como escolher a ordem da base K;
- Quanto maior K, melhor tende a ser o ajuste da suavização, mas pode levar a sobreajustes da curva.
- Já, quanto menor K, tendemos mais a um subajuste da curva.
- Para determinar K, analisa-se o balanço entre o viés e variância da suavização
- O erro quadratico médio é decomposto em função do viés e variância da seguinte maneira: $EQM(\hat{x}(t)) = \text{Vies}(\hat{x}(t))^2 + \mathbb{V}[\hat{x}(t)]$

Escolher K para as funções de base

- Validação cruzada *leave-one out* para escolher K
  - Remover um par (t_i, y_i);
  - Ajustar a curva nos dados restantes $\hat{x}_{-i}(t_i)$ para vários K
  - Escolher o K que minimiza $VC(\boldsymbol{y}) = \sum_{i=1}^{n}(y_i - \hat{x}_{-i}(t_i))^2$
- Outra maneira de escolher o K é realizar seleção de variáveis, como o *stepwise*, em uma base com K grande;

Exemplo utilizando validação cruzada

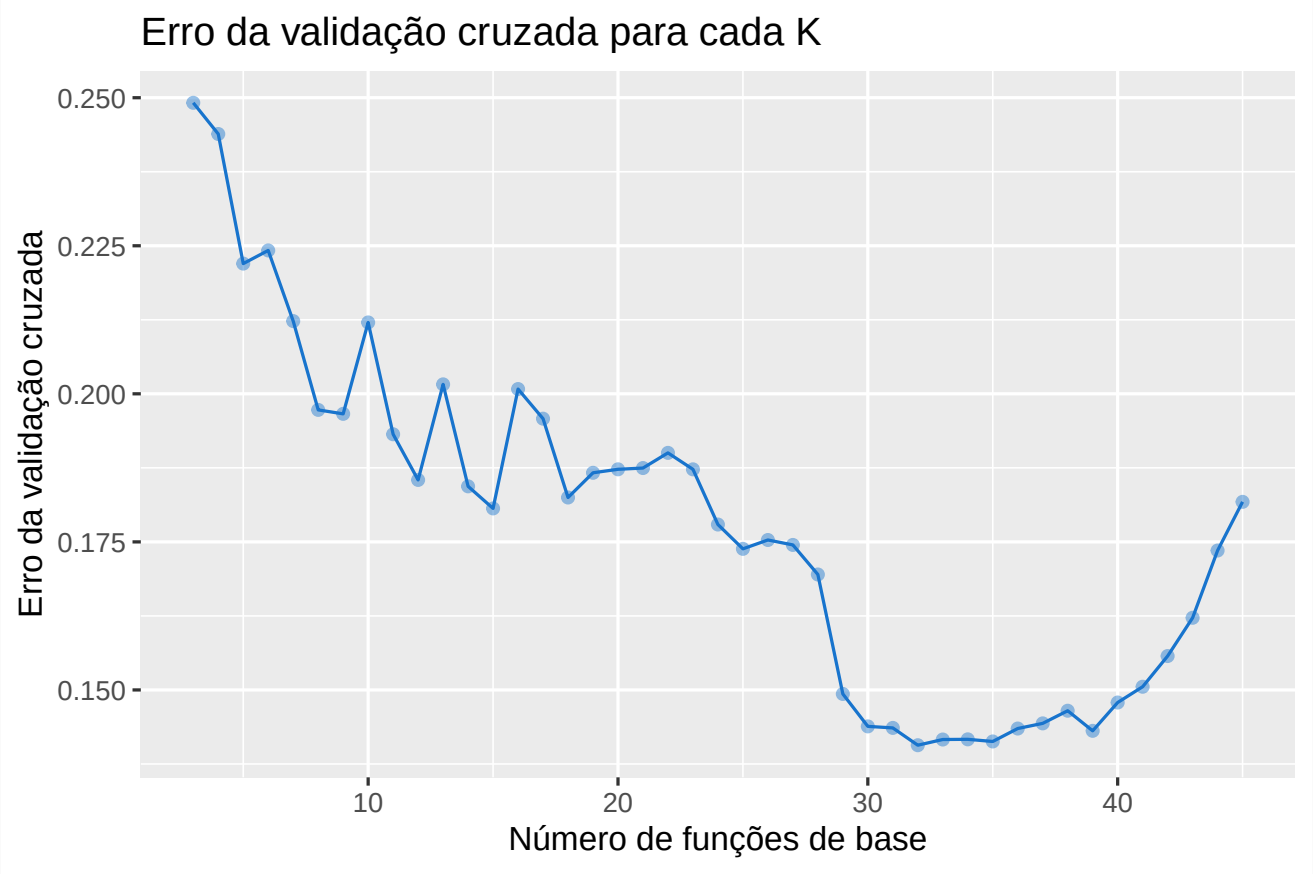

Figura 14: Validação cruzada para escolher K nos dados de precipitação em Vancouver

Suavizando dados funcionais por mínimos quadrados locais

Mínimos quadrados locais

- Intuição: a estimação da função no ponto t deve ser principalmente influenciada por observações próximas ao ponto t;
- Essa propriedade já existe de forma implícita no estimador de mínimos quadrados, e é explicita em estimadores de pesos locais;
- A idéia desses tipos de estimadores é estimar a função $x(t)$ para cada t_j como uma ponderação dos valores y_j: $x(t_j) = \sum_{i=1}^{n} w_i y_i$
- As observações mais próximas a t_j teriam um peso maior que observações mais distantes a esse.

Mínimos quadrados locais

- Podemos obter os pesos locais w_i através de kernels centralizados em t_j:
  - Uniforme: $K(u) = 0.5\ \mathbb{I}(|u| \leq 1)$
  - Quadrático: $K(u) = 0.75(1 - u^2)\ \mathbb{I}(|u| \leq 1)$
  - Gaussiano: $K(u) = (2\pi)^{-1/2} \exp\left(-u^2/2\right)$
- Mais especificamente, fixado um t_j, teremos: $w_i = K\left(\frac{t_i - t_j}{h}\right)$;
- Valores grandes de w_i estão atrelados a proximidade de t_i nas vizinhanças de t_j;
- O parâmetro h é o parâmetro de largura de banda que controla o grau de concentração da vizinhança de t_j:
  - Valores grandes implicam em pesos parecidos entre valores distantes ou próximos a t_j
  - Valores pequenos implicam em pesos maiores apenas para valores mais próximos a t_j

Suavização por kernel

- Estimador mais comum que usa pesos locais: estimador por kernel
- $\hat{x}(t) = \sum_{i=1}^{n} S_i(t) y_i$
- $S_i(t)$: Função de pesos baseada em kernel
- Estimador de Nadaraya-Watson: $S_i(t) = \frac{K((t_i - t)/h)}{\sum_{j=1}^{n} K((t_j - t)/h)}$
- Pesos normalizados de forma que $\sum_{i=1}^{n} S_i(t) = 1$

Exemplo: temperatura diária de Vancouver

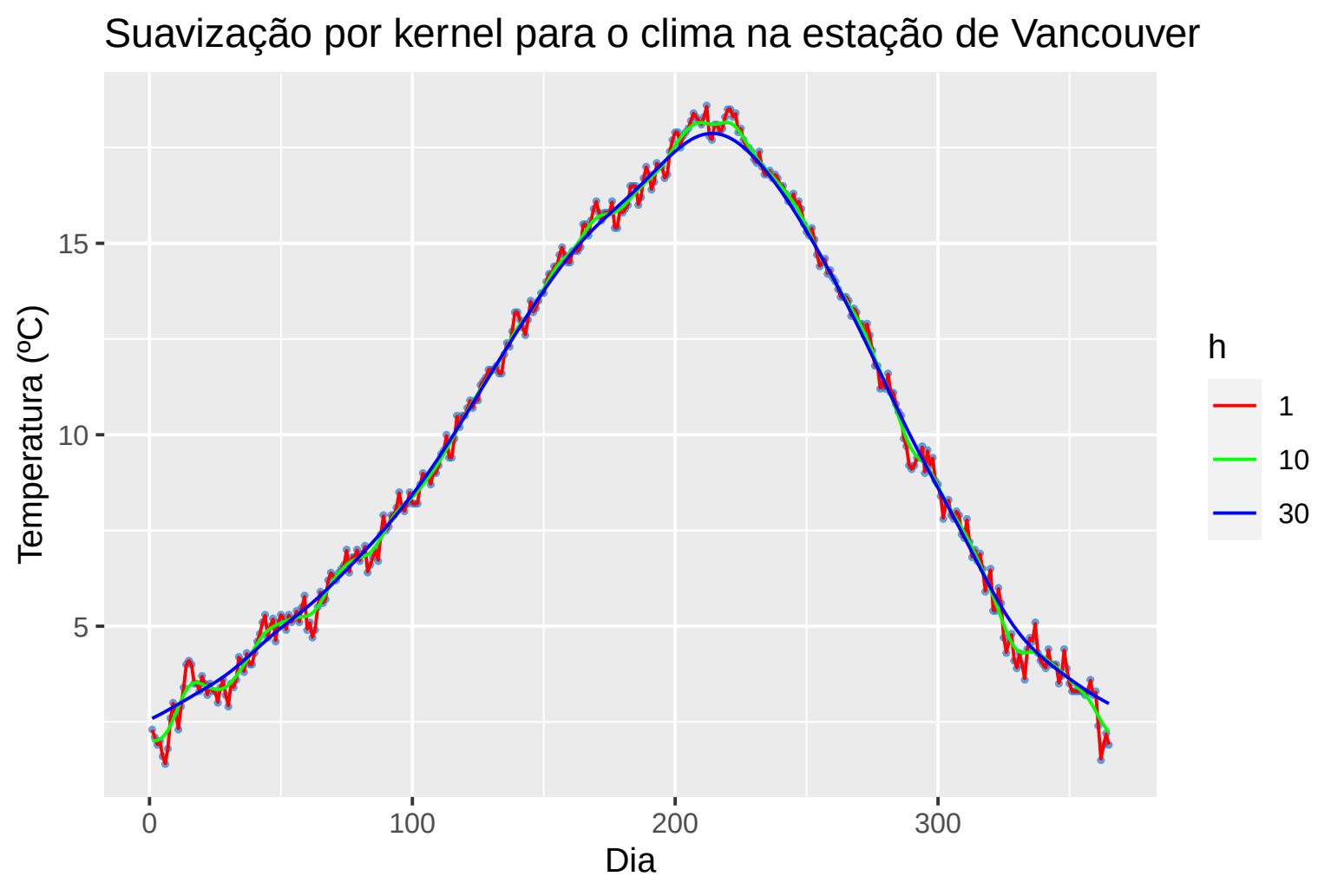

Figura 15: Suavização da temperatura diária em Vancouver para 3 diferentes larguras de bandas

Temperatura diária para outras estações

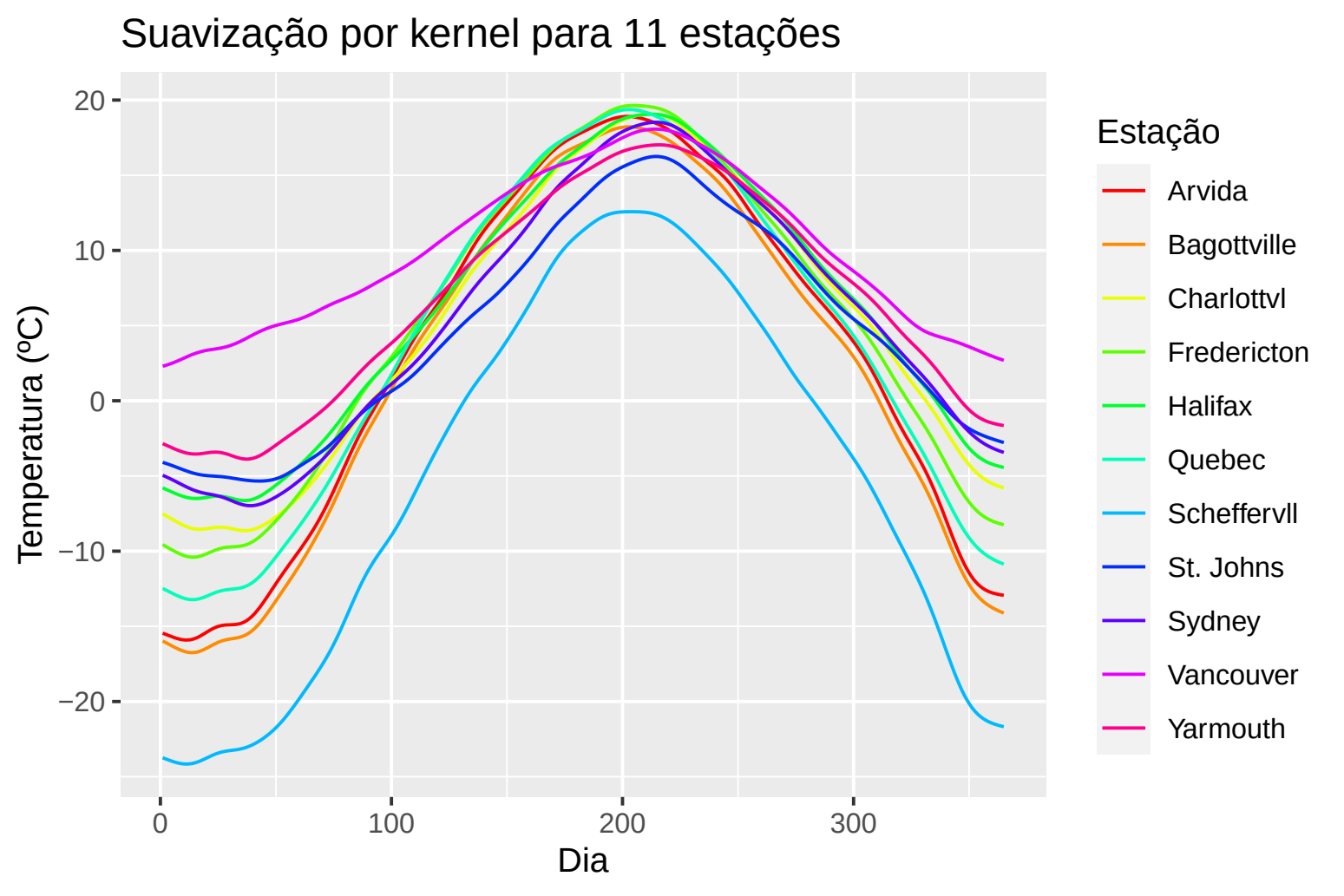

Figura 16: Suavização da temperatura diária para 11 diferentes estações tomando $h = 20$

Escolher h para os estimadores por kernel

- Como podemos chegar a melhor largura de banda h?
- Problema de sobreajuste e subajuste similar ao visto para o parâmetro K das funções de base;
- Similarmente, podemos usar a validação cruzada *leave-one-out* para escolher h ótimo;
- Há várias outras técnicas de escolha automática de h, geralmente no espirito da validação cruzada, mas não se pode confiar cegamente em nenhuma delas;
- Ramsay and Silverman [2008] sugerem testar uma variedade de valores h e por análise gráfica escolher um h interessante

Exemplo utilizando validação cruzada

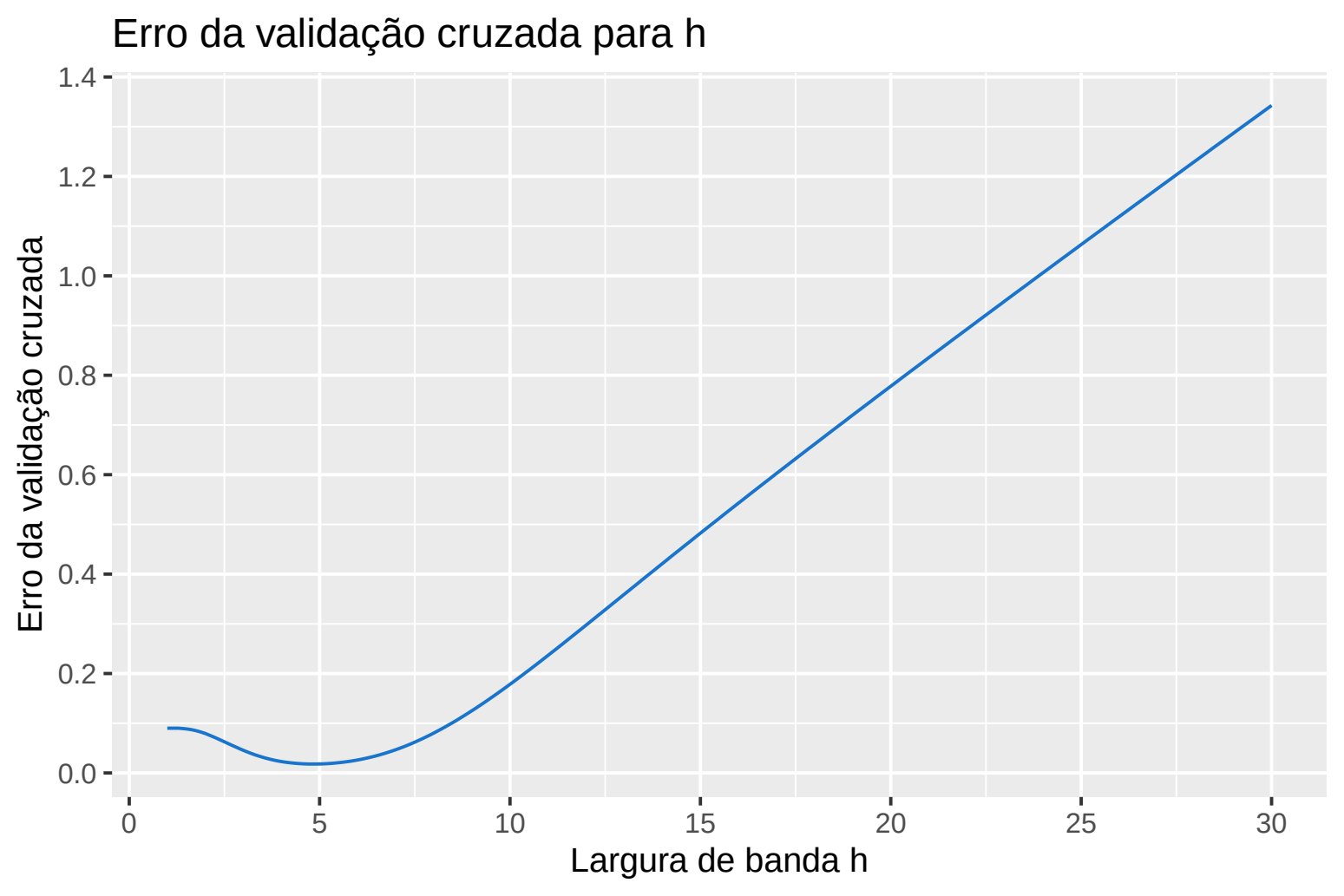

Figura 17: Validação cruzada para encontrar *h* ótimo

Mínimos quadrados locais com funções de bases

- Idéia: combinar o estimador local baseado em kernel com o estimador baseado em funções de bases.
- Isso é feito ao estender os mínimos quadrados ordinários para uma fórmula de erro local:

$$SQT_t(\boldsymbol{y}|\boldsymbol{c}) = \sum_{i=1}^{n} w_i(t) \left[y_i - \sum_{j=1}^{K} c_j \phi_j(t_i) \right]^2 = (\boldsymbol{y} - \boldsymbol{\Phi c})^t \boldsymbol{W}(t)(\boldsymbol{y} - \boldsymbol{\Phi c})$$

- Minimizando SQT_t, obtemos o estimador dos coeficientes locais: $\hat{c}(t) = [\boldsymbol{\Phi}^t \boldsymbol{W}(t) \Phi]^{-1} \boldsymbol{\Phi}^t \boldsymbol{W}(t) \boldsymbol{y}$
- O estimador por kernel de Nadaraya-Watson é um caso especial dos minimos quadrados locais com funções de bases, tomando $K = 1$ e $\phi_j(t) = 1$.

Caso particular: suavização polinomial local

- Tomar como base as funções de base de monômios
- Minimizar em particular:

$$SQT_t(\boldsymbol{y}|\boldsymbol{c}) = \sum_{i=1}^{n} K_h(t_i, t) \left[y_i - \sum_{j=0}^{L} c_j (t - t_i)^k \right]^2$$

- Dessa maneira, conseguimos utilizar a função de base polinomial para estimar mais facilmente as derivadas.

Exemplo: temperatura diária de Vancouver

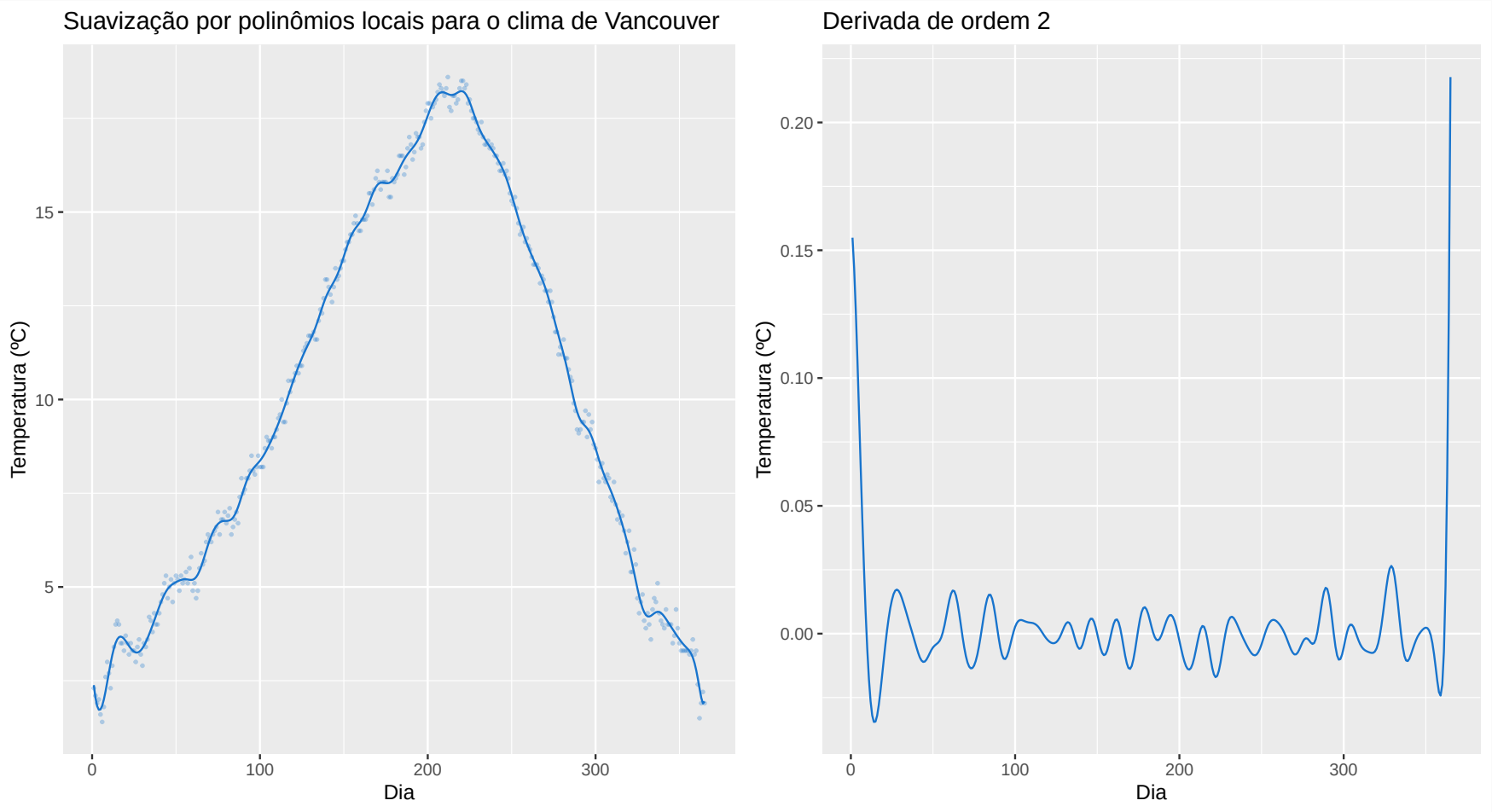

Figura 18: Ajuste por polinômio local com $h = 6$ e segunda derivada estimada

Suavização por penalização

Suavização com penalização

- Erro Quadrático Médio = Viés^2 + Variância Da Amostra
- Intuição
  - Pode-se aumentar um pouco o Viés para reduzir a variância;
  - Predição varia "gentilmente" de um valor ao outro;
  - Utiliza-se a suavidade entre os vizinhos (observações) mais próximos;
- Pode-se expressar uma penalização como

$$PEN_m(x) = \int [D^m x(s)]^2 ds$$

- Sabe-se que $x(t) = \sum_k^K c_k \phi_k(t) = \boldsymbol{c}'\boldsymbol{\phi}(t)$
  - $\boldsymbol{\phi}$ é vetor de tamanho k das funções de base;
  - $\boldsymbol{c}$ é vetor de tamanho k dos coeficientes;

Suavização com penalização

- Pode-se expressar uma penalização como

$$\begin{aligned} PEN_m(x) &= \int [D^m x(s)]^2 \, ds = \int [D^m \boldsymbol{c}'\boldsymbol{\phi}(s)]^2 \, ds = \\ &= \int \boldsymbol{c}' D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}' \boldsymbol{c} \, ds = \\ &= \boldsymbol{c}' \left(\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}' \, ds \right) \boldsymbol{c} = \\ &= \boldsymbol{c}' \boldsymbol{R} \boldsymbol{c}. \end{aligned}$$

- Somando a penalização $PEN_m(x)$ e soma de quadrados total $SQT_t(\boldsymbol{y}|\boldsymbol{c})$, tem-se

$$(\boldsymbol{y} - \boldsymbol{\Phi c})^t \boldsymbol{W}(t)(\boldsymbol{y} - \boldsymbol{\Phi c}) + \lambda \boldsymbol{c}' \boldsymbol{R} \boldsymbol{c}.$$

- Derivando em relação a $\boldsymbol{c}$ e igualando a 0, conclui-se que

$$\hat{\boldsymbol{c}} = [\boldsymbol{\Phi}^t \boldsymbol{W}(t) \Phi + \lambda \boldsymbol{R}]^{-1} \boldsymbol{\Phi}^t \boldsymbol{W}(t) \boldsymbol{y}$$

Suavização com penalização (4 métodos)

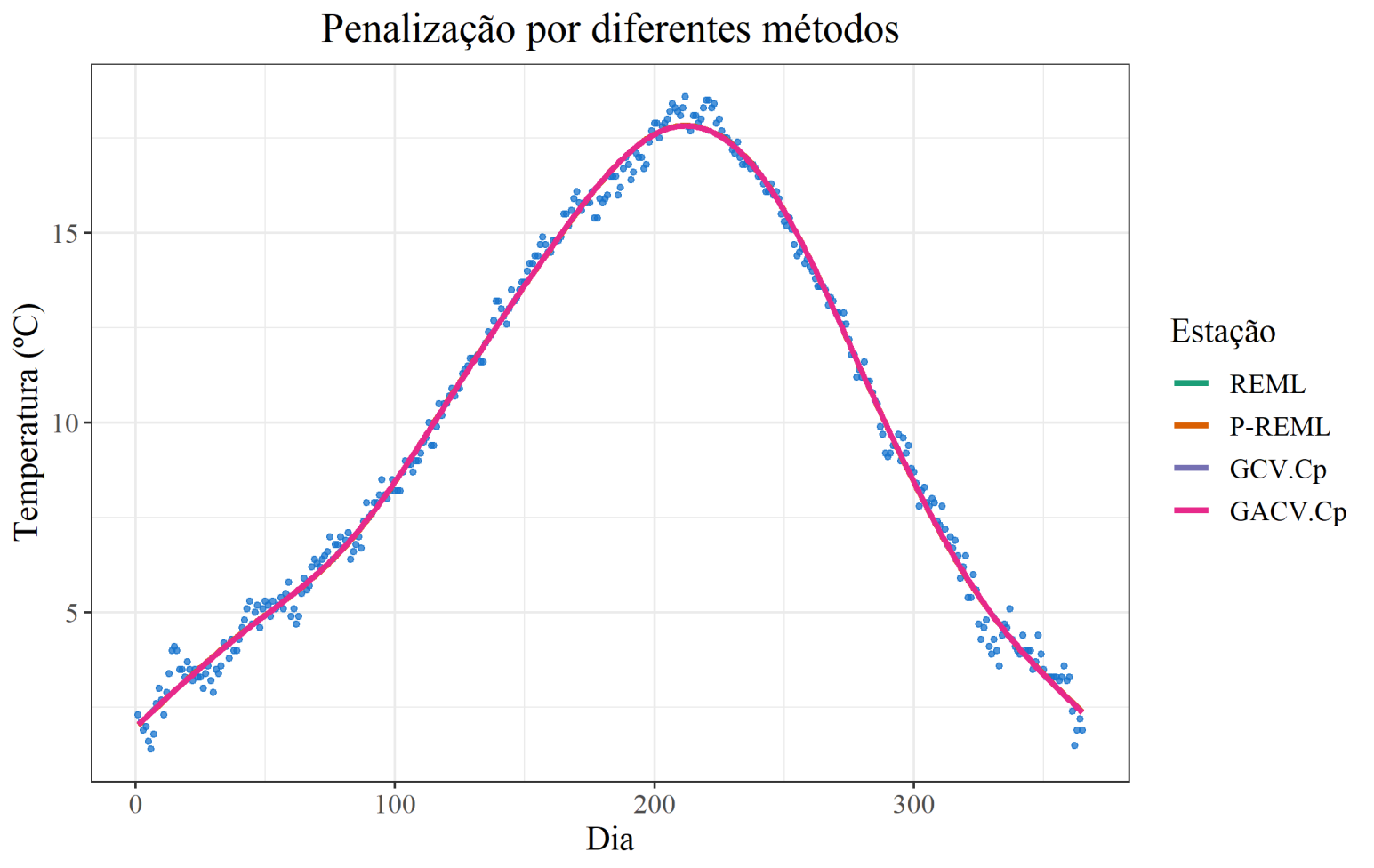

Figura 19: Suavizações para a estação de Vancouver com 4 métodos de estimação.

Suavização com penalização (4 métodos)

Tabela 4: Valores preditos das Suavizações para a estação de Vancouver com 4 métodos de estimação.

	Vancouver	day	REML	P-REML	GCV.Cp	GACV.Cp
jan01	2.30000	1	2.04601	2.04590	2.05941	2.05900
jan02	2.10000	2	2.11015	2.11005	2.12251	2.12213
jan03	1.90000	3	2.17428	2.17418	2.18561	2.18525
jan04	2.00000	4	2.23838	2.23829	2.24867	2.24835
jan05	1.60000	5	2.30244	2.30237	2.31171	2.31142
...						

Suavização com penalização (derivadas)

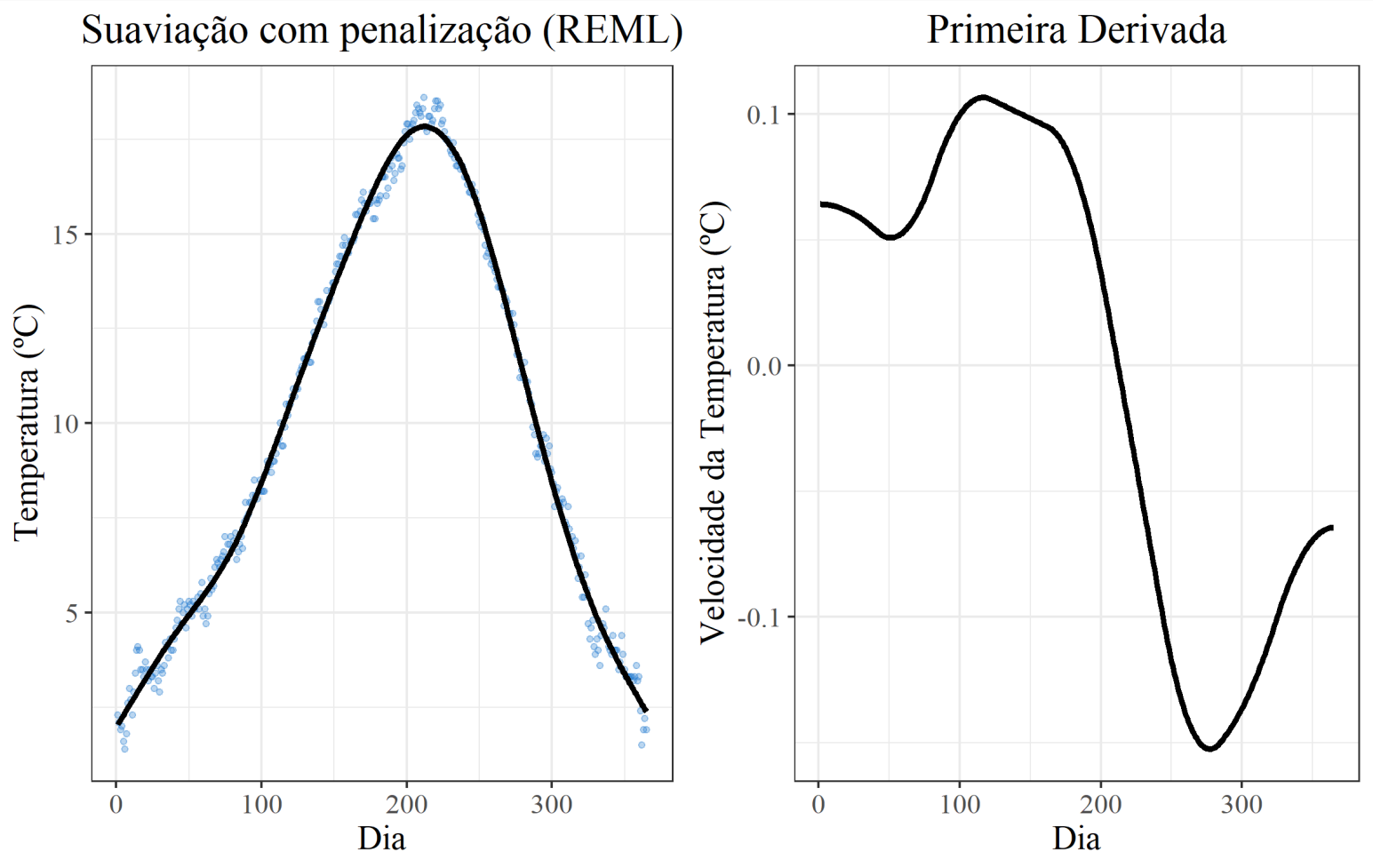

Figura 20: Suavização para a estação de Vancouver (estimação por REML) e primeira derivada.

Suavização com penalização (derivadas)

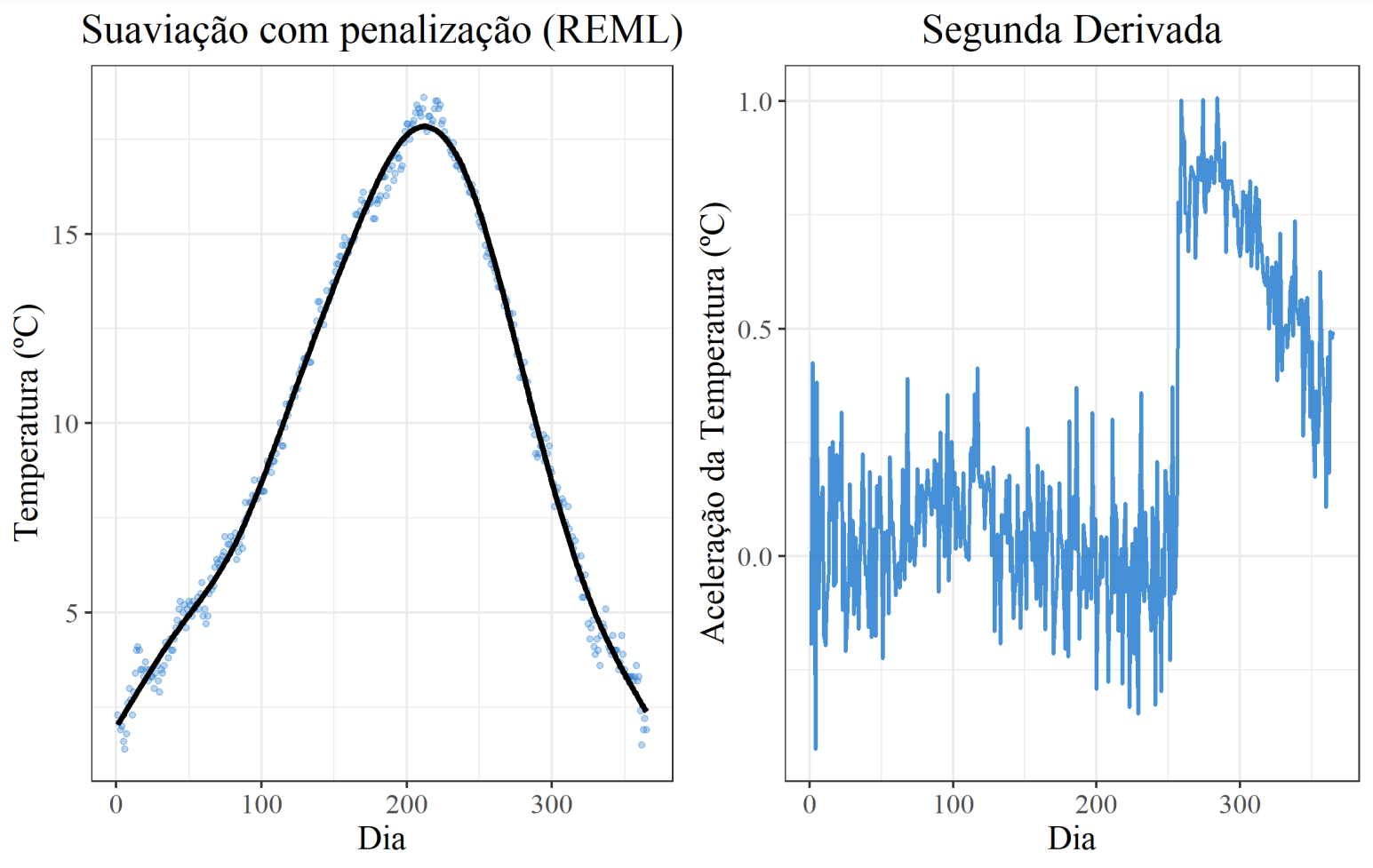

Figura 21: Suavização para a estação de Vancouver (estimação por REML) e segunda derivada.

Suavização com penalização (derivadas)

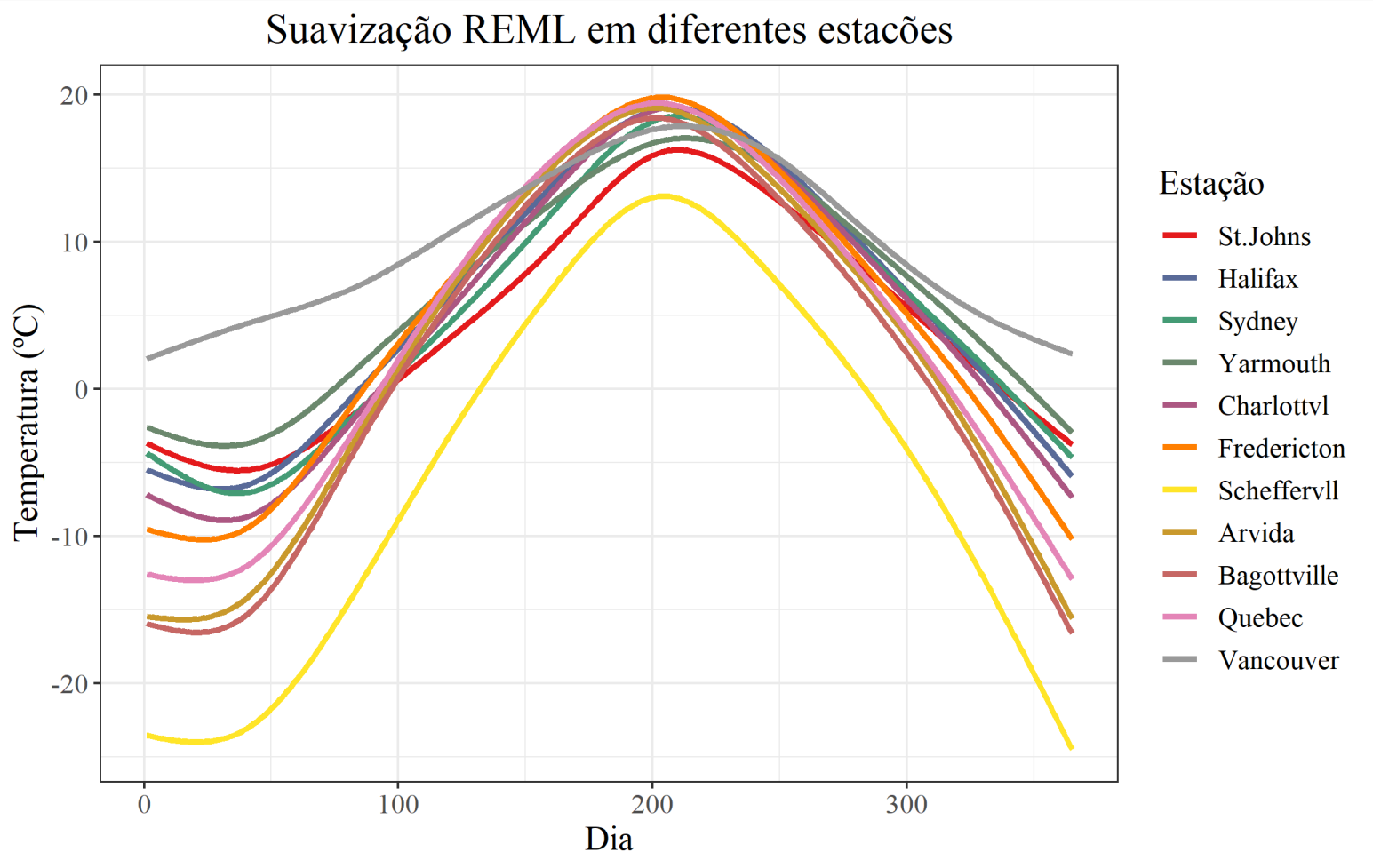

Figura 22: Suavizações para 11 estações (estimação por REML).

Extensões da análise de dados funcionais

Extensões

- Análise de componentes principais funcional
  - Matriz de covariância $\sum \implies$ Superfície de covariância $\sigma(s,t)$;
  - Decomposição por valores singulares $\implies$ decomposição Karhunen-Loève;
- Exploratória de dados funcionais
  - Suavização;
  - ACP Funcional;
  - Covariância Funcional;
- Regressão funcional
  - $y_i = \alpha + \sum \beta_j x_i(t_j) + \epsilon$
  - ACP Funcional + Regressão
  - $y_i(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_i(t) x_{ij}$;
- E ainda mais;

Referências

C de Boor. A practical guide to splines. revised edition. new-york: Springer. 2001.

Ulf Grenander. Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3):195–277, 1950.

Jürgen Kleffe. Principal components of random variables with values in a seperable hilbert space. *Mathematische Operationsforschung und Statistik*, 4(5):391–406, 1973.

James O Ramsay and Bernhard W Silverman. Functional data analysis. *İnternet Adresi: http*, 2008.

JO Ramsay, Giles Hooker, and Spencer Graves. Introduction to functional data analysis. In *Functional data analysis with R and MATLAB*, pages 1–19. Springer, 2009.