



Programa Ares-Rio
Ar e Saúde
Rio de Janeiro

Instituto de Medicina Social
Universidade do Estado
do Rio de Janeiro



Imputação de dados faltantes em séries temporais de poluição atmosférica

Washington Junger Antonio Ponce de Leon

Laboratório de Estatística Aplicada com
ênfase em Dados Dependentes - LEADD

VII Congresso Brasileiro de Epidemiologia

Porto Alegre, 23 de setembro de 2008

Problema

Motivação em epidemiologia ambiental:

Considere um estudo ecológico. As concentrações de material particulado são monitoradas em vários pontos da cidade a cada hora e médias diárias são calculadas. Alguns monitores apresentam lacunas. A exposição é estimada como a média das concentrações diárias nos monitores.

Dia	Centro	Jacarepaguá	Nova Iguaçu
01/02/2001	37.4167		39.1250
02/02/2001	53.0417	39.0000	44.9583
03/02/2001		33.7917	
04/02/2001		25.9167	31.2500
05/02/2001		27.3750	45.2500
06/02/2001		23.2917	30.7500
07/02/2001	45.9583		42.6667
08/02/2001	32.5833	29.5833	57.2083

Como estimar a exposição com o menor erro possível?

Procedimentos usuais

- **Análise dos registros completos**
 - **Imputação com a média incondicional**
 - **Imputação com a mediana**
 - **Imputação pelo vizinho mais próximo univariado**
 - **Imputação com a média condicional**
-

Algoritmo de imputação

- Baseado no algoritmo EM para a distribuição normal multivariada
- Seja x_t ($t=1, \dots, n$) a t -ésima realização de um vetor aleatório X com distribuição normal multivariada de dimensão p com m elementos não observados.

$$\mathbf{x}_t = \left(x_{t1}, \dots, x_{tm}, x_{t(m+1)}, \dots, x_{tp} \right)^T = \left(\mathbf{x}_{t1}, \mathbf{x}_{t2} \right)^T$$

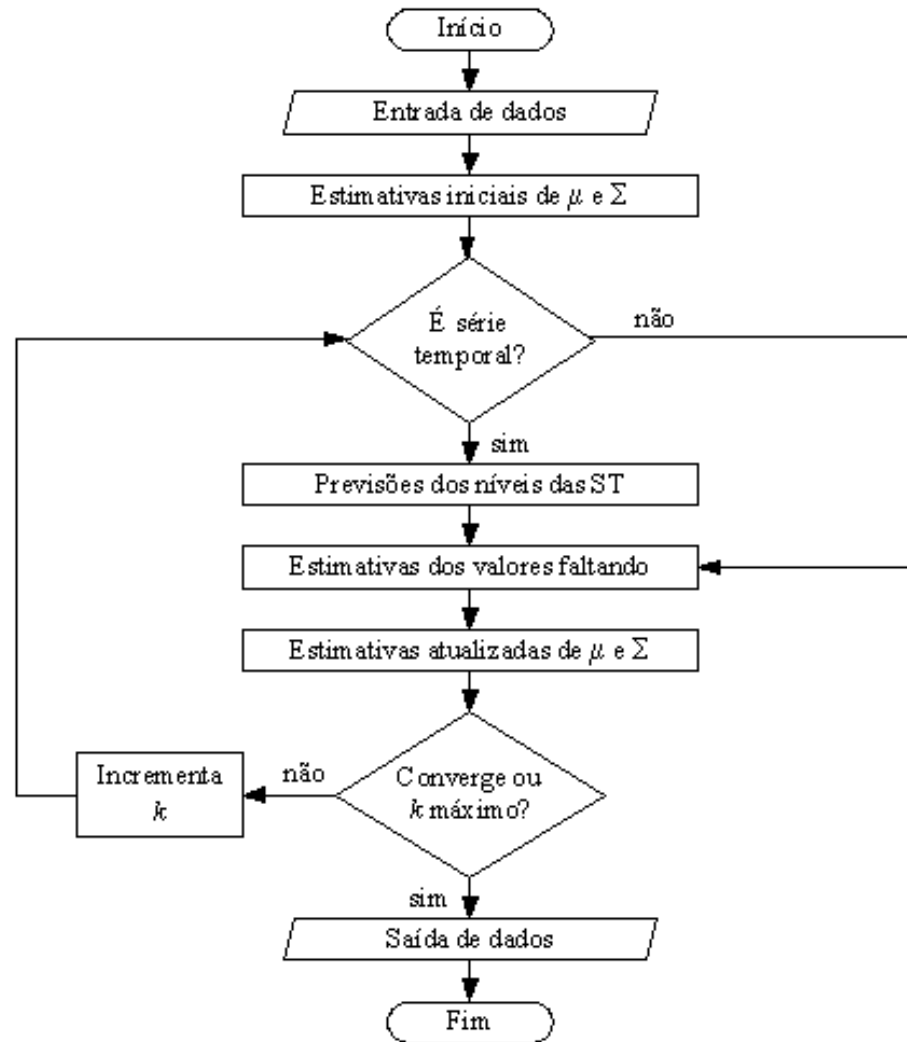
- Considere B janelas de com diferentes regimes de variância ao longo do tempo. O vetor média e a matriz de covariâncias em um instante de tempo t na janela b ($b=1, \dots, B$) são

$$\tilde{\boldsymbol{\mu}}_t = \begin{bmatrix} \tilde{\mu}_{t1} \\ \tilde{\mu}_{t2} \end{bmatrix} \quad \tilde{\boldsymbol{\Sigma}}_b = \begin{bmatrix} \tilde{\Sigma}_{b11} & \tilde{\Sigma}_{b12} \\ \tilde{\Sigma}_{b21} & \tilde{\Sigma}_{b22} \end{bmatrix}$$

- Após imputar os dados com a média condicional e calcular a contribuição, o vetor média e matriz de covariâncias em uma iteração do algoritmo são dados por

$$\tilde{\boldsymbol{\mu}}_b = \sum_{t=1}^{n_b} \tilde{\mathbf{x}}_{bt} / n_b \quad \tilde{\boldsymbol{\Sigma}}_b = \sum_{t=1}^{n_b} \tilde{\mathbf{x}}_{bt} \tilde{\mathbf{x}}_{bt}^T / n_b - \tilde{\boldsymbol{\mu}}_b \tilde{\boldsymbol{\mu}}_b^T$$

Algoritmo de imputação



Componente temporal

A cada iteração é preciso estimar o nível μ_t da série temporal

- **Modelo ARIMA(p, d, q)**

$$\nabla^d x_{jt} = \phi_1 x_{jt-1} + \phi_2 x_{jt-2} + \dots + \phi_p x_{jt-p} + a_{jt} - \theta_1 a_{jt-1} - \theta_2 a_{jt-2} - \dots - \theta_q a_{jt-q}$$

$$\tilde{\mu}_{jt} = E \left[X_{jt} \mid x_{j(t-1)}, x_{j(t-2)}, \dots \right]$$

- **Spline cúbica natural**

$$S(g_j) = \sum_{k=1}^K \{X_t - g(v_k)\}^2 + \lambda \int_a^b \{g''\}^2 dx$$

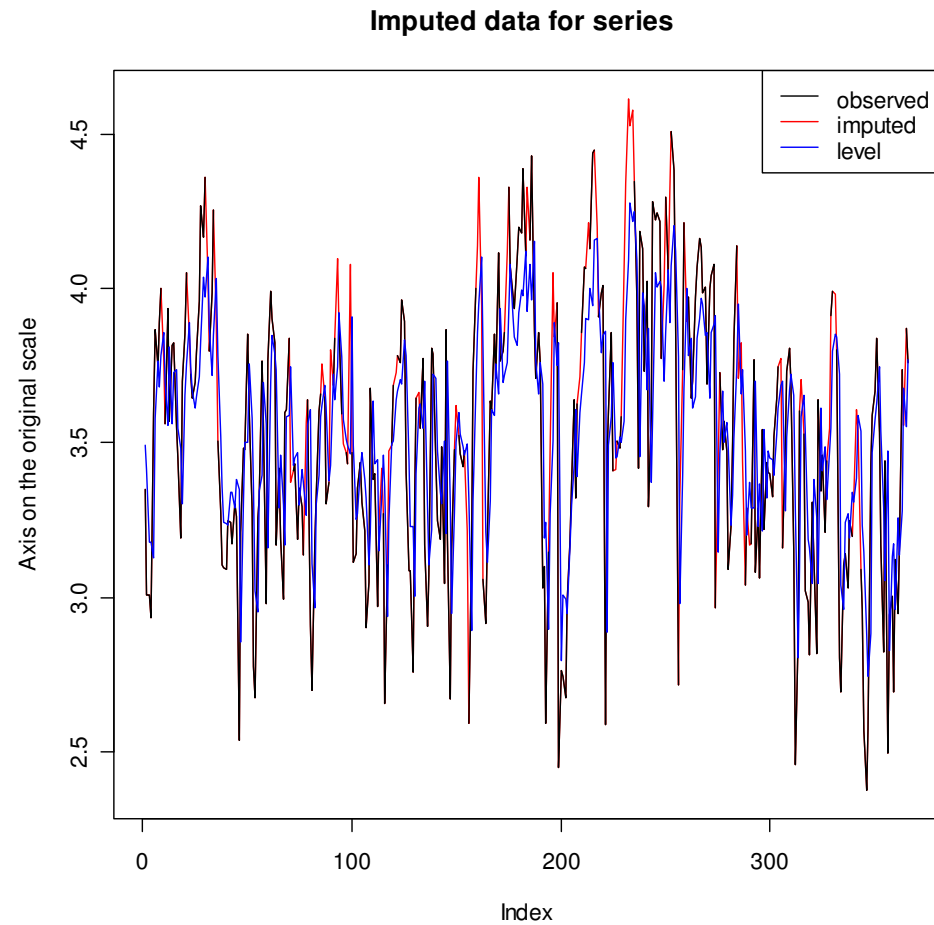
$$\mu_{jt} = g(x_{jt})$$

- **Modelo de regressão (GAM)**

$$\mu_{jt} = \beta_0 + \sum_u \beta_u Z_u + \sum_v g_v(Z_v)$$

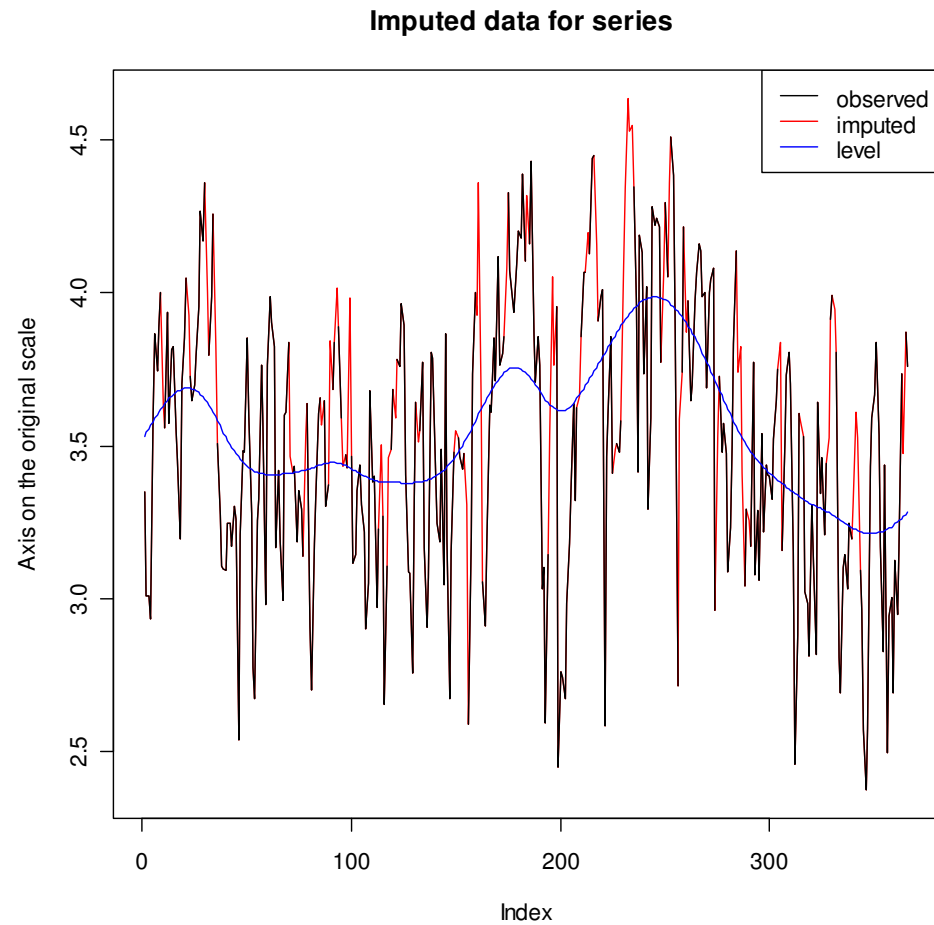
Componente temporal

Exemplo de ajuste do componente temporal usando *ARIMA*



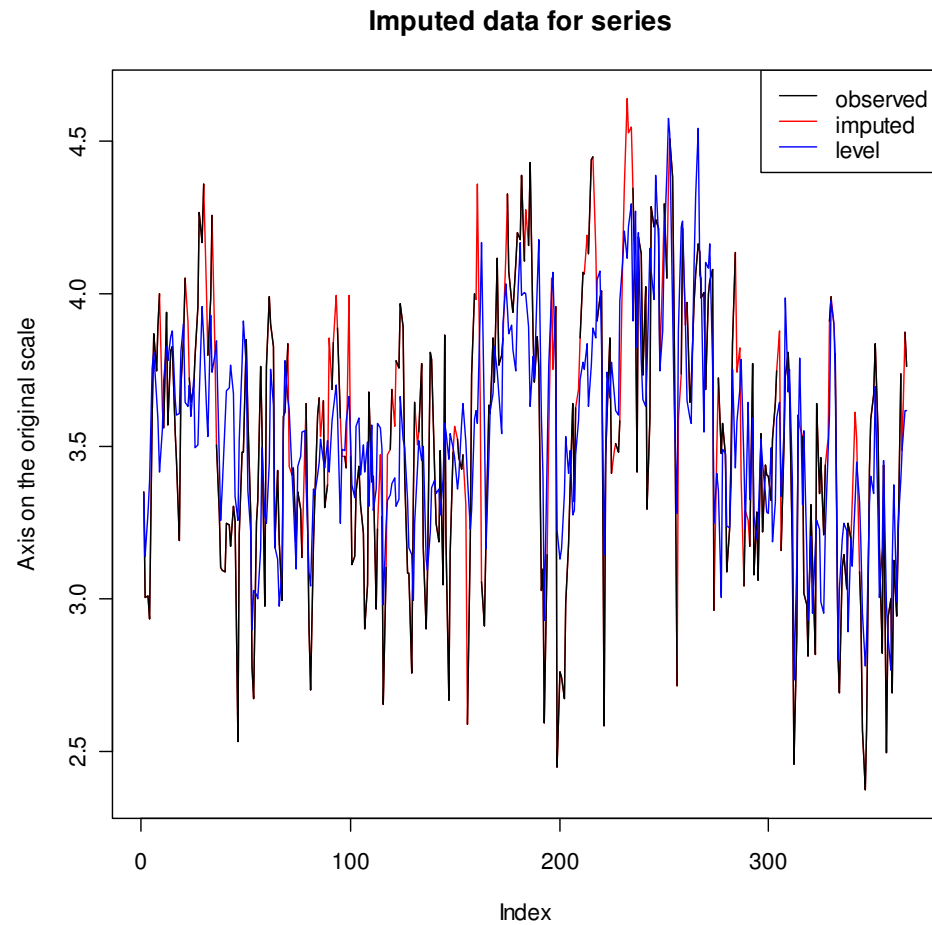
Componente temporal

Exemplo de ajuste do componente temporal usando *spline*



Componente temporal

Exemplo de ajuste do componente temporal usando *GAM*



Estudo de simulação

Para avaliar a efetividade do método, este foi submetido ao seguinte desenho de simulação

- **Um conjunto de dados de concentrações de PM10 da cidade de São Paulo sem variáveis faltantes composto de 10 estações e 366 observações**
- **Modelo de regressão Poisson semi-paramétrico para internação de crianças até 5 anos de idade por doenças respiratórias, controlando por tendência, sazonalidade, dias da semana, feriados temperatura e umidade**
- **O risco relativo (percentual) estimado para a média de PM10 foi 0,433% com IC de 95% igual a (0,224;0,643). O coeficiente estimado foi 0,004321 (0,001063)**
- **Para avaliar a validade da imputação, a estratégia da simulação foi gerar 100 padrões de dados incompletos de diversos tamanhos e configurações na matriz de concentrações de poluentes usando os mecanismos MCAR, MAR e MNAR; imputar os valores faltantes, estimar o efeitos com os dados imputados e calcular medidas sumárias das estimativas**
- **Para avaliar a performance, apenas uma replicação foi usada e os seguintes indicadores calculados: RMSD, MAD, BIAS, PV, r e d2**
- **Também foi avaliado o impacto de um fator de penalização pela informação perdida sobre os coeficientes do modelo de associação**

Resultados

Validade: MAR

	CD	IM	MD	VP	CM	EM	EM Spline	EM ARIMA	EM MAG	EM-MR Spline	EM-MR ARIMA	EM-MR MAG
MAR												
5%	0,432 (0,006)	0,445 (0,012)	0,445 (0,012)	0,436 (0,008)	0,433 (0,005)	0,433 (0,004)	0,433 (0,004)	0,433 (0,005)	0,433 (0,004)	0,432 (0,004)	0,432 (0,006)	0,432 (0,004)
10%	0,431 (0,010)	0,463 (0,017)	0,463 (0,017)	0,439 (0,010)	0,432 (0,009)	0,435 (0,007)	0,433 (0,006)	0,435 (0,009)	0,433 (0,007)	0,432 (0,007)	0,435 (0,012)	0,432 (0,007)
20%	0,429 (0,015)	0,490 (0,028)	0,490 (0,028)	0,441 (0,017)	0,430 (0,014)	0,436 (0,009)	0,434 (0,009)	0,438 (0,013)	0,433 (0,009)	0,433 (0,010)	0,437 (0,016)	0,433 (0,010)
30%	0,422 (0,018)	0,518 (0,040)	0,519 (0,041)	0,447 (0,021)	0,422 (0,018)	0,438 (0,013)	0,434 (0,012)	0,437 (0,020)	0,433 (0,011)	0,434 (0,013)	0,436 (0,022)	0,434 (0,012)
40%	0,419 (0,022)	0,551 (0,047)	0,552 (0,048)	0,454 (0,025)	0,419 (0,022)	0,443 (0,015)	0,435 (0,015)	0,439 (0,024)	0,433 (0,017)	0,435 (0,016)	0,437 (0,028)	0,433 (0,018)

Resultados

Performance : MAR

	Ind.	IM	MD	VP	CM	EM	EM Spline	EM ARIMA	EM MAG	EM-MR Spline	EM-MR ARIMA	EM-MR MAG
MAR												
5%	RMSD	0,421	0,416	0,344	0,124	0,150	0,138	0,155	0,155	0,133	0,159	0,159
	MAD	0,346	0,342	0,256	0,076	0,117	0,108	0,119	0,119	0,103	0,117	0,117
	BIAS	0,286	0,281	0,059	-0,008	0,001	-0,010	0,006	0,006	-0,011	0,005	0,005
	PV	0,094	0,114	1,633	0,783	0,804	0,961	0,988	0,988	0,910	0,979	0,979
	r	0,499	0,504	0,660	0,885	0,903	0,920	0,900	0,900	0,925	0,895	0,895
	d2	0,526	0,534	0,790	0,937	0,946	0,959	0,948	0,948	0,960	0,945	0,945
40%	RMSD	0,517	0,521	0,403	0,039	0,206	0,186	0,191	0,191	0,181	0,197	0,197
	MAD	0,422	0,425	0,304	0,007	0,156	0,140	0,143	0,143	0,136	0,150	0,150
	BIAS	0,383	0,390	0,136	-0,002	0,001	-0,002	0,019	0,019	0,004	0,019	0,019
	PV	0,077	0,095	1,658	0,770	0,694	0,827	0,876	0,876	0,852	0,927	0,927
	r	0,416	0,419	0,640	0,889	0,839	0,871	0,866	0,866	0,879	0,860	0,860
	d2	0,500	0,503	0,767	0,971	0,906	0,930	0,928	0,928	0,935	0,925	0,925

Resultados

Penalização: coeficiente e erro-padrão sob MAR com pesos $w_i = 1 - 0.5(m_i/p)$

Valor de referência: 0,004321 (0,001063)

%	Penali- zação	Estatística	EM Spline	EM ARIMA	EM MAG	EM-MR Spline	EM-MR ARIMA	EM-MR MAG
5%	não	β	0,004316	0,004248	0,004330	0,004295	0,004261	0,004314
		EP(β)	0,001061	0,001060	0,001062	0,001061	0,001058	0,001061
	sim	β	0,004301	0,004236	0,004313	0,004281	0,004247	0,004299
		EP(β)	0,001066	0,001066	0,001067	0,001066	0,001063	0,001066
10%	não	β	0,004324	0,004202	0,004321	0,004307	0,004238	0,004305
		EP(β)	0,001061	0,001063	0,001061	0,001063	0,001071	0,001061
	sim	β	0,004343	0,004229	0,004340	0,004328	0,004264	0,004325
		EP(β)	0,001077	0,001079	0,001076	0,001079	0,001086	0,001077
20%	não	β	0,004334	0,004137	0,004307	0,004285	0,004115	0,004277
		EP(β)	0,001057	0,001058	0,001055	0,001057	0,001059	0,001054
	sim	β	0,004340	0,004162	0,004314	0,004295	0,004146	0,004287
		EP(β)	0,001088	0,001089	0,001086	0,001088	0,001091	0,001085
30%	não	β	0,004362	0,004103	0,004394	0,004335	0,004097	0,004398
		EP(β)	0,001064	0,001056	0,001063	0,001063	0,001066	0,001064
	sim	β	0,004352	0,004124	0,004375	0,004326	0,004118	0,004376
		EP(β)	0,001113	0,001107	0,001112	0,001112	0,001117	0,001113
40%	não	β	0,004436	0,004050	0,004473	0,004426	0,004006	0,004473
		EP(β)	0,001074	0,001050	0,001074	0,001073	0,001031	0,001074
	sim	β	0,004466	0,004126	0,004494	0,004456	0,004091	0,004492
		EP(β)	0,001137	0,001118	0,001136	0,001136	0,001103	0,001136

Considerações

- Apesar de ser desenvolvido sob normalidade, o método poder ser utilizado em qualquer problema que possa ser formulado como um problema de estimação de parâmetros de uma normal multivariada
 - O principal objetivo do algoritmo é realizar imputação com dependência temporal, entretanto pode ser utilizado com dados transversais
 - O uso de um fator de penalização reduz a superestimação da precisão dos estimadores
 - O algoritmo está implementado numa biblioteca para o *R* chamada *mtsdi* de fácil utilização e páginas de ajuda
 - É a metodologia de imputação de dados do Projeto ESCALA
 - A imputação de dados não deve ser vista como uma forma de fabricar dados que não existem, e sim permitir que se faça inferências usando o máximo possível da informação contida nos dados
-

www.ims.uerj.br/ares-rio

wjunger@ims.uerj.br



Programa Ares-Rio

Ar e Saúde - Rio de Janeiro



Instituto de Medicina Social

Universidade do Estado do Rio de Janeiro
