# Amostragem Preferencial: Modelação de dados autocorrelacionados

**Andreia Monteiro**
*Universidade do Minho & CIDMA & CMUM, andreiaforte50@gmail.com*

Raquel Menezes
*Universidade do Minho & CMUM, rmenezes@math.uminho.pt*

Maria Eduarda Silva
*Faculdade de Economia da Universidade do Porto & CIDMA, mesilva@fep.up.pt*

**Palavras–chave**: Amostragem Preferencial; Séries Temporais; Processos Autoregressivos em Tempo Contínuo

**Resumo**: A necessidade de modelos espaço-temporais, aparece em vários contextos, como por exemplo nas ciências ambientais, agricultura, ecologia entre outros. Tradicionalmente a modelação espacial e temporal assume que as localizações das amostras (no tempo ou no espaço) são selecionadas sem terem em conta os valores do próprio processo em estudo. Contudo, em muitas situações há uma dependência estocástica entre os locais de amostragem e o próprio processo subjacente ao estudo.

A amostragem igualmente espaçada é, provavelmente, o esquema mais frequentemente utilizado na prática, mesmo quando os dados são recolhidos ao longo do tempo numa determinada região. No entanto, dados amostrados de forma irregular podem ocorrer em diversas situações. Um caso particular de dados recolhidos de forma irregular é o de dados em que a sua recolha ao longo do tempo, depende, por determinadas razões, dos próprios valores observados. Por exemplo, um determinado indicador médico do estado de saúde de um indivíduo pode ser medido em diferentes intervalos de tempo e com diferentes frequências, dependendo do próprio estado de saúde. Num contexto completamente diferente, os tempos de ocorrências das transações nos mercados financeiros dependem em larga medida do valor dos ativos subjacentes. Desta forma, informação adicional do fenómeno em estudo é obtida a partir da frequência ou dos tempos de ocorrência das observações. Nestas situações, há uma dependência estocástica entre o processo que vai ser modelado e os tempos das observações.

Este problema foi primeiramente identificado no contexto da estatística espacial, por [2], que lhe atribuiu o nome de amostragem preferencial. Diggle e os seus coautores demonstraram que ignorar a natureza preferencial da amostragem pode levar a estimativas enviesadas. O nosso trabalho estende o conceito de amostragem preferencial e o modelo apresentado por [2] à componente temporal.

Desta forma, se $Y_i$ denotar o valor medido no tempo $t_i$, um modelo simples para os dados assume a forma

$$Y_i = \mu + S(t_i) + Z_i,\ i = 1, ..., n$$

onde $Z_i$ é um erro de medição, com média zero e variância $\tau^2$ e $S(.)$, um processo estocástico não observado, é um processo autoregressivo em tempo contínuo de ordem 1, CAR(1), [1]. Tem-se assim que:

- $Y = (Y_1, ...Y_n)^t \sim MVN\left(\mu_y \mathbf{1}, \Sigma_y\right)$

  com $\mu_y = \mu\mathbf{1}$ e $\Sigma_y = \frac{\sigma_w^2}{2\alpha_0} R_y(\alpha_0) + \tau^2 I_n$ em que $\mathbf{1}$ é um vetor de 1's, $I_n$ é a matriz identidade $n \times n$ e $R_y(\alpha_0)$ é uma matriz $n \times n$ em que o elemento $(i,j)^{th}$ é $\rho(|t_i - t_j|)$ definido por $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = e^{-\alpha_0|h|}$

- $S = (S(t_1), ..., S(t_N))^t \sim MVN\left(0, \Sigma_s\right)$

  em que $\Sigma_s = \frac{\sigma_w^2}{2\alpha_0} R_s(\phi)$, onde $R_s(\phi)$ é a matriz de correlação $N \times N$ do processo CAR(1).

O modelo para lidar com amostragem preferencial é definido através da distribuição conjunta $[S, T, Y]$ escrita como $[S][T\,|S][Y\,|S(T)]$, onde [.] significa a "distribuição de" e escreve-se $S = \{S(t) : t \in \mathbb{R}\}$, $T = (t_1, ..., t_n)$, $S(T) = \{S(t_1), ..., S(t_n)\}$ and $Y = (Y_1, ...Y_n)$. São necessárias também algumas suposições:

- $[S]$ é um processo Gaussiano estacionário;
- $[T\,|S]$ é um processo de Poisson não homogéneo com intensidade $\lambda(t) = \exp\{\alpha + \beta S(t)\}$;
- Condicionado a $S$ e $T$, $Y$ é um conjunto de variáveis gaussianas mutuamente independentes, $[Y_i|S(t_i)] \sim N\left\{\mu + S(t_i), \tau^2\right\}$, com $\tau^2$ igual à variância do erro de medição.

O modelo é estimado por máxima verossimilhança e com recurso a simulações de Monte Carlo. Neste trabalho, realizamos um estudo de simulação para mostrar os benefícios do modelo proposto relativamente à abordagem "clássica" e aplicamos o modelo a um conjunto de dados reais.

Como trabalho futuro, pretendemos aplicar o conceito de amostragem preferencial a dados que apresentem uma estrutura espacial e temporal. Um exemplo desse tipo de dados aparece no contexto dos projetos das "cidades inteligentes". Nestes projetos, os dados são recolhidos de diferentes formas e são analisados para que os sistemas urbanos sejam melhor compreendidos e a qualidade de vida melhorada. Estes dados tipicamente exigem o desenvolvimento de ferramentas estatísticas capazes de lidar com uma grande quantidade

de dados recolhidos através de sensores de monitorização. Estes dados espaço-temporais podem não satisfazer algumas das suposições usuais em análise de dados. De facto, por questões práticas o desenho amostral da rede de sensores pode não representar uniformemente a região de observação, por exemplo se mais sensores são colocados em áreas consideradas mais críticas isto leva a uma amostragem preferencial no espaço. De modo similar, a recolha dos dados ao longo do tempo pode depender dos valores observados, por exemplo pode ter-se decidido monitorizar mais frequentemente quando um valor, considerado crítico para a saúde humana, é excedido.

**Bibliografia**

[1] Brockwell, P.J., Continuous-time arma processes. In D.N.Shanbhag and C.R.Rao, editors, Handbook of Statistics 19 - Stochastic Processes: Theory and Methods, pp 249-276, Elsevier, 2001.

[2] Diggle P.J., Menezes R., Su T., Geostatistical inference under preferential sampling. Applied Statistics, 59(2), pp 1-20, 2010.