

3 Dados

As séries temporais tem $60 \times 24 \times 7 = 10080$ pontos. De fato elas correspondem a 7 dias de uma medida tomada minuto a minuto. Cada série temporal tem um código que corresponde a um dado sujeito. São 96 sujeitos. Cada sujeito está classificado em 3 categorias. O arquivo “timeseries.NEW.csv” tem as séries temporais dispostas como 96 colunas com um cabeçalho correspondendo ao código do sujeito e o arquivo “timeseries.classification.csv” tem duas colunas a primeira o código do sujeito e a segunda sua classificação.

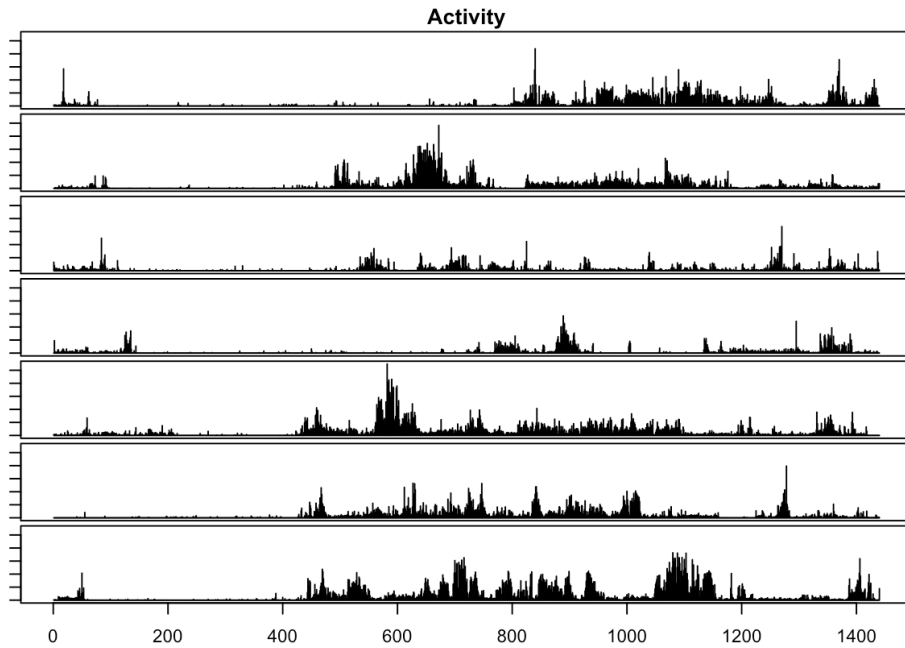


Figure 1:

Na figura 1 é possível ver uma destas séries temporais divididas em 7 segmentos correspondentes a um dia cada um (1440 minutos).

A proposta do trabalho é usar métricas (medidas) das propriedades da série temporal como preditor das categorias dos sujeitos que geraram elas. Eu falei em aula que umas 10 métricas que não fosse muito parecidas ou correlacionadas já seria um bom começo. Mas claro, quanto mais melhor.

Considere que a série temporal seja representada por uma variável discreta tempo t_i e uma variável contínua a_i , correspondente ao dado da série coletado ao tempo t_i .

Por exemplo, como citado em aula a partir da série podemos imaginar algumas métricas para servirem de preditores das 3 classes. São métricas possíveis

por exemplo, o valor médio da série temporal

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i \quad (1)$$

onde $n = 10080$. Tem também a variância

$$\text{var}(a) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})^2 \quad (2)$$

que pode dar uma ideia muito superficial quanto o sujeito é irregular. Uma possibilidade mais interessante para capturar irregularidade semanal é calcular a variância da média entre as semana. Ou seja para cada semana se calcula uma média

$$\bar{a}_s = \frac{1}{1440} \sum_{i \in S} a_i \quad (3)$$

e a variância semanal

$$\text{var}_s(a) = \frac{1}{7} \sum_{s=1}^7 (\bar{a} - \bar{a}_s)^2 \quad (4)$$

Outras possibilidades são por exemplo:

- $a_{max} = \max(a_i)$. O valor máximo e possíveis combinações com os valores máximos semanais.
- $\Delta t_{>f} = \frac{1}{n} \sum_i H(f * a_{max} - a_i)$ onde $H(x)$ é a função Heaviside ou função degrau e f é um número entre 0 e 1. Esta métrica diz a fração de tempo o sujeito ficou com atividade acima de $f*100\%$ do seu valor máximo de atividade.
- Pode-se fazer a mesma coisa usando o valor médio $\Delta t_{>f} = \frac{1}{n} \sum_i H(f * \bar{a} - a_i)$ onde $H(x)$ é a função Heaviside ou função degrau e f agora pode variar de 0 até um certo número. Esta métrica diz a fração de tempo o sujeito ficou com atividade acima de $f*100\%$ do seu valor médio de atividade.