Proposta de solução para o desafio Eldorado.

Israel Gonçalves de Oliveira

Instituto de Pesquisas Eldorado
Parque Científico e Tecnológico da PUCRS

Porto Alegre, fevereiro de 2018.

Contents

Introdução

- Base de dados no formato CSV, contendo 21.120 amostras com 24 características (*features*), sendo a primeira o alvo (*target*), a segunda a hora, seguida de mês e ano, a terceira e quarta, respectivamente. As demais 20 são variáveis não caracterizada (ou, nenhuma informação a priori). Todos as características são números, algumas amostras com informação faltante, nomeado com 'NA'.
- Objetivo: "realizar a previsão do Target para todos os meses de 2018 e para o ano de 2018".
- Com uma análise visual com evidências no domínio frequencial e com base nos histogramas, optou-se por utilizar apenas os valores de alvo e temporais de mês e ano. A solução proposta é baseada na esperança matemática, ou seja, usado o valor médio.
- A aplicação para análise e solução foi desenvolvida em MATLAB [1] rodando no GNU/Linux Ubuntu 16.04.3 LTS.

Análise: pré-formatação.

- Os dados foram disponibilizados no formato CSV, arquivo texto. Como os valores usavam vírgula com separador decimal, foi necessário uma substituição para o ponto, para que fosse lido corretamente (não apenas pelo MATLAB, mas para outros softwares testados).
- Os valores faltantes continham os caracteres 'NA', e foram substituídos por 'nan', para que o MATLAB lesse como um valor não numérico (NaN: *Not a Number*).
- Código utilizado:

```
$ cat data.csv |sed -e 's/\,/\./g' \
    |sed -e 's/NA/nan/g' >data_mod.csv
```

- Para tal processo poderia se usar um editor de texto, visto que o arquivo da base tinha tamanho de 5 MB, todavia isso seria inviável para arquivos muito maiores.

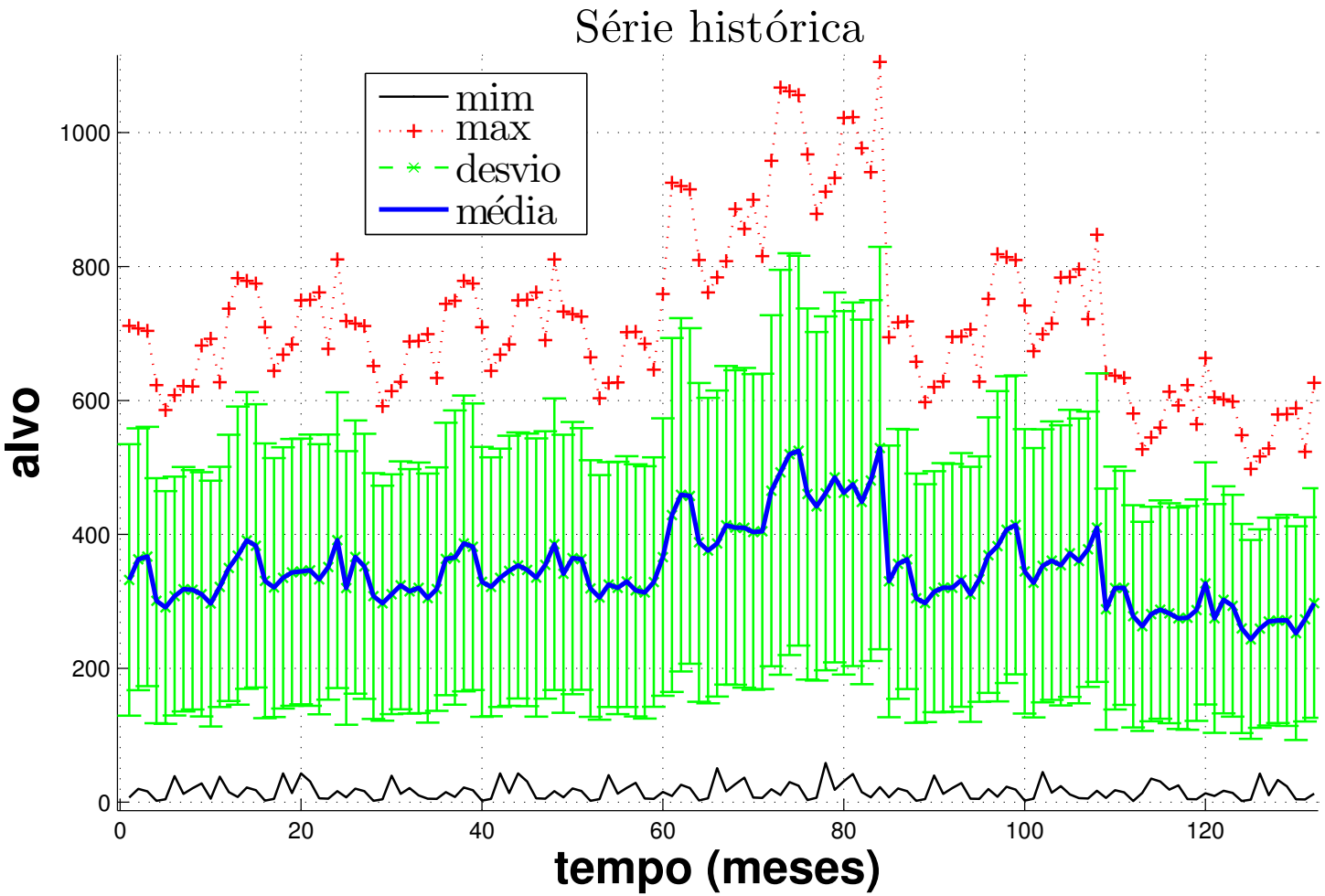
Análise: pré-formatação.

- Para as amostras com informação faltante apenas no mês e ano, foi realizada uma substituição do valor 'nan' pelo valor da amostra anterior. Esse processo contempla a grande maioria das amostras com 'NA' nos campos mês e ano de forma correta, visto que as amostras estavam em sequência temporal. Avaliou-se que o ganho de amostras válidas para uso traria um incremento de informação relevante com o custo de raras as amostras com valor possivelmente errado e mês e ano em uma unidade. Não foram aproveitadas as amostras com 'NA' no campo alvo.
- De um total de 21.120 amostras, foram aproveitadas 18.870, perdendo 2.250 (representando 10,6% da base original). Das amostras aproveitadas, 1.895 tiveram o ajuste no campo mês e 1.792 no campo ano, representando 10% e 9,5% das amostras aproveitadas, respectivamente.

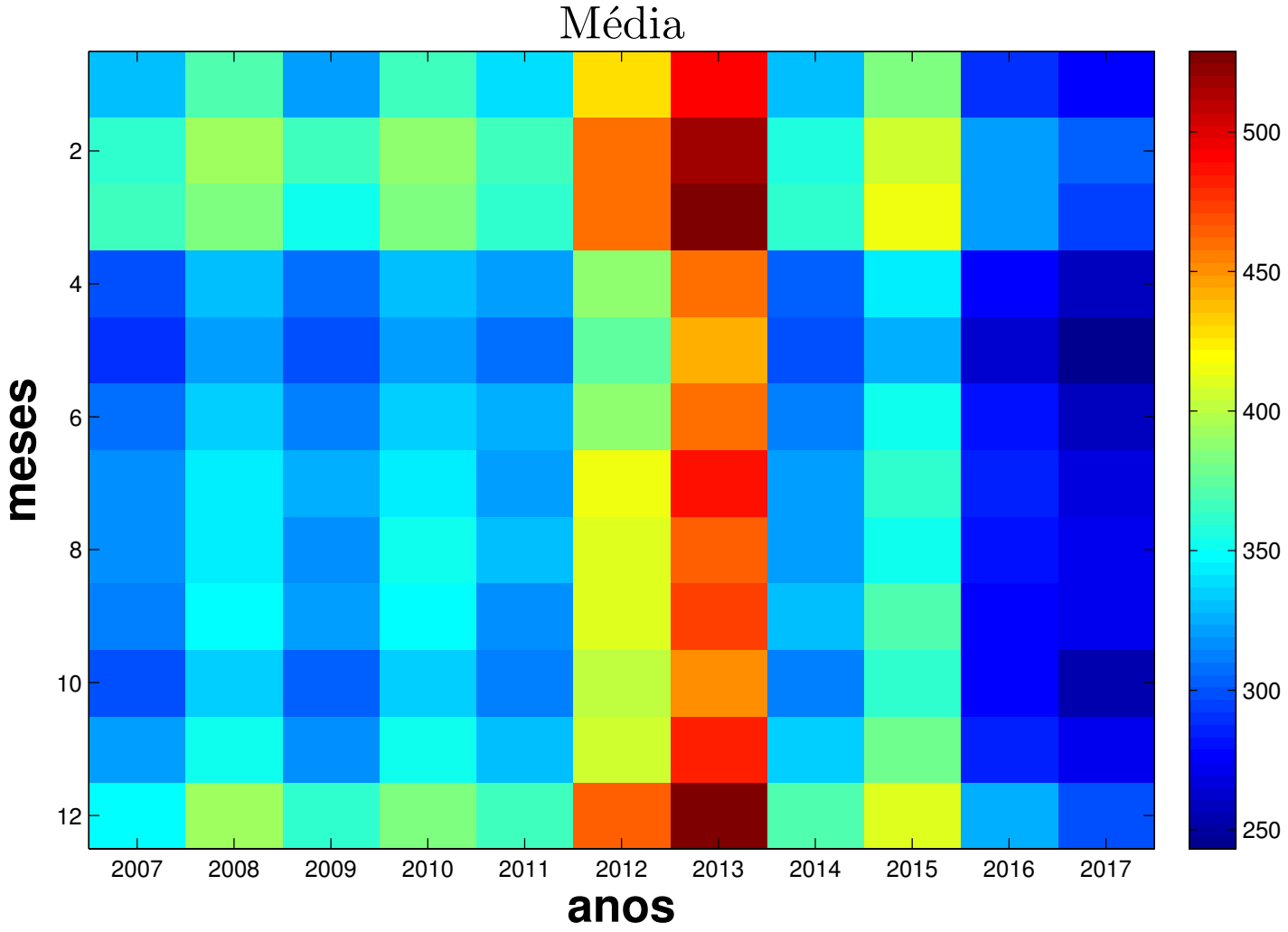
Análise: visualização

- Verificar a média, desvio padrão, máximo e mínimo mensal.
- Verificar a relação mensal e anual.
- Sem um padrão anual significativo.
- No gráfico dos valores mínimos há um padrão mensal aparente.
- Os anos de 2012 e 2013 parecem atípicos e os últimos dois anos apresentam uma queda contínua. Todavia, um padrão é identificado entre esses pares: o ano anterior se mostra como tendência.

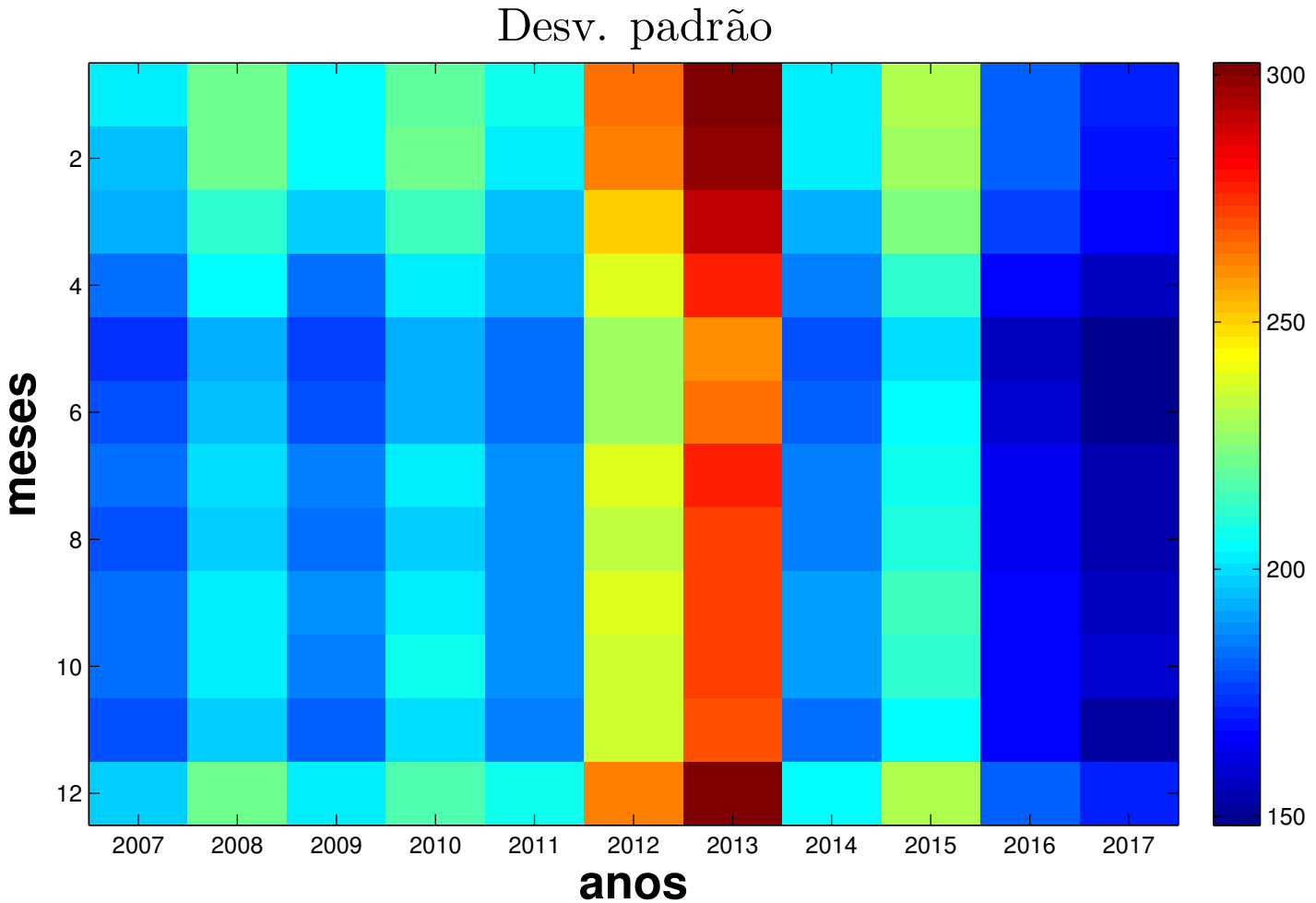
Análise: visualização



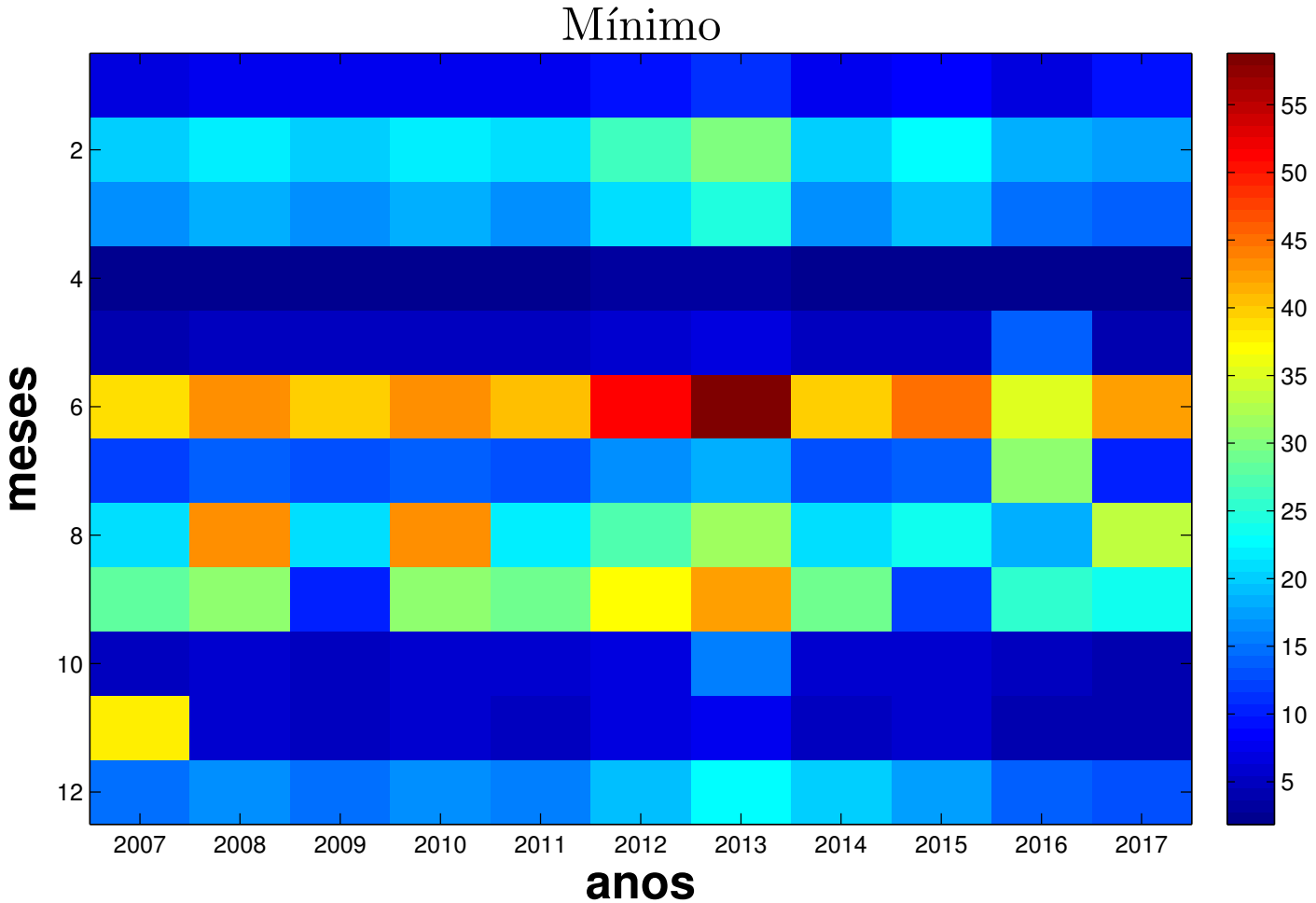
Análise: visualização



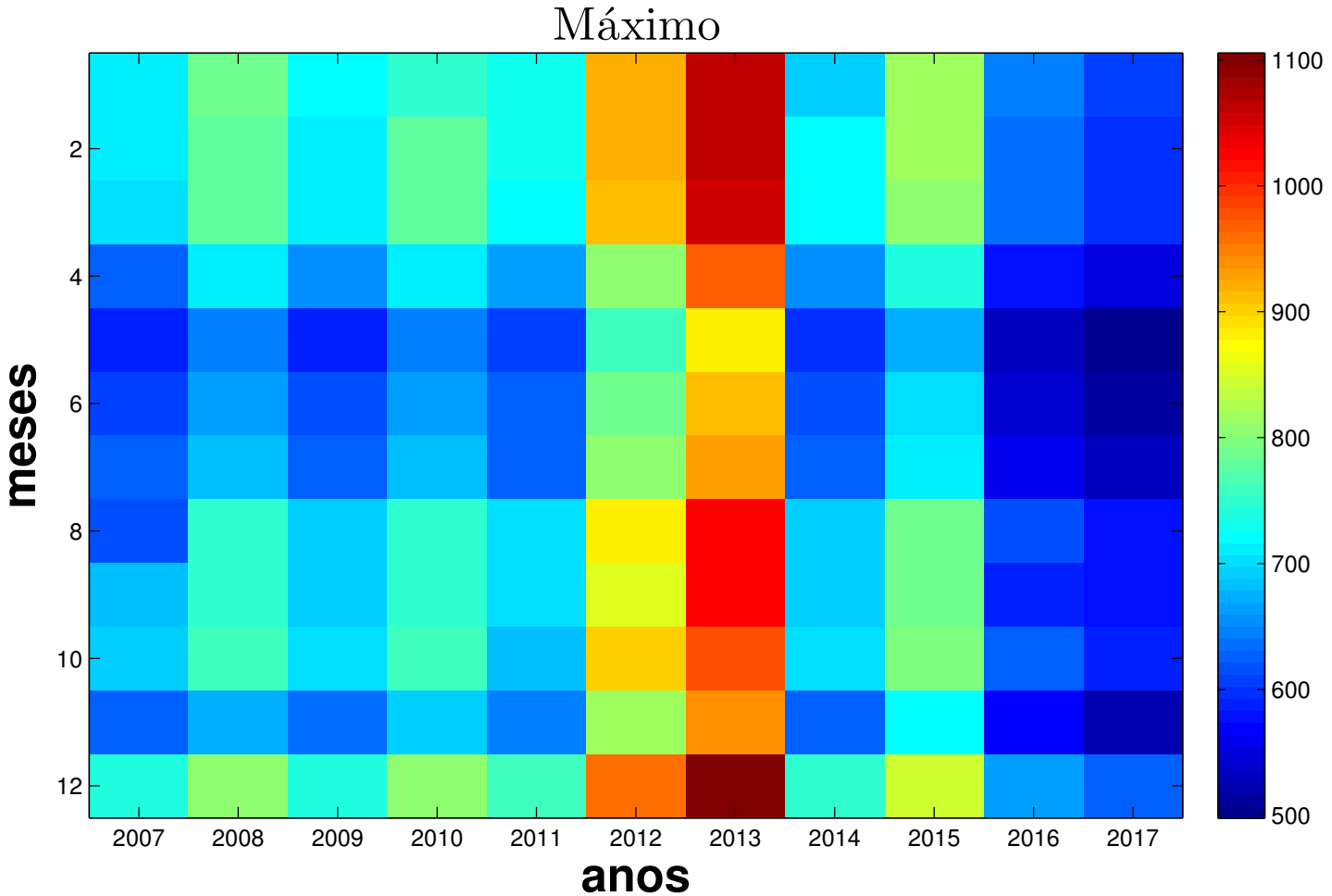
Análise: visualização



Análise: visualização



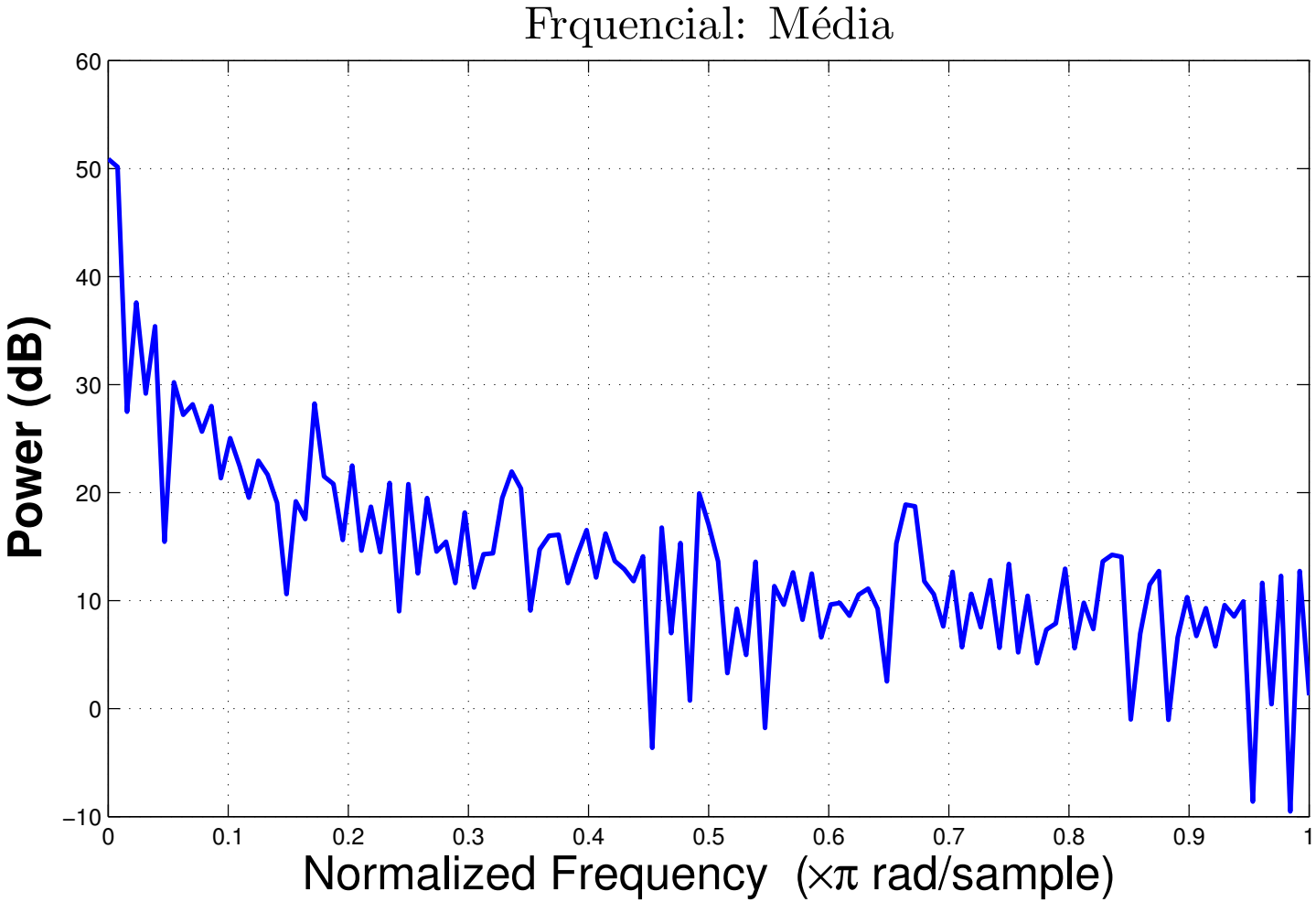
Análise: visualização



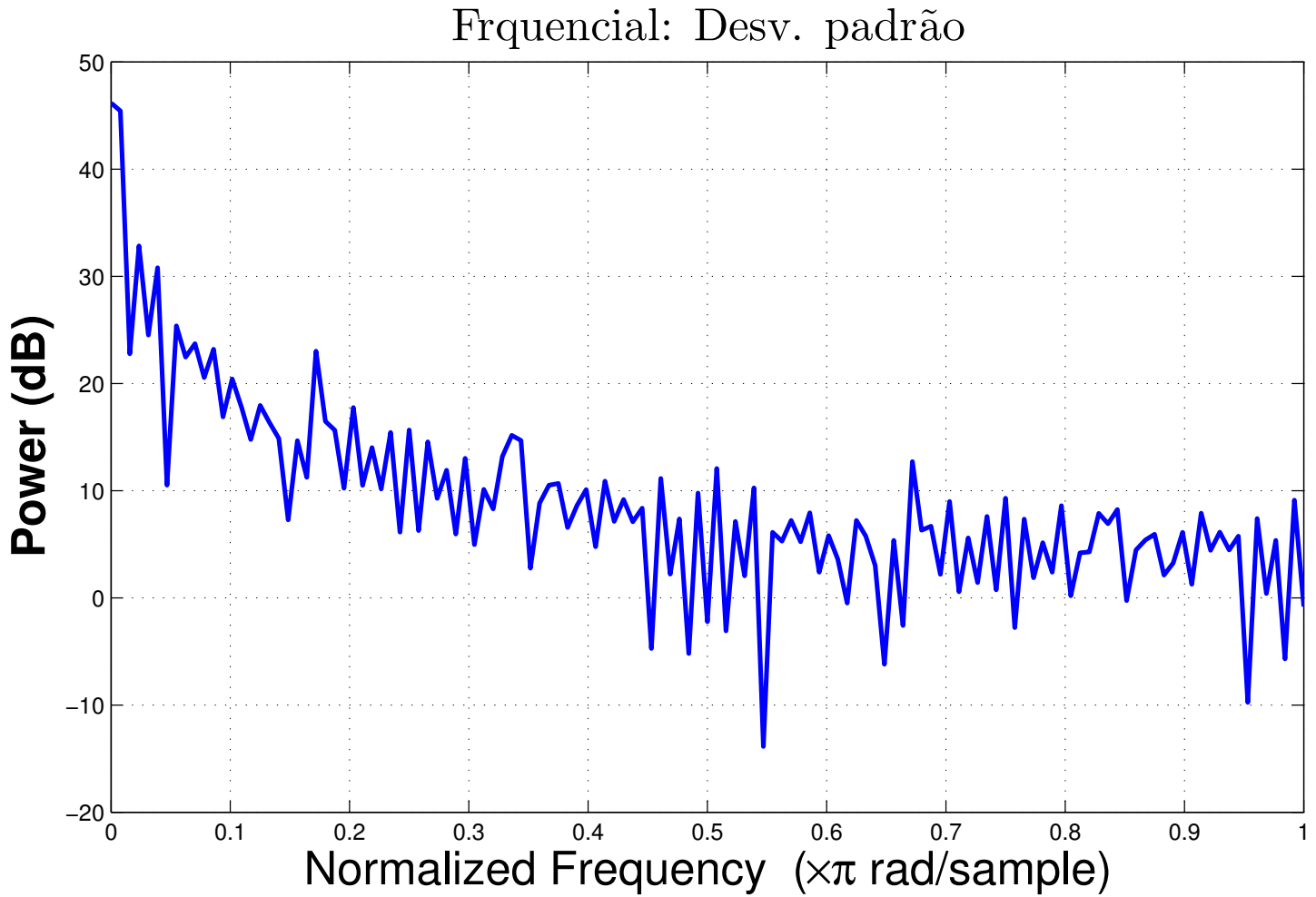
Análise: Visualização no domínio frequencial

- A fim de se obter evidências numéricas quanto aos padrões, pode-se usar uma análise frequencial. O grau de periodicidade de um sinal é relacionado a boa estimativa do sinal utilizando uma expansão de *Fourier*. Não apenas para estimativa mas também pode-se evidenciar o grau de ruído. Um alto desvio padrão pode ser ocasionado pelo ruído.
- Conforme a análise, a periodicidade mensal é predominante e não há aparente periodicidade anual.
- Componentes de baixa frequência são dominantes, sugerindo que o uso de valor médio poderia oferecer melhor estimativa.

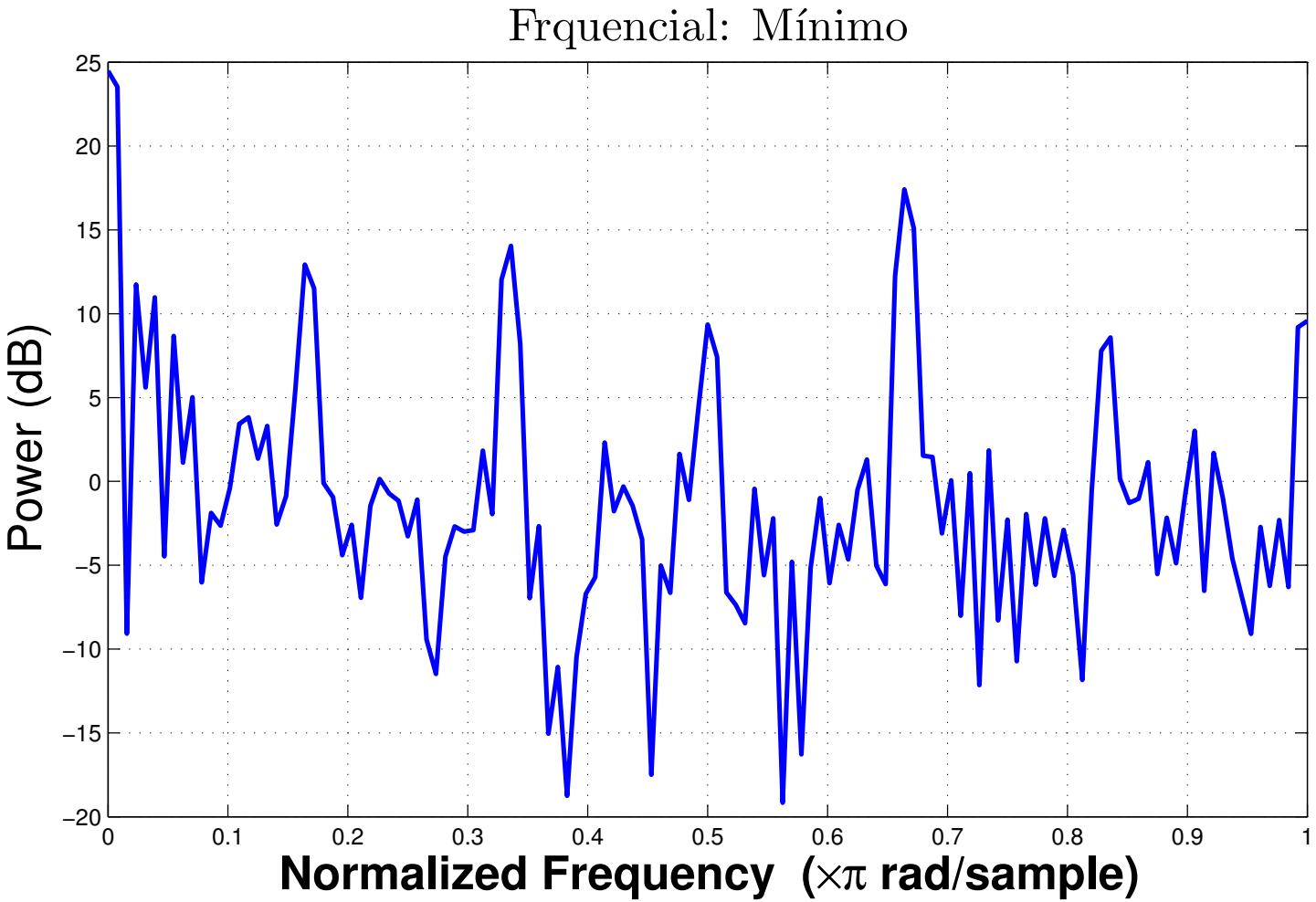
Análise: Visualização no domínio frequencial



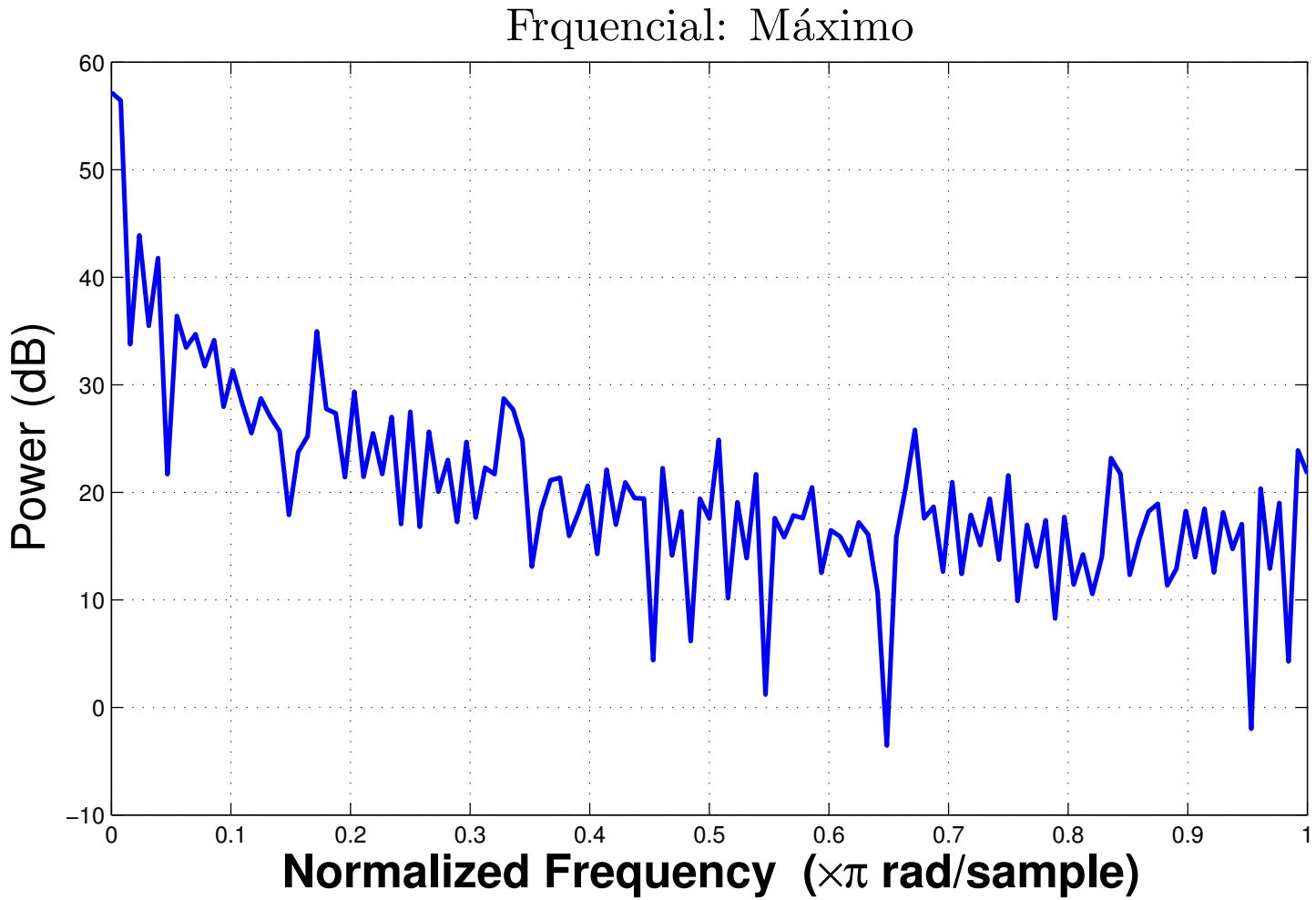
Análise: Visualização no domínio frequencial



Análise: Visualização no domínio frequencial



Análise: Visualização no domínio frequencial

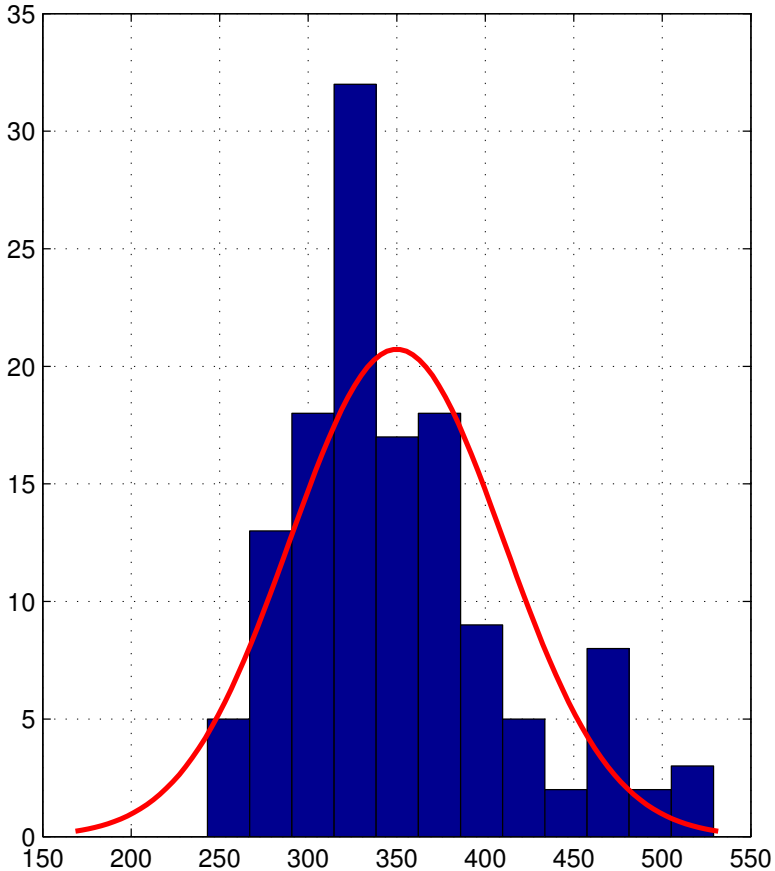


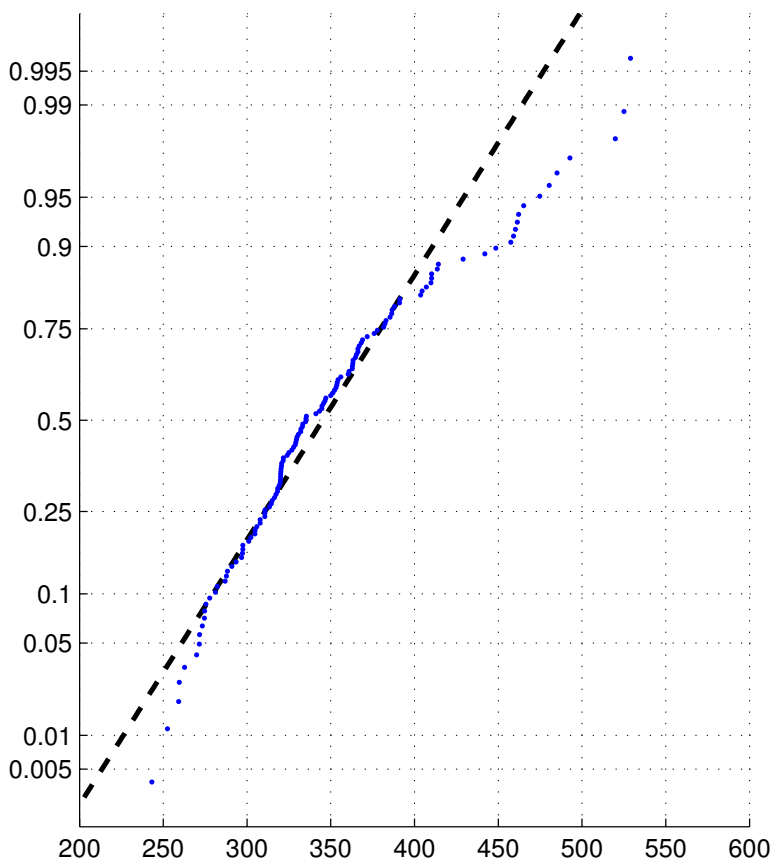
Análise: histogramas

- Identificar grupo de valores que ao serem considerados podem causar distorções (*bias*) na estimativa, diminuindo sua qualidade.
- Conforme primeiras visualizações, o ano de 2013 apresenta valores médios atipicamente altos. Também se observa que os anos de 2012 e 2017 valores mais destoantes.
- Os valores de médias mais altas aparecem de forma saliente nos histogramas e muito acima da curva gaussiana. Já os valores mais baixos, como os valores dos anos de 2016 e 2017 estão mais próximos da curva gaussiana.

Análise: histogramas

Hist. e Prob.: Média





Análise: Conclusões

- Não há padrões anuais dos quais pode-se valer para melhorar a estimativa.
- O padrão mensal é significativo e pode ser usado para se obter uma melhor estimativa.
- Reforçado pela análise frequencial, uma estimativa usando a média seria o mais indicado.
- As amostras dos anos de 2012 e 2013 adicionam significativa distorção e quanto aos anos de 2016 e 2017 não.
- Há uma relação de tendência nos pares 2012/2013 e 2016/2017.

Solução: base teórica

Basicamente, pode-se utilizar a noção de valor esperado. Para o caso discreto, considerando o valor de alvo mensal como uma variável aleatória X, a esperança dessa variável é obtida com [2]:

$$E[X] = \sum_{i=1}^{N} x_i p(x_i) = \mu_X \tag{1}$$

sendo x_i todos os i valores possíveis e sua respectiva probabilidade $p(\cdot)$, e o valor médio μ_X. Como o modelo utilizado é uma distribuição normal (gaussiana), a melhor expectativa para o valor de alvo mensal é dada pela média simples

$$\mu_X = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

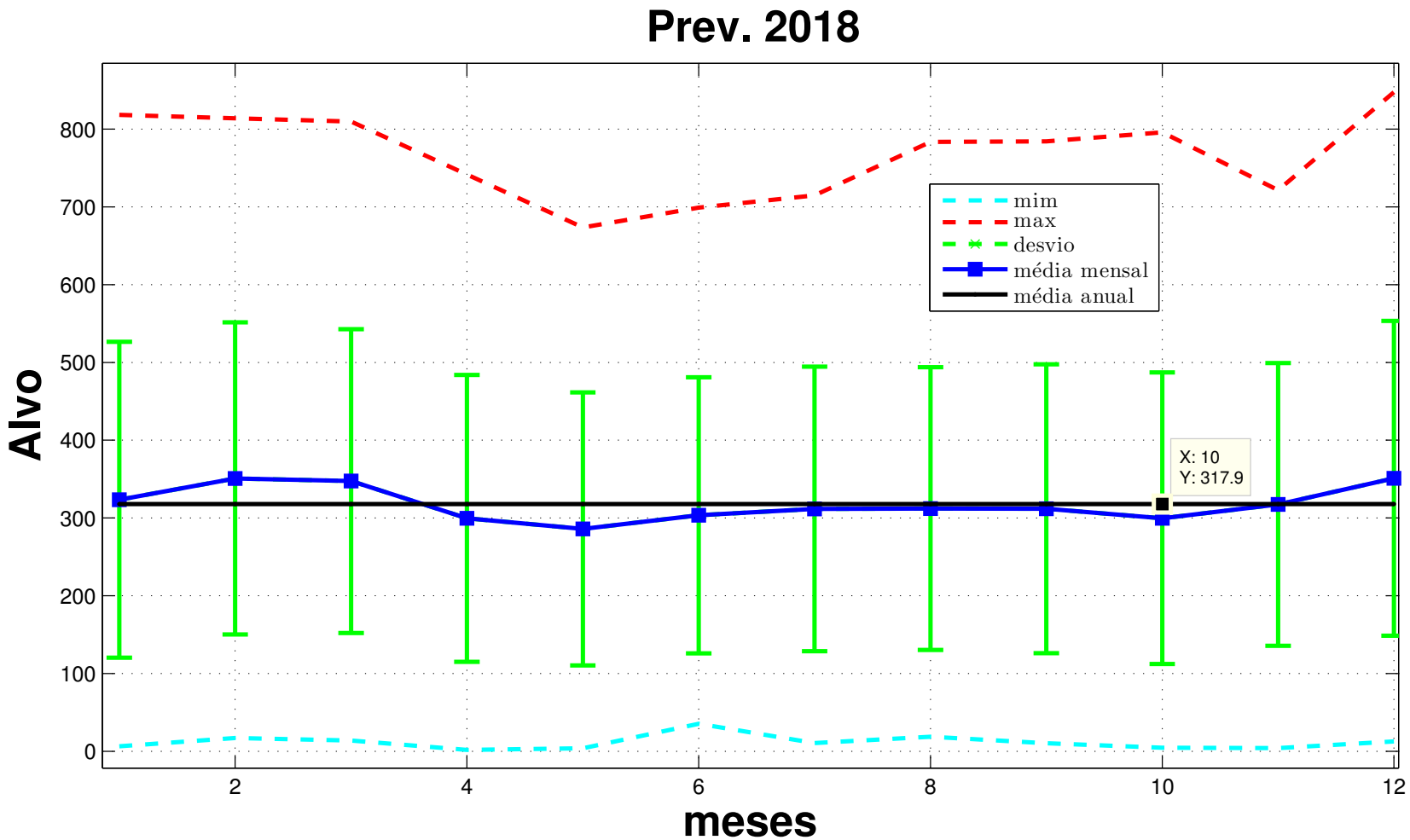
com o desvio padrão obtido com

$$\sigma_X = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_X)^2} \tag{3}$$

Solução: proposta

- Como resposta ao desafio, calcula-se as médias mensais de todos os anos da série histórica, exceto dos meses dos anos 2012 e 2013, como uma estimativa para os meses de 2018. Sendo a estimativa desse ano a média das médias mensais.

Solução: resultados

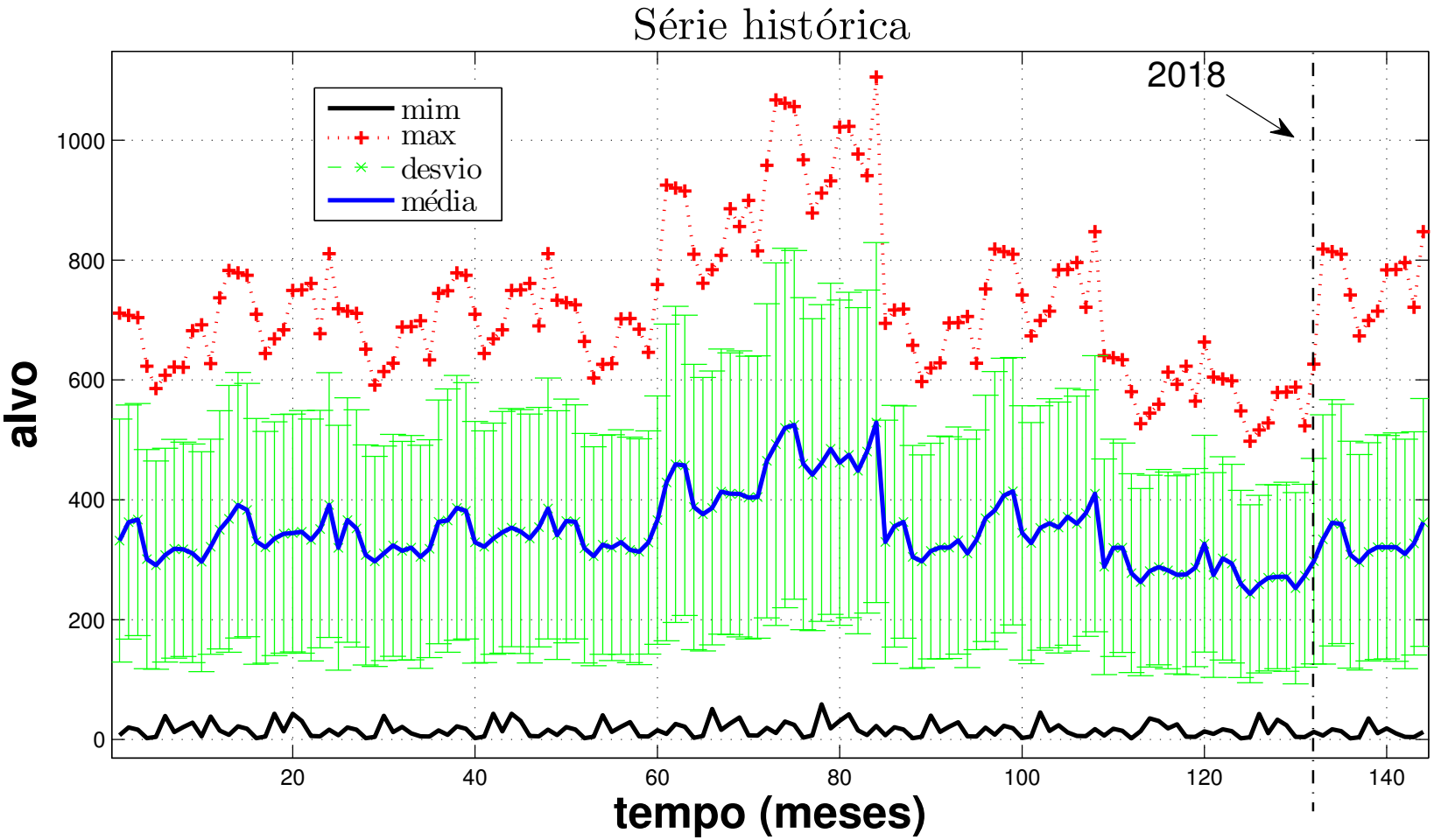


Solução: resultados

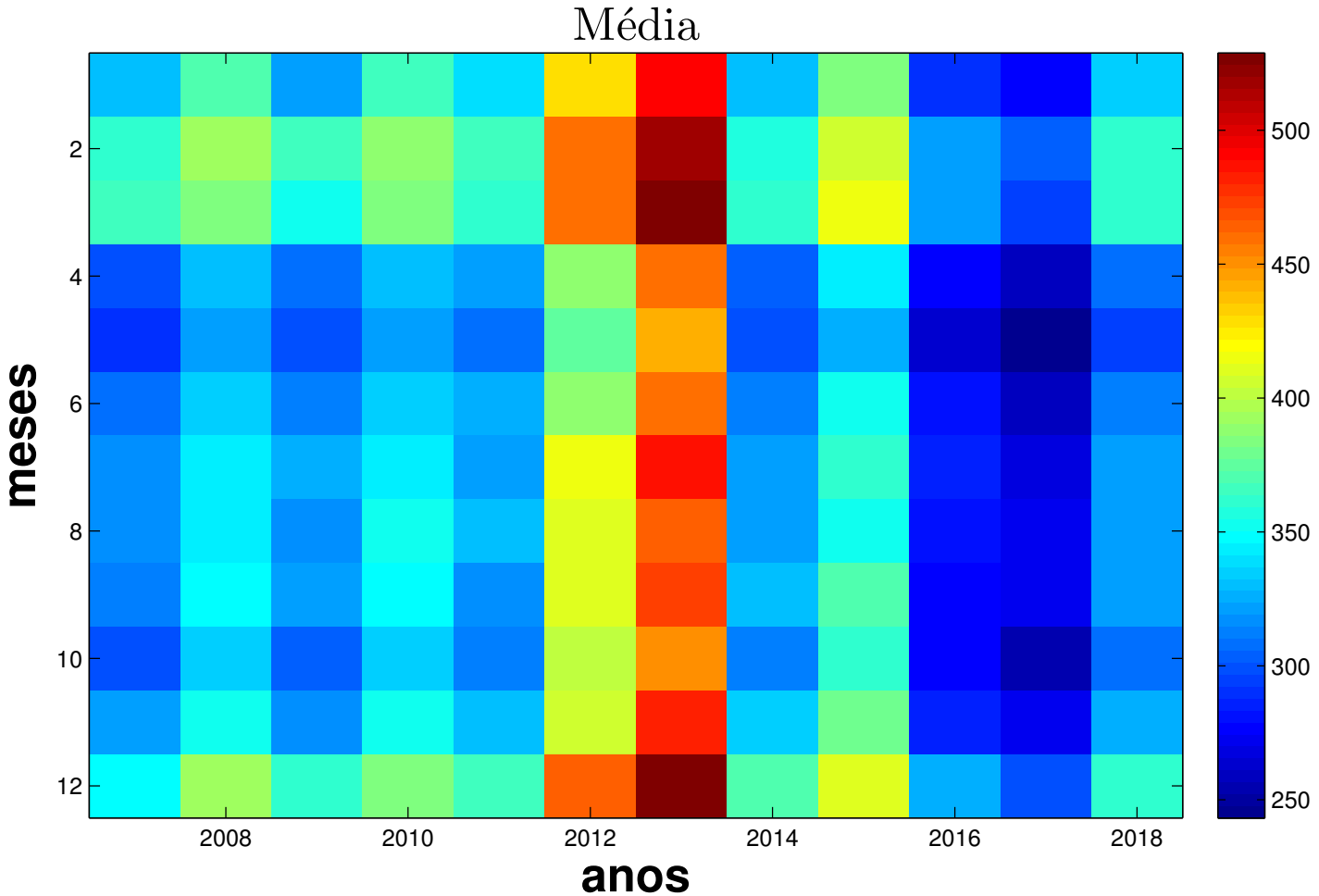
Tabela: estimativa mensal para 2018

Mês.	estimativa (min/max)
Janeiro	333,97 ± 207,91 (6,4651 / 818,33)
Fevereiro	361,64 ± 205,51 (17,05 / 814,00)
Março	359,70 ± 199,61 (13,88 / 809,72)
Abril	308,43 ± 189,20 (1,84 / 741,75)
Maio	295,85 ± 179,84 (3,83 / 673,59)
Junho	313,20 ± 182,04 (35,27 / 698,98)
Julho	321,03 ± 187,44 (10,54 / 714,81)
Agosto	320,92 ± 186,30 (18,74 / 783,57)
Setembro	321,06 ± 190,52 (10,27 / 784,36)
Outubro	309,65 ± 191,39 (4,51 / 795,84)
Novembro	327,28 ± 186,24 (4,20 / 721,41)
Dezembro	362,26 ± 206,62 (12,77 / 847,56)

Solução: resultados



Solução: resultados



Conclusão

- Considera-se como respondido o desafio frente a análise dos dados disponíveis, seguida de uma proposta de estimativa com embasamento teórico e em evidências técnicas.
- Devido ao que foi exigido no desafio e pelo princípio da parcimônia, os dados referentes às variáveis não identificadas foram ignorados.

Notas

- Poder-se-ia avaliar os histogramas mensais, evitando a análise frequencial, para se concluir que há um padrão mensal e esse pode ser modelado por uma distribuição normal.
- Numa análise de histograma das 20 variáveis não identificadas, observou-se na sua maioria uma distribuição gaussiana. Por meio de uma análises nos domínios temporal e frequencial, observou-se sinais altamente ruidosos. Também foi observado que 9 dessas variáveis apresenta valores dispares no ano de 2007.
- Quanto ao fato de 9 variáveis terem valores díspares apenas no ano de 2007 e os valores de Alvo não serem díspares, pode-se concluir que tais variáveis não afetam significativamente o valor Alvo.
- Pode-se concluir, então, que o uso das variável não identificadas para a estimativa traria melhora não significativa ou até nula, não compensando o alto custo computacional e complexidade da aplicação para usá-las.

Obrigado pela atenção e pela oportunidade.


The MathWorks Inc., "Matlab, version 8.3.0.532 (r2014a)," Natick, Massachusetts, 2014.


T. Soderstrom and P. Stoica, *System Identification*, ser. Prentice Hall International Series In Systems And Control Engineering. Prentice Hall, 1989.