

MODELAGEM E PREVISÃO DO NÚMERO DE CASOS DE INTERNAÇÃO POR DOENÇAS DO APARELHO RESPIRATÓRIO NA CIDADE DE SÃO PAULO VIA LSTM

Luís Philipe Craveiro Mendes e Matheus Jun Onishi da Silva

Introdução

Compreende-se que há uma estreita relação entre fatores climáticos e a saúde humana, especialmente no que se refere ao sistema respiratório. Condições climáticas adversas podem aumentar a ocorrência de internações hospitalares e, em casos mais graves, levar a óbitos (Arbex 2012). No Brasil, muitas cidades estão em processo de urbanização ou já apresentam intensa urbanização, o que agrava os problemas de saúde pública devido à formação de ilhas de calor, poluição atmosférica e outros fatores climáticos que contribuem para o aumento de doenças respiratórias (Anenberg 2020).

Na área de saúde pública, prever o número de internações esperadas em um determinado período é essencial para o planejamento de recursos e a formulação de políticas. Além disso, compreender o impacto das variáveis explicativas no modelo é crucial para a identificação de medidas preventivas. Neste trabalho, limitamo-nos ao desenvolvimento de um modelo capaz de prever o número de internações, utilizando uma classe de redes neurais apropriada para séries temporais e utilizando covariáveis climáticas.

O estudo concentra-se na modelagem do número de internações por doenças do sistema respiratório na cidade de São Paulo entre os anos de 2018 e 2019. Foram utilizadas como covariáveis a umidade relativa do ar, a temperatura e a concentração de material particulado fino ($PM_{2.5}$). Os dados de internações foram obtidos no DATASUS, enquanto as informações climáticas foram extraídas do Sistema Integrado de Serviços Ambientais (SISAM). Para a modelagem, foi empregado o modelo de redes neurais de Memória Longa de Curto Prazo (LSTM) (Sepp Hochreiter 1997), devido à sua capacidade de capturar padrões temporais complexos nos dados.

Materiais e Métodos

Memória Longa de Curto Prazo (LSTM)

O LSTM é um tipo de rede neural da família de Redes Neurais Recorrentes que funcionam como uma espécie de diversos “mensageiros” que processam os dados que receberam, e transmitem isso para outros mensageiros. Isso cria uma espécie de memória que é compartilhada entre eles e pode fornecer bons resultados como saídas. Essa classe de redes neurais é utilizada em diversos contextos como trabalhos de tradução, reconhecimento de voz etc. (“Understanding LSTM Networks,” n.d.).

Redes Neurais Recorrentes (RNNs) conseguem realizar boas previsões considerando o impacto de informações recentes, entretanto, quando há informações importantes disponíveis no histórico dos dados que impactam em previsões futuras em intervalos cada vez maiores, essa classe de redes neurais se torna ineficiente na tarefa de realizar previsões, visto que não conseguem resgatar esse termo longínquo que impacta a previsão desejada com eficiência. Ou seja, quanto maior o *lag* entre a informação que deseja ser prevista e a informação relevante para tal previsão, menos capaz essa classe de redes realizar previsões precisas. (Bengio 1994) explora essas ideias com mais detalhes. Porém, para o LSTM, o problema de resgatar dados históricos longínquos é solucionado.

O principal objetivo do LSTM é justamente conseguir manter a influência de dados passados em estimativas futuras. Diferentemente de redes neurais recorrentes padrões, ele tem uma estrutura mais complexa. Veja abaixo a ilustração de uma RNN padrão com a única função de ativação tangente hiperbólica:

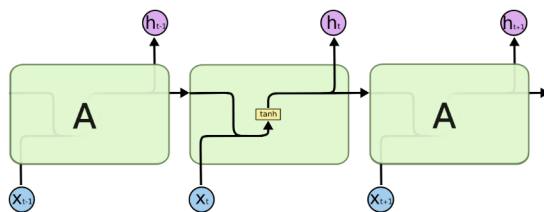


Figure 1: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Agora veja a estrutura de um módulo do LSTM que será abordado com mais detalhes a diante:

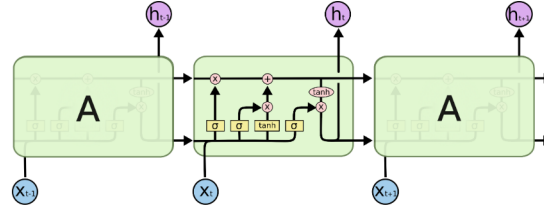


Figure 2: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

A Interconexão das Informações

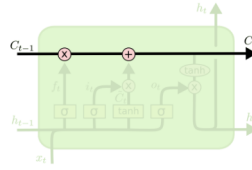


Figure 3: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Note, pela primeira figura acima considerando C_{t-1} uma informação que está passando por um módulo, que C_{t-1} passando por meio de transformações lineares. Primeiramente ele é multiplicado por uma saída de uma função de ativação sigmóide, ou seja, tal informação ganha um peso considerando valores entre 0 e 1, já que estes valores são a imagem da função sigmóide. Em seguida, C_{t-1} passa por mais duas operações que veremos com mais detalhes futuramente. Por enquanto, foquemos nessa primeira operação.

A partir do que foi discutido, podemos entender que a função sigmóide filtra parte da informação contida em C_{t-1} ao multiplicar por valores entre 0 e 1. Veja abaixo o destaque para essa operação inicial:

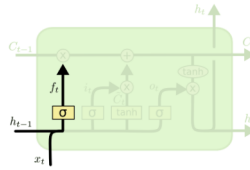


Figure 4: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Considerando isso, com base nos valores h_{t-1} e x_t , ele decide o quanto de informação de C_{t-1} irá passar por meio da função sigmóide, isto é, $f_t = \sigma(w_f[h_{t-1}, x_t] + \theta_f)$, o qual w_f é um peso sináptico e θ_f é um viés. Após esse passo de filtro de quanta informação é útil, temos

o segundo passo que é o processamento de uma nova informação que deve ser acrescentada a C_{t-1} .

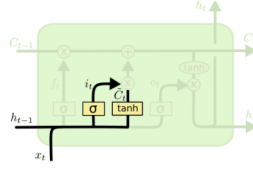


Figure 5: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Esse novo processamento é feito por meio da função de ativação hiperbólica, isto é, $\tilde{C}_t = \tanh(w_C[h_{t-1}, x_t] + \theta_C)$, com peso sináptico e viés, respectivamente, w_C e θ_C que é, assim como C_{t-1} , ponderada por meio da saída da função sigmóide $i_t = \sigma(w_i[h_{t-1}, x_t] + \theta_i)$, com peso sináptico e viés, respectivamente, w_i e θ_i que decide o quanto dessa nova informação será acrescentada a antiga ponderada $C_{t-1}f_t$. Após todo esse processo, ambas informações são somadas, gerando a nova informação $C_t = f_t C_{t-1} + i_t \tilde{C}_t$ que está ilustrada abaixo.

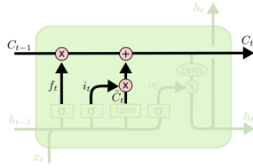


Figure 6: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

A partir da obtenção de C_t , temos tudo que precisamos para emitir a saída final, ela será o C_t filtrado. Primeiramente, C_t passa por uma função sigmóide a qual decide que informações estarão disponíveis na saída, ou seja, há um filtro do que é menos e mais provável, já que a imagem desta função é para valores entre 0 e 1. Ao mesmo tempo, o mesmo C_t passa por uma função tangente hiperbólica que tem imagem nos valores de -1 a 1. A saída da função tangente hiperbólica é multiplicada pela saída da função sigmóide, consequentemente a informação dada pela função tangente hiperbólica é filtrada. Esse processo final é ilustrado na figura abaixo.

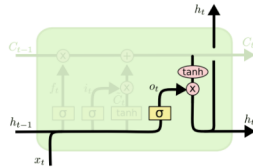


Figure 7: Fonte: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Tal saída é disposta como uma resposta do modelo para um dado ponto e também é replicada para o módulo seguinte onde o processo irá se repetir para o tempo $t + 1$.

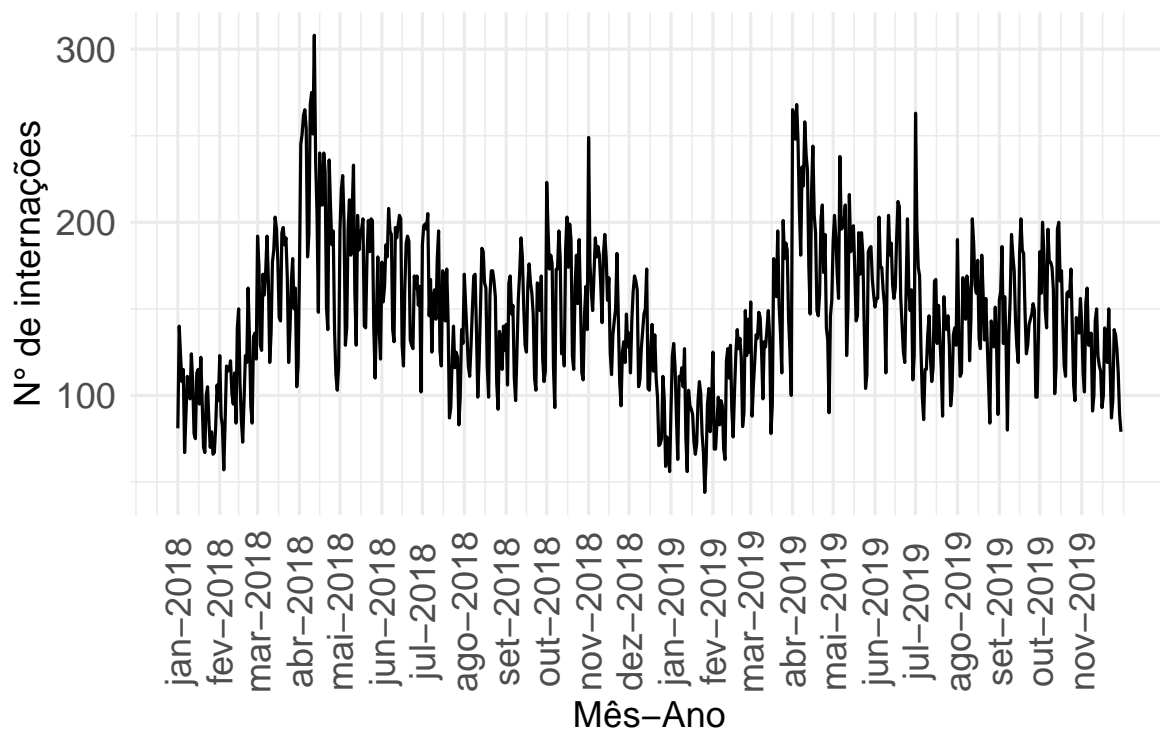
Resultados

Nesta aplicação, dedicamo-nos a apresentar o efeito da poluição, mais especificamente o efeito do material particulado fino ($PM_{2.5}$), juntamente com a umidade relativa do ar e a temperatura, no número de internações de pessoas com doenças do aparelho respiratório na cidade de São Paulo - SP, no período de 01/01/2018 a 30/11/2019. Os dados ambientais foram coletados por meio do Sistema Integrado de Serviços Ambientais (SISAM) e os dados sobre internações por doenças respiratórias foram obtidos por meio do DATASUS.

Ademais, na base de dados, o $PM_{2.5}$ era registrado a cada 12 horas em 2019 por meio das estações meteorológicas, enquanto a umidade e a temperatura, outras covariáveis utilizadas no modelo, eram registradas a cada 6 horas bem como o $PM_{2.5}$, mas somente em 2018. Considerando isso, calculamos a média diária de cada uma delas.

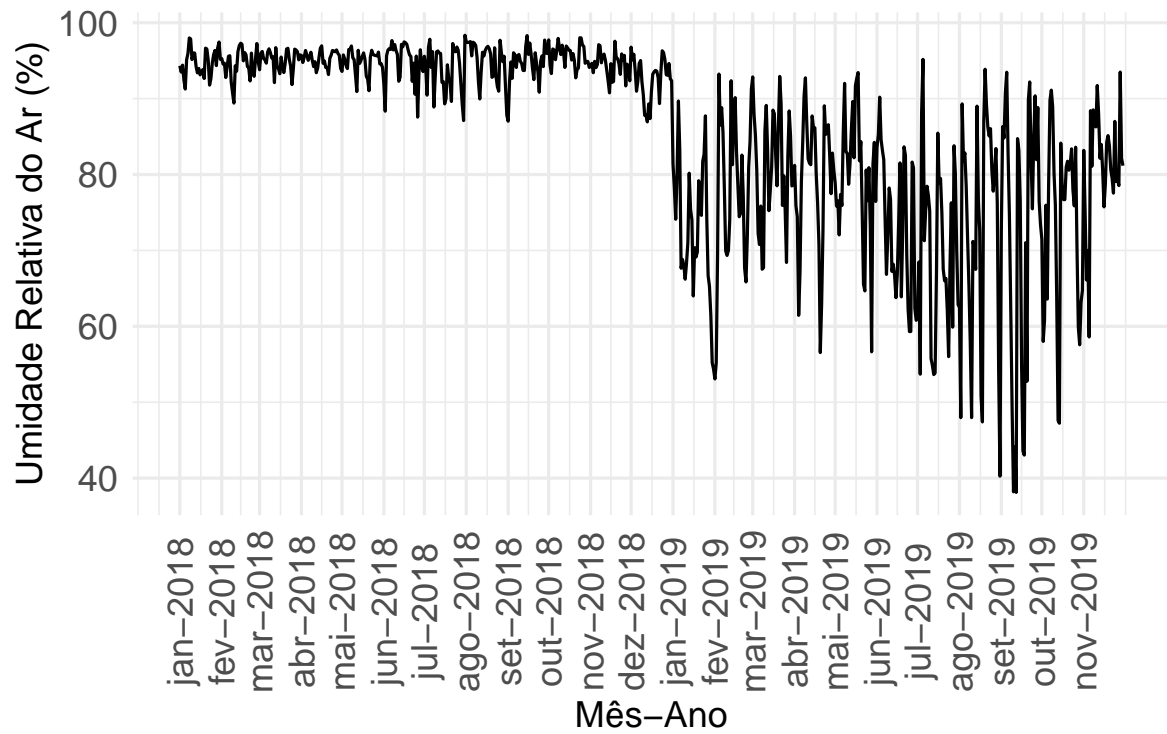
Análise Descritiva

Abaixo pode ser visto o comportamento do número de internações de pessoas com doenças do aparelho respiratório.



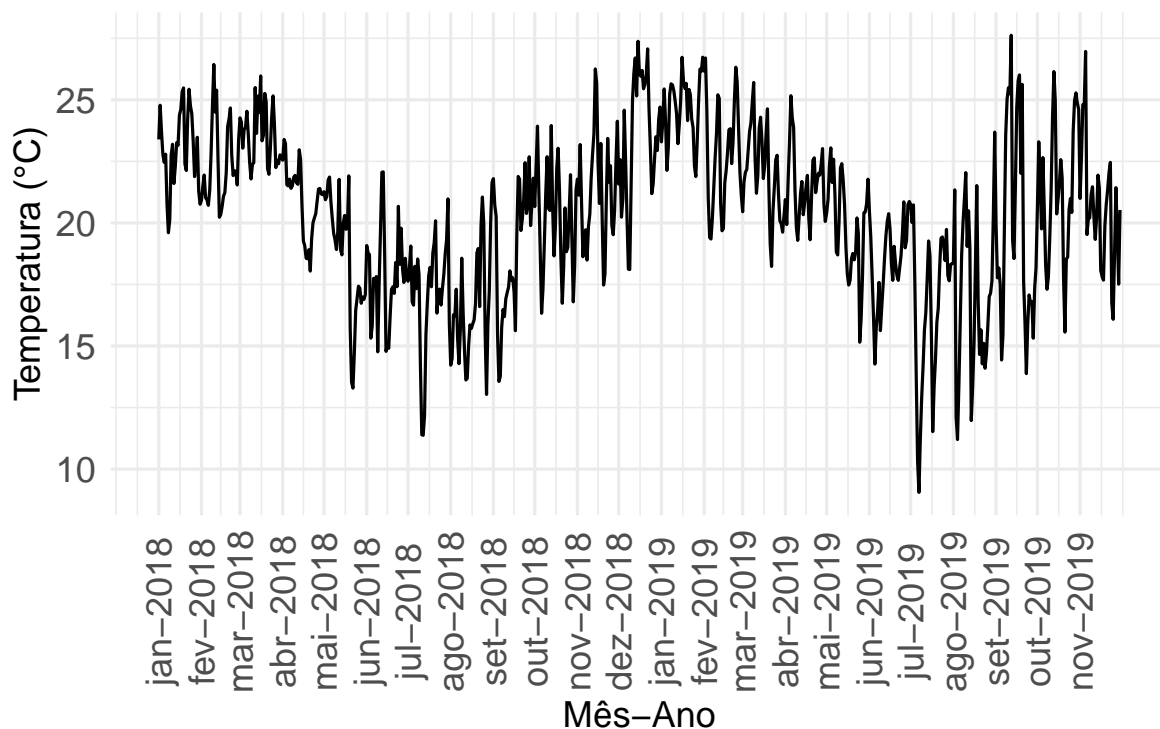
Nota-se que os períodos de maiores concentrações dos poluentes estão nos meios dos anos.

Abaixo temos a umidade relativa do ar:



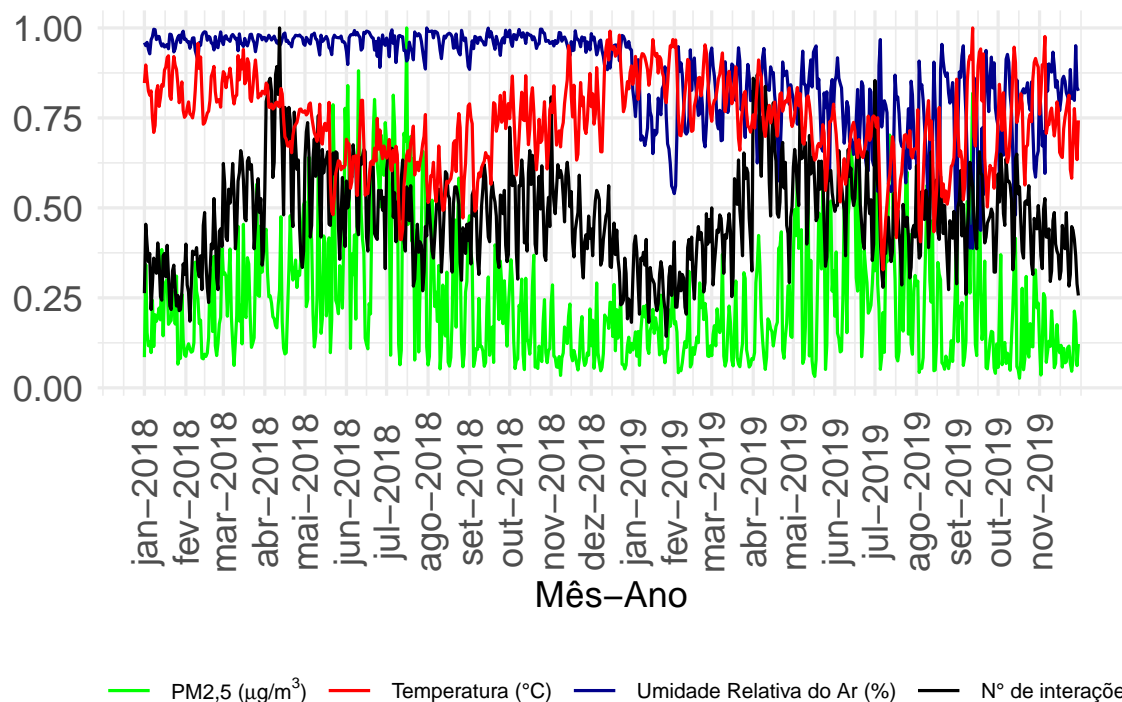
Com altas porcentagens principalmente no ano de 2018, com um comportamento atípico.

Já para a temperatura média diária, temos:



Curiosamente níveis intermediários de temperatura parece ter uma relação positiva com o número de internações, pois parece que há um aumento dos números de internações nesses períodos sem temperaturas mais extremas.

Por fim, podemos visualizar essas variáveis em um mesmo gráfico, onde as séries estão padronizadas pelos seus máximos, dessa forma elas ficam no intervalo $[0,1]$:



Realmente parece ter uma relação entre temperaturas intermediárias e o número de interações, o gráfico acima torna isso mais evidente.

#Conclusão

Referências

- Anenberg, Susan C. 2020. “Synergistic Health Effects of Air Pollution, Temperature, and Pollen Exposure: A Systematic Review of Epidemiological Evidence.” *Environmental Health* 19 (1). <https://doi.org/10.1186/s12940-020-00681-z>.
- Arbex, Marcos Abdo. 2012. “A Poluição Do Ar e o Sistema Respiratório.” *Jornal Brasileiro de Pneumologia* 38 (5): 643–55. <https://doi.org/10.1590/s1806-37132012000500015>.
- Bengio, Y. 1994. “Learning Long-Term Dependencies with Gradient Descent Is Difficult.” *IEEE Transactions on Neural Networks* 5 (2): 157–66. <https://doi.org/10.1109/72.279181>.
- Sepp Hochreiter, Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- “Understanding LSTM Networks.” n.d. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.