Séries Temporais: Um Estudo Sobre A Predição de Medidores Populacionais Brasileiros.

Lucas Santiago de Oliveira[1], **Rafael Amauri Diniz Augusto**[1], **Thiago Henriques Nogueira**[1]

[1]Instituto de Informática – Pontifícia Universidade Católica de Minas Gerais (PUCMG)

{lucas.oliveira.1201561,radaugusto,thiago.nogueira}@sga.pucminas.br

Abstract. *Every year, the population of Brazil grows by a given factor, increasing demand for a higher operating budget in cities, states and the country. The factor that denotes said growth is given by a combination of different variables such as access to sanitation, universal healthcare and better infrastructure, which are the foundation for supporting a bigger population. Discovering this growth factor can help a number of government institutions, planning departments and cities in taking into account projected growth when reporting future operating budgets to the federal government of Brazil. This study aims to develop a machine learning model that is able to easily and reliably project said population growth, especially in small towns that don't have enough data to train a multivariable model.*

1. Introdução

O acelerado crescimento populacional e econômico experienciado por inúmeros países no século XX demandou uma expansão da capacidade de processamento e interpretação de informações em diferentes setores da sociedade, com uma das técnicas mais relevantes sendo a análise de séries temporais devido à sua ampla aplicação em diferentes setores da sociedade. Séries temporais são definidas como uma coleção de observações expandidas ao longo de um determinado período de tempo, e sua representação consiste em uma série de pontos indexados e ordenados ao longo de um intervalo constante de tempo [BOX et al. 2015]. Tendo isso em vista, diferentes conjuntos de dados podem ser entendidos e representados como séries temporais: a quantidade de objetos produzidos por uma fábrica, uma representação numérica do número semanal de acidentes em uma estrada, precipitação ao longo de um ano, observações de hora em hora acerca de um processo químico, e diversos outros. Exemplos de usos de séries temporais também podem ser encontrados em setores como finanças, geografia, engenharia, ciências naturais e ciências sociais [S. and GERSHENFELD 1994].

Tendo isto em vista, é natural que séries temporais se tornem cada vez mais presentes no contexto de análise de dados a fim de diversificar o número de técnicas utilizadas e prover alternativas mais adequadas a cada cenário. O presente trabalho visa explorar essa aplicação de séries temporais no contexto da predição do crescimento populacional brasileiro.

A análise de séries temporais é uma categoria de técnicas capazes de produzir um modelo que leva em consideração dados passados da série para fazer uma predição do valor de um dado desconhecido. Estas técnicas são úteis à medida que estimar valores relevantes em determinados contextos pode se provar imperativo para tomadas de decisões bem-sucedidas, com essas previsões podendo ser aplicadas nos mais diferentes contextos

e cenários, como explicitado anteriormente. Adicionalmente, como os modelos prevêem novos valores utilizando como base valores passados, é importante que haja coerência na ordem de observação a fim de possibilitar que o modelo note a presença de características e padrões na série. Esta característica pode ser aproveitada para mapear e mensurar diferentes métricas que estão ordenadas ao longo de um período de tempo. Com isto em mente, o presente artigo tem como objetivo utilizar a análise de séries temporais para avaliar a evolução de diferentes medidores populacionais brasileiros de 1960 a 2020, levando em conta dados históricos.

2. Trabalhos Correlatos

Historicamente, modelos lineares têm dominado previsões de séries temporais pela sua simplicidade de implementação, baixo custo computacional e facilidade para entender seu funcionamento, o que faz com que eles sejam bem conhecidos e efetivos na resolução de uma ampla gama de problemas, como pode ser visto em diferentes estudos que envolvem crescimento populacional [Green and Sparks 1999]. Apesar disso, contextos específicos trazem uma demanda por ferramentas mais robustas, baseadas em diferentes algoritmos. Dentro das técnicas mais modernas com *Deep Learning*, são notáveis o uso de *Convolutional Neural Networks* [Ban et al. 2020] e *Long Short-Term Memory Networks* [Lindemann et al. 2021] aplicadas a séries temporais. Esses modelos são originalmente voltados a análises multivariadas com bases de dados massivas, o que faz com que as suas características mais notáveis sejam a alta resistência a ruídos na base de dados e a habilidade de aprender e extrair automaticamente as características principais de uma série. Apesar de apresentarem grandes vantagens, para funcionar corretamente e com alto grau de confiabilidade, também é demandada uma base de dados extensiva, além do custo computacional ser muito maior. Essa característica é levada em consideração em diferentes contextos, inclusive em problemas que não possuem uma base de dados extensa o bastante. Este ponto faz com que o presente trabalho busque evitar tal barreira, visto que nem sempre cidades e entidades governamentais possuem tal base de dados capaz de suportar um modelo de *Deep Learning*. Com isso em mente, o presente trabalho apresenta uma busca por um modelo de *Machine Learning* capaz de realizar predições de medidores populacionais apenas com dados passados, em forma de séries temporais, já que é entendido que para previsões de médio prazo, que são previsões de 10 anos no futuro, não são necessárias tais detalhamentos.

3. Metodologia

A primeira etapa do presente trabalho consistiu na obtenção de um conjunto de dados disponibilizados pelo Banco Mundial [Group 2022], que é composto por uma coletânea de diferentes valores para métricas sociais, socioeconômicas e populacionais brasileiras, coletadas entre 1960 e 2020.

Foram escolhidos três indicadores que representam diferentes aspectos do desenvolvimento populacional brasileiro e que são adequados para modelagens de séries temporais. Isto é, indicadores que apresentam um padrão de crescimento característico, consistente e previsível. Os indicadores escolhidos são “SP.POP.TOTL”, que representa a população total do Brasil em milhões de pessoas, “SP.RUR.TOTL.ZS”, que representa a porcentagem da população brasileira que vive em áreas rurais e “SP.URB.TOTL.IN.ZS”, que representa a porcentagem da população brasileira que vive em centros urbanos. O

objetivo do grupo ao escolher os indicadores explicitados é demonstrar a aplicação de séries temporais em dados populacionais e a modelagem de um preditor capaz de projetar um crescimento realista dentro de um período de dez anos.

Para este fim, o conjunto de dados foi separado em conjuntos de treino, validação e teste, com o objetivo de predizer os valores para os indicadores escolhidos para dez anos no futuro com alto grau de confiabilidade. Para isso, foi estabelecida uma taxa de 83% dos dados sendo usado para treinamento do modelo e 2% dos dados sendo usados para o conjunto de validação, com 15% dos dados sobrando para o conjunto de teste do modelo. A escolha por trás dessa alta taxa de treinamento foi motivada principalmente pelos valores observados no crescimento populacional brasileiro serem afetados por incontáveis fatores como políticas públicas, contextos históricos, fenômenos sociais e desenvolvimentos no âmbito socioeconômico brasileiro ao longo de anos, e por isso se faz imperativo treinar o preditor com dados mais recentes possíveis a fim de suprimir o ruído causado por todas estas variáveis.

Os dados fornecidos para cada indicador são medidos em unidades diferentes, relativas ao indicador que está sendo analisado. Na tabela 1 segue um exemplo de valores para seis diferentes anos no banco de dados envolvendo os indicadores descritos.

Ano	SP.POP.TOTL	SP.RUR.TOTL.ZS	SP.URB.TOTL.IN.ZS
1960	72.179235	53861	46139
1961	74.311338	52878	47122
1962	76.514329	51901	48099
1963	78.772647	50921	49078
1964	81.064572	49941	50059
1965	83.373533	48963	51037

Tabela 1. Amostra de valores dos primeiros 6 anos para os indicadores selecionados

O presente trabalho analisa os dados obtidos como séries temporais, uma vez que eles se encaixam na definição encontrada em "Time Series Analysis: Forecasting and Control" [BOX et al. 2015]. Neste estudo a ordem de observação dos dados é de extrema relevância, pois ela aponta para características e padrões na série e permite com que o preditor consiga modelar a evolução dos valores naquela série.

Também conhecida como suavização exponencial, o método de Holt é utilizado para calcular previsões em séries temporais que apresentam tendência [WINTERS 1960]. Para se prever valores em Y a partir de um conjunto de dados com o método de suavização exponencial de Holt, é necessário saber a tendência e o nivelamento da série temporal gerada. Essa expressão é dada por:

$$F(t) = L(t) + T(t) + R \tag{1}$$

Figura 1. Expressão que rege a série temporal

onde $L(t)$ é o nivelamento da série, $T(t)$ representa a inclinação da linha na qual os dados estão distribuídos e R é o ruído presente nos dados. O nível da série temporal,

ou $L(t)$, representa o valor em Y da série temporal em um instante T. Para estimar esse valor, é necessário utilizar uma Equação de Atualização de Nível, que é expressada por:

$$L(t) = \alpha * (\frac{Y(t)}{S(t)}) + (1 - \alpha) * (L(t-1) + T(t-1)) \quad (2)$$

Figura 2. Equação de nível da série

Na figura 2, é escolhido um valor para alpha (α), o que afeta a representatividade dos valores de níveis-base passados. Valores próximos a um reduzem o peso de valores muito antigos, e valores próximos a zero dão o mesmo peso para todos os valores presentes na série. Dessa forma, o modelo aprende qual será o nível referente a um novo ano que foi inserido na série. Também é importante notar que a série apresenta tendência, ou $T(t)$, que é entendida como a angulação da linha da série temporal [WIENER 1949]. Essa tendência foi identificada como sendo de tendência aditiva, visto que os dados se aproximam de uma evolução linear.

A segunda parte do estudo consistiu na obtenção dos dados temporais para os indicadores escolhidos e a passagem dessa série para o modelo de Holt utilizando o valor 1.4 para alpha a fim de dar um peso maior para valores mais recentes da série, já que é desejável que valores observados mais recentemente ditem o sentido de evolução do modelo. O valor 1.4 foi descoberto como valor ideal após ajustar o modelo para ele prever bem o conjunto de dados de validação.

Os valores previstos para cada ano de teste, que consiste nos anos de 2010 a 2020, foram em seguida armazenados e modelados com a biblioteca matplotlib para visualização e comparação com os valores reais observados para este período. Para a etapa de avaliação do modelo, foram escolhidas as métricas de avaliação *MAE* (*Mean Absolute Error*) e o *R2 Score*. Tendo em vista que os indicadores sociais escolhidos são facilmente modeláveis por modelos lineares e não apresentam características que dificultam o aprendizado de séries temporais como sazonalidade e trends excêntricas, uma variância aceitável é definida pelo grupo na tabela 2.

Indicador	Valor ideal para *MAE*	Valor ideal para *R2*
SP.POP.TOTL	8	0.75
SP.RUR.TOTL.ZS	4	0.7
SP.URB.TOTL.IN.ZS	4	0.7

Tabela 2. Resultados considerados ótimos para cada métrica e para cada indicador populacional

4. Experimentos

Ao final da etapa de previsões foi gerada uma imagem para cada indicador contendo os valores previstos para o período de teste, bem como os valores reais referentes aos períodos de treinamento, validação e teste, vide as figuras 1, 2 e 3. Como extensão dessa etapa, o modelo também foi avaliado de acordo com as métricas de avaliação definidas pela metodologia do estudo, vide as tabelas 3, 4, 5.

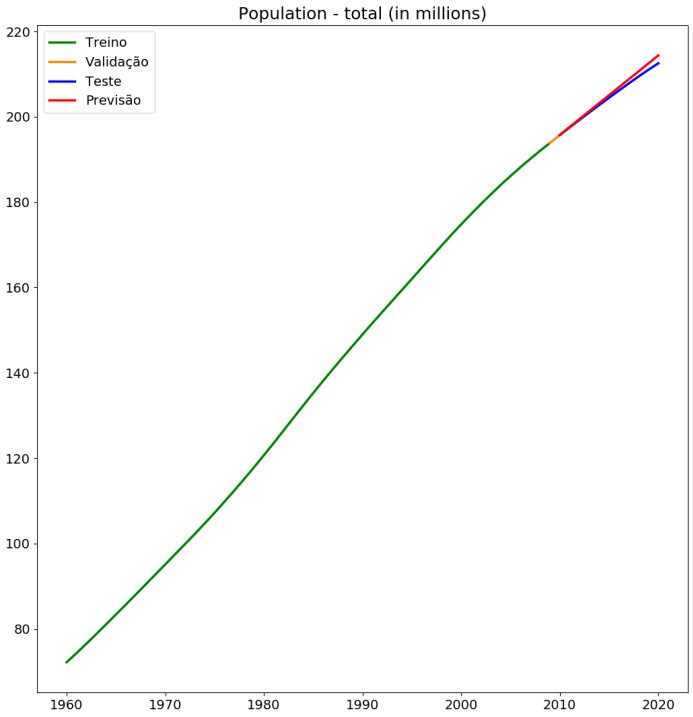

Figura 3. Modelagem da previsão para o indicador "SP.POP.TOTL"

Na figura 3 é apresentada a modelagem da previsão para o indicador "SP.POP.TOTL", que apresenta a população brasileira ao longo dos anos, em milhões de pessoas. A tabela 3 apresenta os valores obtidos para as métricas de avaliação do modelo. A partir do exposto, é possível ver como a modelagem performou dentro do objetivo estabelecido nas duas métricas, e apresentou um desvio negligenciável.

Métrica de avaliação para a previsão de "*Population - total (in millions)*"	Valor
MAE Score	0.7305939401983349
R2 Score	0.9703244706047911

Tabela 3. Avaliação da performance do modelo no conjunto de teste com as métricas *MAE* e *R2*

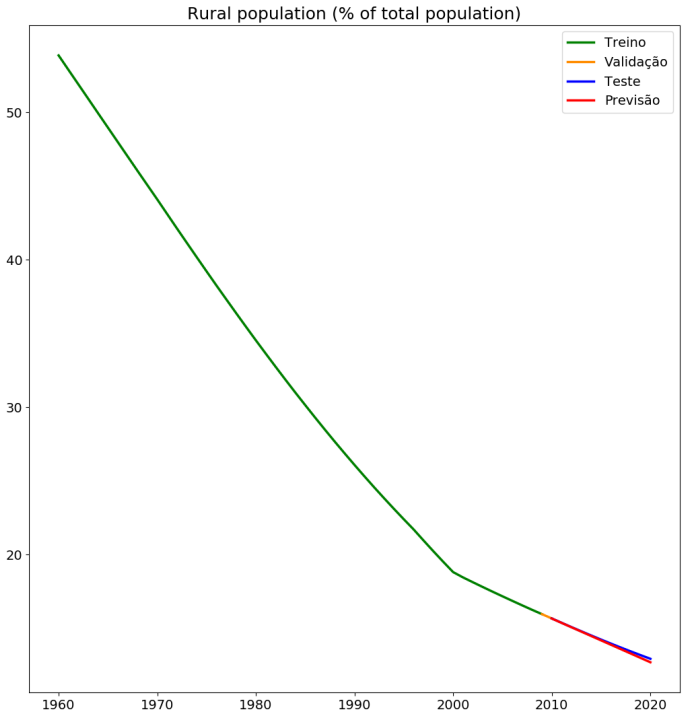

Figura 4. Modelagem da previsão para o indicador "SP.RUR.TOTL.ZS"

Na figura 4 é vista a modelagem da previsão para o indicador "SP.RUR.TOTL.ZS", que apresenta a porcentagem da população brasileira que vive em áreas rurais ao longo dos anos. A tabela 4 apresenta os valores obtidos para as métricas de avaliação do modelo em relação a esse indicador.

Métrica de avaliação para a previsão de *"Rural population (% of total population)"*	Valor
MAE Score	0.0829616118683835
R2 Score	0.9832934099951631

Tabela 4. Avaliação da performance do modelo no conjunto de teste com as métricas *MAE* e *R2*

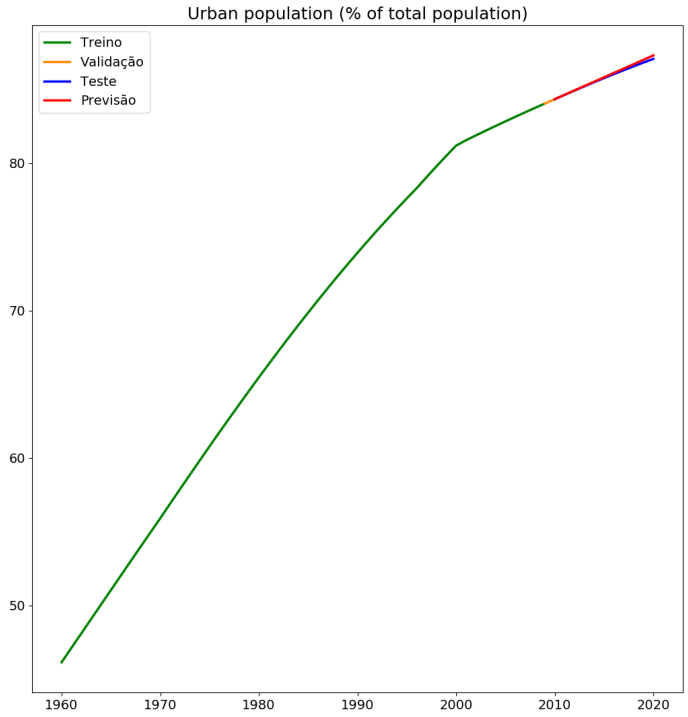

Figura 5. Modelagem da previsão para o indicador "SP.URB.TOTL.IN.ZS"

Na figura 5 é vista a modelagem da previsão para o indicador "SP.URB.TOTL.IN.ZS", que apresenta a porcentagem da população brasileira que vive em centros urbanos ao longo dos anos. A tabela 5 apresenta os valores obtidos para as métricas de avaliação do modelo em relação a esse indicador.

Métrica de avaliação para a previsão de Urban population "(*% of total population*)"	Valor
MAE Score	0.08296161186839222
R2 Score	0.9832934099951601

Tabela 5. Avaliação da performance do modelo no conjunto de teste com as métricas *MAE* e *R2*

5. Discussão

É relevante notar como as previsões para todos os indicadores apresentaram um resultado de variância dentro do que foi definido como objetivo pelo grupo. Como exemplo, na figura 5 pode ser vista a previsão para o indicador 3 (SP.URB.TOTL.IN.ZS), que se mostrou muito bom para ser previsto com séries temporais, pois, mesmo tendo um ponto de amenização que diminui a taxa de crescimento, essa amenização acontece na década de 1990, o que ainda está dentro do período de treinamento, ou seja, o modelo consegue se alterar para levar essa mudança em conta na hora de prever valores futuros para esse indi-

cador. A performance do modelo neste indicador também foi considerada extremamente satisfatória para o grupo. Os bons resultados vistos para este indicador também podem ser vistos no indicador 2, que também apresenta uma amenização na mesma época, o que faz sentido, visto que os indicadores 2 e 3 representam dois opostos: A população que vive em centros urbanos e a população que vive em zonas rurais, logo, é esperado que os dois apresentem a mesma taxa de crescimento, seja ele positivo ou negativo.

Da mesma forma, na figura 3 pode ser visto que o crescimento populacional brasileiro segue a mesma taxa de crescimento desde 1960, o que é suportado por cada vez mais pessoas viverem em centros urbanos, tendo mais acesso a condições melhores de vida, à saúde, saneamento básico e infraestrutura. Esse suporte social garante que a população continue crescendo e que a taxa de mortalidade infantil abaixe, o que explica essa taxa estar em constante crescimento.

6. Conclusão

Ao final do projeto, foram obtidas previsões para três diferentes indicadores sociais brasileiros nos anos de 2011 a 2020 com pouco nível de desvio, o que implica na possibilidade do uso do modelo desenvolvido para prever os valores populacionais futuros no Brasil e auxiliar a tomada de decisões governamentais em nível federal, estadual e municipal. É relevante notar como as três métricas apresentaram excelentes resultados para predição, o que é esperado quando a evolução futura dos dados segue o mesmo padrão de crescimento que foi apresentado durante a fase de treinamento do modelo.

Apesar de excelentes resultados, é importante relembrar que existem mais fatores que afetam a evolução dos indicadores avaliados neste estudo do que apenas os seus valores passados e uma tendência de evolução. Políticas públicas, outros indicadores sociais e ruídos inexplicáveis devem ser levados em conta, e por este motivo o grupo recomenda que para trabalhos futuros que foquem no aspecto de uma evolução de longo prazo, como 40 anos ou mais, esse problema seja tratado como uma evolução multivariada com séries temporais, levando em conta outros indicadores e como eles afetam os indicadores analisados aqui.

Referências

[Ban et al. 2020] Ban, Y., Zhang, P., and A, N. (2020). Near real-time wildfire progression monitoring with sentinel-1 sar time series and deep learning. Technical report, Sci Rep 10, 1322, https://doi.org/10.1038/s41598-019-56967-x.

[BOX et al. 2015] BOX, G., JENKINS, G., REINSEL, G., and LJUNG, G. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley and Sons Inc., 5th edition.

[Green and Sparks 1999] Green, A. G. and Sparks, G. R. (1999). *Population growth and the dynamics of Canadian development: a multivariate time series approach*. Economic History Working Papers 22414, London School of Economics and Political Science, Department of Economic History.

[Group 2022] Group, W. B. (2022). *Health Nutrition and Population Statistics*. HealthStats.

[Lindemann et al. 2021] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., and Weyrich, M. (2021). *A survey on long short-term memory networks for time series prediction*. Procedia CIRP, volume 99, pages 650-655, issn 2212-8271 edition.

[S. and GERSHENFELD 1994] S., W. A. and GERSHENFELD, N. A. (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Taylor & Francis Inc, 1st edition.

[WIENER 1949] WIENER, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press.

[WINTERS 1960] WINTERS, P. R. (1960). *Forecasting Sales by Exponentially Weighted Moving Averages*. INFORMS.