Revisão de Inteligência Artificial (2022) 55:1253–1287
https://doi.org/10.1007/s10462-021-09991-1

Previsão de séries temporais altamente granulares e de curto prazo de PM2,5 poluente atmosférico - uma revisão comparativa

Rituparna Das1 · Asif Iqbal Middya1 · Sarbani Roy1

Publicado on-line: 29 de março de 2021

Resumo

A previsão de séries temporais adquiriu imensa importância de pesquisa e possui vastas aplicações na área de monitoramento da poluição do ar. Este trabalho tenta investigar as habilidades de várias técnicas existentes quando aplicadas para previsão de séries temporais altamente granulares de curto prazo de PM2.5. Mais especificamente, foi realizado um estudo comparativo, levando em consideração tanto os modelos popularmente utilizados quanto os menos utilizados nesta área. O estudo foi realizado considerando dez modelos bem definidos que são ARIMA (média móvel integrada auto-regressiva), SARIMA (ARIMA sazonal), SES (suavização exponencial simples), DES (suavização exponencial dupla), TES (suavização exponencial tripla) , ANN (rede neural artificial), DT (árvore de decisão), kNN (k-vizinho mais próximo), LSTM (memória de curto prazo longo) e MCFO (primeira ordem da cadeia de markov). Foi construído um framework que categoriza os modelos, os implementa em ambientes de execução idênticos e prevê valores sucessivos. A implementação foi realizada em cinco conjuntos de dados de séries temporais de poluição do ar do mundo real, que são coletados de cinco estações de monitoramento de configurações governamentais localizadas em um período de 1 ano (julho de 2018 a junho de 2019). Foi realizada uma análise estatística rigorosa que fornece uma visão da natureza e variabilidade desses dados de séries temporais. A previsão foi realizada em uma base de curto prazo, com foco em alta granularidade, enquanto três diferentes comprimentos de horizonte de previsão (1 dia, 1 semana e 1 mês) foram testados. Eventualmente, os modelos foram comparados em termos de suas unidades de medição de desempenho associadas, a saber, RMSE (média do erro quadrático médio), MAE (erro absoluto médio) e MAPE (erro percentual absoluto médio). Os resultados comparativos verificados com vários conjuntos de dados mostram que todos os modelos possuem menos erros para um horizonte de previsão mais curto, onde o LSTM fornece o melhor desempenho. A superioridade dos modelos de aprendizado de máquina e aprendizado profundo é encontrada no caso de maior comprimento do horizonte de previsão com kNN alcançando a melhor precisão, enquanto a degradação significativa do desempenho do ARIMA é encontrada para horizonte de previsão mais longo. Além disso, TES, DT, kNN, LSTM, MCFO são bem adotados em relação à forma e variabilidade dos dados. Observe que o desempenho em vários comprimentos de horizonte de previsão granular alto foi estudado em vários conjuntos de dados que agregam valor a este trabalho.

Palavras -chave Séries temporais de poluição do ar · PM2.5 · Previsão de curto prazo · Alta granularidade

* Sarbani Roy
sarbani.roy@jadavpurniversity.in

Informações estendidas sobre o autor disponíveis na última página do artigo

1. Introdução

A rápida urbanização resultou no aumento de poluentes atmosféricos que incluem poluentes gasosos como **O3,** NO, **NO2, *NOx,* SO2** e partículas (PM). Os efeitos adversos desses poluentes na saúde humana têm sido motivo de preocupação global há muito tempo e a condição é visivelmente ruim em países em desenvolvimento como a Índia. Em 2015, o Global Burden of Disease (GBD) classificou o PM2,5 (partículas com diâmetro inferior a 2,5 micrômetros) como o quinto entre os principais elementos nocivos. Isso causou 4,2 milhões de mortes e perda de anos de vida ajustados por incapacidade (DALYs) de 103,1 milhões (Yang et al. 2020). Isso, por sua vez, representa um total mundial de mortes de 7,6% e 4,2% dos DALYs universais. Esta enorme mortalidade, bem como a morbidade causada por poluentes atmosféricos são as razões por trás do crescente interesse de pesquisa de monitoramento e previsão desses poluentes nocivos. Os órgãos governamentais ambientais estão construindo rigorosamente modelos adequados (Qiao et al. 2019) para esse fim. Entre estes, a análise de séries temporais e a previsão de tais poluentes têm ganhado popularidade significativa.

Os dados de séries temporais são gerados quando os valores de uma variável são observados em intervalos de tempo discretos e iguais. A modelagem de uma série temporal tenta aprender o contexto subjacente e suas dependências em diferentes etapas de tempo. Esta é uma técnica muito útil de explicar estatisticamente os dados, que ajuda a monitorar e fazer previsões. A principal diferença entre os modelos de previsão de séries temporais e outros modelos de previsão é que, nos primeiros, os valores anteriores ou valores de defasagem são o único parâmetro de previsão enquanto, em outros modelos de previsão, as dependências de valores de defasagem não são consideradas. Além disso, outras variáveis preditoras são levadas em consideração para fazer a previsão e as observações dos dados podem não ser regulares com o tempo. Nesse contexto, os modelos de previsão de séries temporais podem ser ditos como modelos orientados a dados que visam estudar a relação de observação de dados atuais da história passada. Novamente, com a crescente implantação de estações de monitoramento de poluição do ar e dispositivos sensores, hoje em dia as informações de poluentes do ar são geradas de forma regular e periódica. Assim, resulta em vasta disponibilidade de dados de séries temporais. Por sua vez, infuencia diretamente a adaptabilidade da modelagem de séries temporais para a previsão de poluentes atmosféricos. Além disso, esses dados são mais apropriados, precisos e contínuos com longo período de tempo quando comparados aos dados baseados em sensoriamento participativo (PS) que são coletados com dispositivos IoT de baixo custo.

A discussão de problemas de previsão de séries temporais leva em consideração dois fatores. Um fator é o horizonte de previsão. É o intervalo de tempo futuro, durante o qual a previsão será realizada. Outro fator é a granularidade da previsão dentro desse horizonte. A granularidade é o intervalo de ocorrências dentro da série temporal e decide o número de etapas de previsão.
Por exemplo, a previsão de um dia (horizonte de previsão=1) pode ser realizada de uma só vez (inter val=24 h, baixa granularidade) ou pode ser feita a cada hora (intervalo=60 min, alta granularidade). Com base no tipo de dados e nesses fatores, a previsão pode ser de dois tipos: previsão de curto prazo e previsão de longo prazo. Não há faixa de valores definida para esses dois tipos (Divina et al. 2019). É determinado com base no problema. No entanto, na prática, o primeiro contém horizontes que podem variar de algumas horas a alguns meses, enquanto o segundo pode variar de alguns meses a mais de um ano. Uma vez que uma maior precisão é alcançada no caso de previsão de curto prazo, muitos estudos abordam isso. Ao mesmo tempo, pode-se dizer que, em vez de tomar decisões de longo prazo, conhecer a exposição em tempo real ou instantânea a esses poluentes nocivos pode nos ajudar em nosso planejamento diário. Para cumprir este propósito, a previsão com curto intervalo de tempo pode ser muito eficaz. Assim, motiva o estudo da previsão granular alta de curto prazo de séries temporais de poluentes.

Trabalhar com modelos de previsão de séries temporais requer três ações. A primeira é a identificação e processamento dos dados, a segunda é a avaliação dos parâmetros do modelo e a terceira é calcular os valores de erro (Dastorani et al. 2016). Principalmente, as séries temporais univariadas foram tratadas por abordagens baseadas em regressão estatística, ou seja, média móvel integrada auto-regressiva (ARIMA) (Qiao et al. 2019) e sua variante sazonal (SARIMA)
(Dastorani et al. 2016). Essas abordagens lidam com componentes como tendência e sazonalidade nos dados e expressam a relação linear do valor atual com o valor anterior e os resíduos. As abordagens baseadas em suavização exponencial (ES) (Qiao et al. 2019) estão entre outros métodos tradicionais baseados em média móvel. Embora essas abordagens sejam populares, elas não funcionam bem no caso de dados não lineares. Assim, para abordar essas abordagens de aprendizado de máquina de dados (ML) e aprendizado profundo (DL) (Qiao et al. 2019) aparentemente começaram a ser usadas. Redes neurais artificiais (RNA) e suas variantes, redes neurais convolucionais (CNN), memória de curto prazo longo (LSTM), rede bayesiana (BN), heterocedasticidade condicional auto-regressiva generalizada (GARCH) têm sido aplicadas por pesquisadores para série temporal prevista. Mas essas técnicas também podem sofrer de otimização local. Os processos estocásticos que são usados para fins de previsão são a cadeia de markov (MC) e o modelo de markov oculto (HMM). Eles são baseados na teoria da estimativa de máxima verossimilhança. Os métodos de previsão baseados em MC são simples e computacionalmente fáceis e foram considerados bem sucedidos para resolver problemas complexos. No entanto, geralmente para calcular a matriz de transição, são usados grandes dados de histórico. Dados de séries temporais do mundo real geralmente exibem ruídos estocásticos que resultam em comportamento não linear, que controlam o desempenho de previsão da maioria dos modelos de previsão. Além disso, muito menos aplicação desses modelos é encontrada no caso de previsão de curto prazo com alta granularidade. Portanto, explorar modelos de previsão neste contexto pode contribuir na área de previsão de séries temporais de poluição do ar e tais estudos devem ser promovidos. Além disso, a escolha do modelo mais adequado é muito importante e depende de vários aspectos como tempo, custo e precisão dos modelos.

Na literatura existente (Caillault e Bigand 2018; Choubin et al. 2018; Khaldi et al.
2018; Das e Ghosh 2018; Domaÿska e Wojtylak 2012; Meryem e Ismail 2014; Nury et ai. 2017; Ye et ai. 2013; Momani e Naill 2009; Brunelli et ai. 2007; Díaz-Robles et ai. 2008; Qin et ai. 2019) todos os modelos mencionados acima foram usados na previsão da poluição do ar. No entanto, esses trabalhos não se concentram em previsões altamente granulares e de curto prazo de dados de séries temporais **PM2,5 .** Para resolver este problema, neste artigo, um estudo comparativo de vários métodos de previsão foi realizado com base em vários conjuntos de dados de séries temporais de alta granularidade (intervalo de amostragem = 30 minutos) de **PM2.5.** A literatura existente não explora como o desempenho dos modelos de previsão depende de vários horizontes de previsão de curto prazo para séries temporais PM2.5. Para resolver o problema, este artigo investigou o desempenho dos modelos em tais dados para comprimento variável do horizonte de previsão de curto prazo.

Observe que, para aplicar os modelos de previsão para previsão de séries temporais altamente granulares e de curto prazo de PM2.5, não seguimos diretamente nenhuma implementação existente.
Em vez disso, criamos nossa própria implementação específica do problema desses modelos. Por exemplo, a seção de resultados (veja a Tabela 5) deste artigo mostra configurações de parâmetros para os modelos ótimos que são construídos com base em cinco diferentes conjuntos de dados do mundo real.

Neste trabalho, vários modelos de previsão de séries temporais foram categorizados e seu estudo comparativo foi enfatizado com base na previsão de curto prazo de um importante poluente atmosférico, o **PM2.5.** A Fig. 1 descreve a estrutura do estudo. Os dados foram coletados de cinco estações de monitoramento localizadas na Índia. Cada um exibe variação temporal contínua e periódica por um ano. Considerando abordagens estatísticas e baseadas em aprendizado de previsão de séries temporais, dez modelos bem definidos foram considerados. Com base na estratégia de trabalho, eles foram categorizados em quatro tipos: baseados em regressão, ES

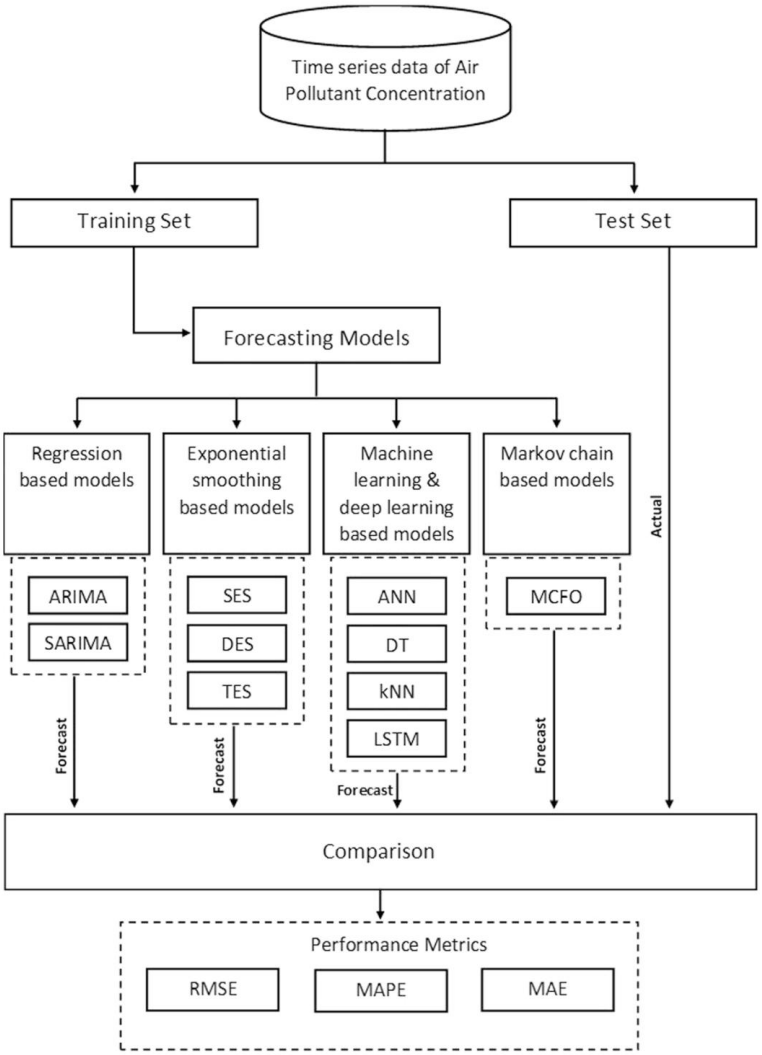


Fig. 1 Estrutura para realizar análise comparativa de modelos de previsão de séries temporais no contexto de dados de poluentes atmosféricos

baseado, baseado em ML e DL e baseado em MC. Todos os cinco conjuntos de dados foram divididos em conjuntos de teste de treinamento que foram discutidos em detalhes na Seção. 4.3. O processo de previsão foi realizado em curto prazo para três diferentes comprimentos de horizontes de previsão que são de 1 dia, 1 semana e 1 mês. O comprimento variável do horizonte de previsão nos ajuda a entender a variação no desempenho dos modelos com passos de tempo de previsão crescentes. Neste trabalho, o intervalo de previsão granular alto é definido como 30 minutos. Observe que a análise geral neste artigo para comparar diferentes métodos é signifcativamente diferente em termos de propósito (ou seja, previsão granular alta e de curto prazo), resultados, insights e relatórios dos trabalhos existentes sobre previsão de séries temporais PM2.5.

Este estudo foi organizado da seguinte forma. Na Sec. 2, algumas literaturas relacionadas ao estudo atual foram discutidas com a devida referência. Seita 3.1 descreve o problema

demonstração. O estudo teórico das quatro categorias dos modelos de previsão de séries temporais foi discutido na Seção. 3.2. Na Seção 4.1, a descrição dos dados e algumas estatísticas importantes dos dados foram fornecidas. A seção 4.2 descreve as métricas de desempenho utilizadas para avaliação dos modelos. As configurações de implementação e os resultados correspondentes dos modelos individuais foram apresentados na Seção. 4.3. Ao final, a conclusão e as direções futuras do trabalho são apresentadas na Sec. 5.

2. Trabalho relacionado

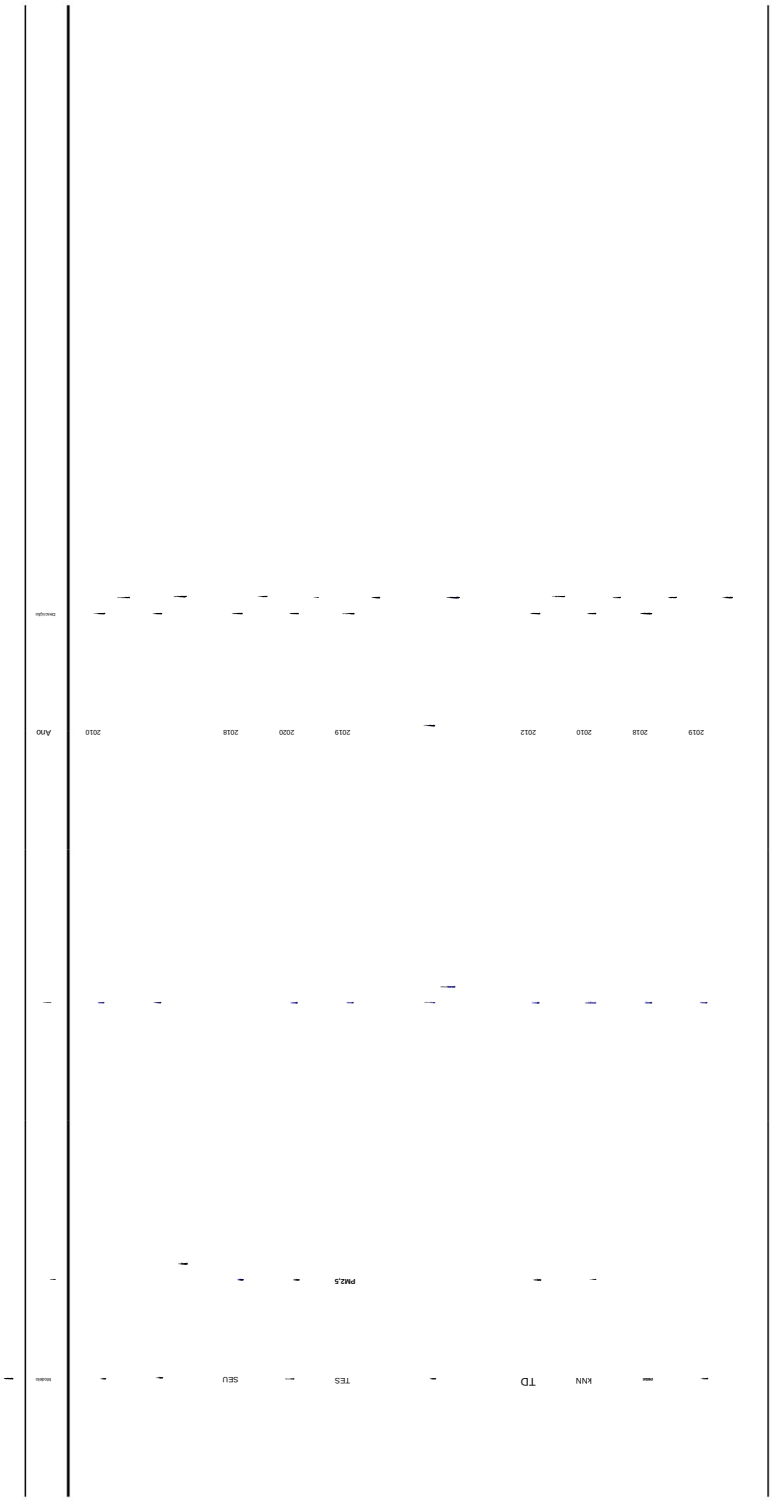
Verificou-se que várias pesquisas estão sendo conduzidas com foco em problemas de previsão de séries temporais. Por exemplo, a literatura (Divina et al. 2019) apresenta uma avaliação comparativa de técnicas de previsão aplicadas ao longo de séries temporais. O objetivo deste trabalho é prever o consumo de energia de edifícios inteligentes. Ele conclui que as abordagens baseadas em ML são mais adequadas. O estudo de Dastorani et al. (2016) foi feito com o objetivo de investigar as potencialidades dos métodos convencionais de previsão de séries temporais no contexto da previsão de dados mensais de chuva em nove estações de monitoramento localizadas no nordeste do Irã. Este trabalho conclui que o melhor modelo pode variar com as mudanças nos dados. Um trabalho recente (Caillault e Bigand 2018) mostra uma estrutura básica de previsão de séries temporais de fenômenos meteorológicos univariados onde modelos tradicionais como suavização exponencial única (SES), modelos SARIMA e ML sazonal-naive (Snaive), feed-forward neural network (FFNN)), séries temporais estruturais bayesianas (BSTS) e imputação baseada em distorção temporal dinâmica (DTWBI) foram comparadas. Outro trabalho (Choubin et al. 2018) trata do modelo de árvores baseadas em classificação e regressão (CART), ou seja, um modelo de árvore de decisão (DT). Além disso, este trabalho revela um estudo relativo do CART com modelo ARIMA e sistema de inferência neuro-fuzzy adaptativo (ANFIS) com base em dados de precipitação. Em Khaldi et ai. (2018), a capacidade de variantes de RNA foi examinada em termos de previsão de diferentes padrões de séries temporais mensais. Abordagens controladas por dados foram mostradas em Das e Ghosh (2018) para prever séries temporais de dados meteorológicos. Este trabalho apresenta o estudo de técnicas de inteligência computacional e suas variantes. Também mostra experimentalmente a superioridade das BNs no contexto da predição de séries temporais de dados meteorológicos. O estudo de Domaÿska e Wojtylak (2012) mostra a necessidade de grandes séries temporais de informações meteorológicas para aplicar conceitos difusos para prever várias concentrações de formigas poluentes do ar. Em Meryem e Ismail (2014), um estudo relativo de regressão linear, máquinas de vetor de suporte e perceptron multicamada foi feito em termos de sua capacidade de prever quatro conjuntos de dados representativos de séries temporais. Os resultados mostram que a regressão linear produziu as melhores previsões. A técnica Wavelet foi combinada com ARIMA e ANN em Nury et al. (2017). Além disso, sua capacidade de previsão foi comparada com base em séries temporais de temperatura no nordeste de Bangladesh. As avaliações neste estudo mostraram melhor desempenho do modelo wavelet-ARIMA em relação ao modelo wavelet-ANN. Um trabalho semelhante (Ye et al. 2013) usou uma abordagem determinístico-estocástica combinada (DSC) sobre o modelo convencional de séries temporais SARIMA para prever a temperatura absoluta mensal da superfície. A metodologia Box-Jen kins foi usada em Momani e Naill (2009) para modelar o ARIMA para prever as chuvas no aeroporto de Amã mensalmente. O principal objetivo é ajudar os tomadores de decisão a priorizar a gestão da demanda de água. Os estudos supracitados mostram a ampla aplicabilidade dos modelos de previsão na área de diversos f

Agora, problemas de previsão dos principais poluentes atmosféricos estão em demanda há muito tempo para regulá-los dentro de limites (Brunelli et al. 2007). Neste trabalho, o modelo com

(RNN) foi proposta para prever as maiores concentrações de **SO2, O3, PM10, NO2,** CO diariamente, em Palermo, Itália. Uma abordagem híbrida combinando ARIMA e RNA foi mostrada em Díaz-Robles et al. (2008) que foi encontrado para melhorar a precisão do modelo ao prever **PM10** diariamente, bem como de hora em hora. As abordagens baseadas em CNN e LSTM foram propostas em Qin et al. (2019) que trata da previsão da concentração urbana de **PM2.5 .** Uma forma estendida de LSTM (LSTME) foi proposta em Li et al. (2017) para previsão de concentração de poluentes atmosféricos. Ajudou a extrair características intrínsecas dos dados históricos. O trabalho em Freeman et al. (2018) treinam um modelo de deep learning com RNN-LSTM, que prevê concentrações médias de **O3** para um intervalo de tempo de 8 h, com alta precisão. Outro trabalho em Wang et al. (2017) propõe uma nova versão híbrida da metodologia GARCH que combina as capacidades individuais de previsão do SVM, bem como o modelo ARIMA. Os modelos foram comprovados pela previsão de séries temporais de **PM2,5** para Shen zhen, China. A previsão foi realizada de hora em hora durante dez dias e os resultados mostram uma boa precisão. Os autores de Oprea et al. (2016) apresentaram resultados comparativos de RNAs e ANFIS quando aplicados para previsão de material particulado **(PM2.5) .** A aplicação da RNA wavelet modificada foi encontrada em Zainuddin e Pauline (2011) para prever séries temporais de dados de poluição do ar. Os resultados mostraram desempenho superior de WNNs quando integrados com outras famílias de wavelets na camada oculta. Novas perspectivas de auto-regressão de rede neural (NNAR) foram mostradas em Mahajan et al. (2017) quando validado com previsão de séries temporais **PM2,5** com intervalo horário. Ele também fornece estudo comparativo de NNAR com Holt-Winters (aditivo), bem como modelos ARIMA usando medidas de desempenho. Uma nova versão cinza do modelo Holt-Winters foi proposta em Wu et al. (2017) que trata de séries temporais estocásticas de natureza não estacionária. O modelo foi finalmente utilizado para prever o índice de qualidade do ar. Os efeitos sazonais foram tratados pelo método de Holt Winters. Padrões semelhantes foram pesquisados nos dados históricos em Lin et al. (2017) usando a técnica de k-vizinho mais próximo (kNN). Esses padrões de séries temporais, juntamente com outras informações de séries temporais, foram usados pelo modelo proposto NARX. O modelo combina o comportamento não linear da rede auto-regressiva com a rede neural (parâmetros exógenos) e finalmente é usado para prever a qualidade do ar a longo prazo. Outro trabalho (Dong et al. 2009) mostra as limitações do HMM ao longo de séries temporais e propõe um framework para modelos ocultos de semi markov (HSMMs) que apresentam melhor desempenho enquanto previsão de concentração de **PM2,5.** Um trabalho inicial (Huang e Cheng 2008) usou o modelo de série temporal baseado no operador OWA que previa a concentração de **O3 .** Outro trabalho inicial em Chelani (2005) utilizou a teoria de sistemas caóticos e a usou sobre o modelo de rede neural para prever os valores de **PM10** para uma área residencial em Delhi, Índia. O classificador da árvore de decisão foi usado em Amato et al. (2008) para reconstrução de dados de séries temporais onde o foco é aprender informações ambientais obtidas por um sistema de monitoramento equipado com múltiplos sensores e distribuídos pela cidade de Taranto.

A Tabela 1 mostra alguns trabalhos relacionados ao cenário atual e séries temporais de diferentes poluentes atmosféricos que vem sendo utilizados.

Esta literatura existente apresentada na Sec. 2 mostram que os problemas de previsão de séries temporais estão agora sendo imensamente focados pelos pesquisadores. Várias técnicas de inteligência computacional e modelos estatísticos estão sendo propostos e aplicados. Além disso, percebe-se claramente que esses modelos estão sendo rigorosamente aplicados em dados de séries temporais de poluição do ar. Nesse contexto, a previsão de curto prazo com alta granularidade está sendo tarefa mais importante para planejamento ou governança instantânea. Muito menos estudos são encontrados para abordar esta área. O presente trabalho tenta descobrir a utilidade de modelos de previsão que se enquadram em diferentes categorias. Com modelos estatísticos de previsão de séries temporais como modelos baseados em ARIMA, SARIMA ou ES, ele examina modelos de ML e DL amplamente utilizados, como ANN ou LSTM. Por outro lado, este trabalho também investiga modelos de ML como DT, kNN juntamente com modelos baseados em MC



que têm menos aplicações na modelagem de séries temporais. A análise comparativa de desempenho foi feita aplicando todos esses modelos em séries temporais de **PM2.5.** A área de pesquisa de previsão de séries temporais ainda não foi explorada com dados de poluentes para esta localidade. Além disso, essa alta granularidade de previsão é rara neste contexto. O estudo estatístico dos dados representa a natureza geral do **PM2,5** para Kolkata, Índia. A granularidade (previsão de 30 minutos) da previsão é muito alta e os resultados obtidos para três horizontes de previsão diferentes ajudam a escolher estratégias que podem ser bem adotadas para esses dados. Assim, a evolução desse estudo deve ser incentivada.

3 Declaração do problema e métodos

Nesta seção, o problema da previsão de séries temporais foi apresentado formalmente. Em seguida, os modelos foram discutidos em suas respectivas categorias que são consideradas para estudo comparativo em relação ao contexto atual.

3.1 Declaração do problema

Dado um ambiente de séries temporais univariadas, o objetivo do estudo é implementar vários modelos de previsão existentes e compará-los para decidir a aplicabilidade dos modelos. Duas etapas são seguidas ao trabalhar com esse problema. O primeiro é representar a série temporal construindo um modelo e o segundo é obter resultados de previsão usando esse modelo (Meryem e Ismail 2014). Em um ambiente típico de séries temporais, as observações são feitas em intervalos regulares equiespaçados de tempo ***ÿt*** (por exemplo, horas, dias, meses, anos). Um valor recente deve ser uma função de valores passados, também chamados de valores de atraso. Portanto, com as ocorrências do evento representadas pela variável V, se ***V(t)*** representa *V* no tempo t, então o modelo deve ser construído da seguinte forma,

$$\boldsymbol{V(t)} = \boldsymbol{F}\ \mathbf{(V(t\ddot{y}1\ddot{y}t),}\ V(t\ddot{y}2\ddot{y}t), ..., V(t\ddot{y}k\ddot{y}t)) + \boldsymbol{\ddot{y}\ddot{y}(t)} \quad (1)$$

onde, **V(tÿ1.ÿt)** é o valor de *V* na ocorrência anterior e ***V(tÿk.ÿt)*** é a ocorrência de V, *k* passos atrás. ***ÿÿ(t)*** é o erro no tempo *t* também chamado de residual. Na próxima etapa o modelo é usado para prever as próximas *k* ocorrências com intervalo ***ÿt .*** Portanto ***k*** ÿ ***ÿt*** é o horizonte total de previsão. Na literatura, verifica-se que, para um determinado valor de ***ÿt,*** nenhum modelo tem um bom desempenho para todos os valores de k. Assim, neste trabalho o objetivo é implementar e comparar o desempenho de diferentes modelos existentes para diferentes valores de k ie diferentes horizontes de previsão. A comparação foi feita com base em dados de séries temporais de **PM2.5.** Com ***ÿt*** = 30 minutos, ou seja, alto nível de granularidade, as três configurações distintas do horizonte de previsão de curto prazo são um dia, uma semana e um mês, ou seja, k=48, 336 e 1440, respectivamente.

3.2 Métodos

O processo de previsão de uma série temporal utiliza um modelo apropriado para decidir valores futuros com base em históricos anteriores. O tipo de dados é diferente de forma que adiciona explicitamente uma dependência ordenada entre as observações. Aqui, uma dimensão de tempo atua como uma restrição, enquanto fornece uma fonte de informações adicionais. Conforme mostrado na Fig. 1 os modelos que tratam do problema de previsão de séries temporais, foram categorizados em quatro tipos. Nas subseções seguintes um estudo teórico desses modelos sob essas

quatro categorias foram dadas. Além disso, as etapas seguidas para trabalhar com essas técnicas foram demonstradas na Fig. 2.

3.2.1 Modelos baseados em regressão

Modelos baseados em regressão que são aplicáveis em séries temporais univariadas são, ARIMA e sua variante sazonal SARIMA. Essa classe de modelos foi introduzida por George Box e Gwilym Jenkins em 1970. Esses são os métodos clássicos de previsão de séries temporais lineares que se concentram principalmente nas relações lineares entre as observações. Antes de aplicar esses modelos é necessário saber se os dados apresentam variações ou tendências de longo prazo, variações de curto prazo ou sazonalidade e movimentos aleatórios Pode-se pensar em dados de séries temporais como uma combinação dessas variações de forma aditiva ou multiplicativa. Dados sem essas variações são referidos como dados estacionários. Agora, para verificar a estacionariedade, testes demográficos bem conhecidos como o teste Dickey-Fuller (ADF) aumentado e o teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS) são realizados nos dados.

- **ARIMA:** Este método geralmente lida com dados não estacionários e aborda a tendência neles. Na forma típica, o ARIMA exibe três componentes, a saber, AR(p), I(d) e MA(q) (Kumar e Jain 2010).

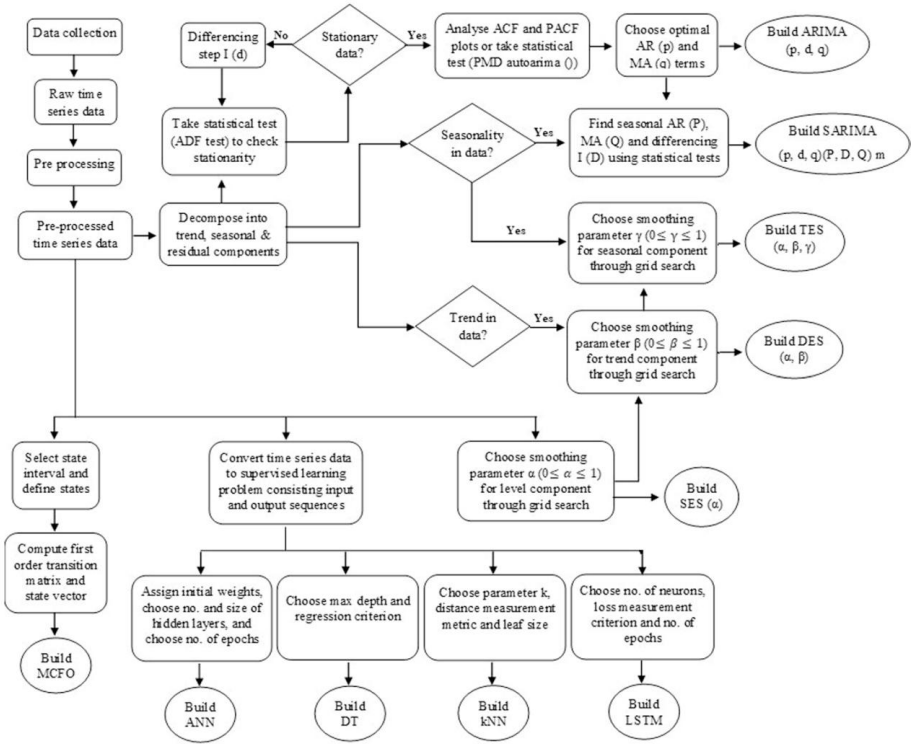


Fig. 2 Blocos de construção dos modelos estudados neste trabalho

1. O componente AR(p) é baseado no processo autorregressivo, ou seja, aqui o valor alvo é gerado como a regressão linear de suas próprias defasagens. O inteiro não negativo p representa os números de termos autorregressivos também chamados de ordem de AR.
2. O I(d) é o componente integrado onde, d é um inteiro não negativo que representa o número de vezes que o processo de diferenciação é realizado para transformar um dado não estacionário em estacionário.
3. O componente MA(q) representa o valor alvo em termos de termos de erro atuais e anteriores e o valor inteiro não negativo q representa o tamanho da janela.

Assim, com ***V(t)*** representando a ocorrência da variável de série temporal no tempo *t* e os subscritos representando o número de valores de defasagem levados em consideração, AR(p) é representado da seguinte forma

$$V(t) = \ddot{y}\ddot{y}0 + \ddot{y}\ddot{y}1V(t \ddot{y} 1) + \ddot{y}\ddot{y}2V(t \ddot{y} 2) + \ldots + \ddot{y}\ddot{y}pV(t \ddot{y} p) + \ddot{y}\ddot{y}(t) \quad (2)$$

aqui, **ÿÿ0** é constante e **ÿÿ1, ..., *ÿÿp*** são coeficientes de defasagem e ***ÿÿ(t)*** é o resíduo no tempo t. Da Eq. 2 pode-se entender que é basicamente um processo de regressão linear e ***V(t)*** depende de seus próprios valores de atraso. Novamente o modelo MA(q) que representa ***V(t)*** em termos de valores de ruído observados no presente e no passado, satisfaz a seguinte equação

$$V(t) = \ddot{y}\ddot{y}(t) + \ddot{y}\ddot{y}1\ddot{y}\ddot{y}(t \ddot{y} 1) + \ddot{y}\ddot{y}2\ddot{y}\ddot{y}(t \ddot{y} 2) + \ldots + \ddot{y}\ddot{y}q\ddot{y}\ddot{y}(t \ddot{y} q) \quad (3)$$

onde **ÿÿ1, ..., *ÿÿq*** são constituídos como os parâmetros de MA(q) e *ÿÿ* representa os termos de erro de acordo com o tempo. Os modelos AR(p) e MA(q) são usados para fins de previsão para séries temporais estacionárias simples. Em geral, os dados de séries temporais são mais complexos e exibem alguma tendência que é denotada pela ordem d do componente I do ARIMA. Assim, para uma série temporal não estacionária ARIMA possui a forma mostrada na Eq. 4.

$$V(t)(1 \ddot{y} \sum_{i=1}^{p} \ddot{y}\ddot{y}i\ddot{y}i)(1 - \ddot{y})^{d} = \ddot{y}\ddot{y}(t)(1 + \sum_{i=1}^{q} \ddot{y}\ddot{y}i\ddot{y}i) + \ddot{y}\ddot{y} \quad (4)$$

onde, ÿ é o operador de atraso que denota a distância do atraso do tempo atual *t* e *ÿÿ* é uma constante. Além disso, a Eq. 4 pode ser reescrito na forma polinomial lag apresentada na Eq. 5 onde, ÿ =*ÿ* **1** ÿ ÿ é o operador de diferenciação.

$$V(t)\,p(\ddot{y})\,\ddot{y}\ddot{y}^{d} = (t)\,q(\ddot{y}) + \quad (5)$$

• **SARIMA:** Com tendência, os dados de séries temporais geralmente incorporam uma variação sazonal ou cíclica. Para resolver este fator, o SARIMA foi desenvolvido a partir do ARIMA. Com as variáveis do ARIMA, o SARIMA adiciona os componentes sazonais AR(P), I(D) e MA(Q). P, D, *Q* são inteiros que são sempre positivos. Esses três parâmetros funcionam no múltiplo de *m ,* que denota o número de passos de tempo em uma estação. Por exemplo, para observações mensais e estação anual, ***m*** = **12.** Em geral, o modelo SARIMA com ordem (p, d, q).(P, D, Q).m e operador de distância de atraso recorrente ***ÿm,*** é representado a seguir Formato

$$V(t)\,p(\ddot{y})AP(\ddot{y}m)\,\ddot{y}\ddot{y}\,d\ddot{y}\ddot{y}_{m}^{D} = (t)\,q(\ddot{y})BQ(\ddot{y}m) + \quad (6)$$

onde, *A* e *B* são parâmetros do componente sazonal AR e MA respectivamente e *ÿÿ* $_{m}$ = **1** - ***ÿm*** é o operador diferencial sazonal.

Para obter melhores resultados, os parâmetros dos modelos ARIMA e SARIMA precisam ser otimizados. Natureza da função de autocorrelação ACF, parcelas parciais de ACF (Gocheva-Ilieva et al. 2014) são examinadas. ACF e PACF descrevem a relação do valor atual na série temporal com seus valores de defasagem e resíduos, respectivamente. Mas esse processo é tranquilo, difícil, demorado e menos preciso. Assim, para este propósito, estão disponíveis testes estatísticos integrados que ajustam esses parâmetros com base no critério de informação akaike (AIC). Modelo com menor valor de AIC é adotado pelos testes estatísticos.

No cenário prático, dados de séries temporais, como concentração de poluentes atmosféricos, geralmente são obtidos em intervalos curtos, de modo que há mais chances de ocorrer uma recorrência cíclica, ou seja, sazonalidade. Assim, o SARIMA é considerado mais adequado e apropriado do que o modelo ARIMA.

3.2.2 Modelos baseados em suavização exponencial

Esses modelos prevêem séries temporais univariadas com tendência sistemática ou mudança periódica. Em vez de atribuir pesos iguais a todos os atrasos, em modelos de previsão baseados em ES, os dados observados recentemente obtêm mais pesos e o peso diminui exponencialmente para observações anteriores. A classe de modelo de previsão baseado em ES inclui SES, DES também conhecido como método de Holt e TES ou método de Holt-Winters.

- **SES:** Este é o mais simples nesta categoria de modelos que é aplicável para dados univariados sem tendência ou sazonalidade (Roy et al. 2018). Na forma simples de média móvel exponencialmente ponderada, ela é apresentada da seguinte forma

$$V_{(t)} = A\ddot{y}t \ddot{y}_{k=1}^{1}(1 - A)^{k} V(t\ddot{y}k) \quad (7)$$

  onde, *A* representa o fator de suavização que controla a taxa de decréscimo exponencial dos efeitos de observação com o tempo. O valor de *A* está entre 0 e 1(0 ÿ ***A*** ÿ 1). Quanto maior o valor de *A* , mais atenção é dada apenas ao passado recente, enquanto quanto menor o valor, mais o histórico é levado em consideração durante as previsões. O SES simplesmente aborda o componente de nível (L) que é um estimador da média local em qualquer ponto do tempo. A suposição é que os dados flutuem em torno de um nível estável.
- **DES:** Esta é uma extensão do SES que foi modelada para tratar dados com tendência linear, mas sem padrão sazonal (Bose et al. 2020). O objetivo é atualizar o componente de tendência ou inclinação por meio de suavização exponencial. DES introduz um novo fator de suavização *B* **(0** ÿ ***B*** ÿ 1) para componente de tendência (T) com componente de nível existente *(L)* de SES e o fator de suavização para componente de nível A. A equação de previsão de DES na forma de componente é

$$V\ddot{y}_{(t+k)|t} = L(t) + k \ddot{y} T(t) \quad (8)$$

  onde, *L* e *T* durante o tempo t, são calculados usando a Eq. 9 e Eq. 10 respectivamente.

$$L(t) = AV(t) + (1 - A)(V\ddot{y}_{(t)}) \quad (9)$$

$$T(t) = B(L(t) \ddot{y} L(t\ddot{y}1)) + (1 \ddot{y} B)T(t\ddot{y}1) \quad (10)$$

  onde, $V_{(t)} = L(t\ddot{y}1) + T(t\ddot{y}1)$ representa a previsão para o tempo t.

• **TES:** Embora o DES amplie a aplicabilidade dos modelos baseados em suavização exponencial abordando dados com tendência, no entanto, faltava sua capacidade de lidar com o efeito sazonal em dados de séries temporais. TES (Ventura et al. 2019) é mais uma extensão do DES que apresenta outro fator de suavização *G* **(0** ÿ ***G*** ÿ 1) que aborda o componente sazonal (S). Novamente, dependendo da natureza do componente sazonal, o TES pode ser do tipo aditivo (as variações sazonais são quase constantes ao longo da série) e multiplicativo (variando a variação sazonal com o nível). A forma de componente do TES para o método aditivo é mostrada abaixo

$$V\ddot{y}_{(t+k)|t} = \boldsymbol{L(t)} + \boldsymbol{k} \ddot{y} \boldsymbol{T(t)} + \boldsymbol{S(t+k\ddot{y}m(d+1))} \quad (11)$$

$$\boldsymbol{L(t)} = \boldsymbol{A}\,(V(t) \ddot{y} S(t\ddot{y}m)) + (1 \ddot{y} A)(L(t\ddot{y}1) + T(t\ddot{y}1)) \quad (12)$$

$$\boldsymbol{T(t)} = \boldsymbol{B}\,(L(t) \ddot{y} L(t\ddot{y}1)) + (1 \ddot{y} \boldsymbol{B})\boldsymbol{T(t\ddot{y}1)} \quad (13)$$

$$\boldsymbol{S(t)} = \boldsymbol{G}\,(V(t) \ddot{y} L(t\ddot{y}1) \ddot{y} T(t\ddot{y}1)) + (1 \ddot{y} \boldsymbol{G})\boldsymbol{S(t\ddot{y}m)} \quad (14)$$

Aqui, *m* é a contagem de valores para uma única temporada. *d* é definido como uma porção inteira de (k-1)/m e é usado para influenciar a previsão coletando amostras de observações finais.

Uma grande falha do DES ou TES aqui é que eles assumem a tendência nos dados de aumentar ou diminuir monotonamente, o que pode não ser o caso sempre no cenário do mundo real. Portanto, para amortecer esse crescimento contínuo, às vezes, novos parâmetros são adicionados à formulação.

3.2.3 Modelos de aprendizado de máquina e aprendizado profundo

Os modelos de ML são orientados por dados e também chamados de modelos de caixa preta. Essa classe de modelos é amplamente aplicada a dados de séries temporais univariadas como uma alternativa aos modelos clássicos gerais de previsão. Eles aprendem a dependência complexa de valores futuros e valores passados dos dados históricos. Conforme mostrado na Fig. 2, aqui os dados univariados precisam ser enquadrados como um problema de aprendizado supervisionado, onde consistirá de uma entrada e uma saída. Nesta família de modelos, discutimos alguns modelos comumente usados para previsão de séries temporais.

• **ANN:** A ideia por trás de uma RNA (Unnikrishnan e Madhu 2019) é persuadida do sistema nervoso biológico que é composto por neurônios. É uma das técnicas de inteligência computacional mais utilizadas que aprendem com as observações passadas, reconhecem um padrão nas observações e as utilizam para prever valores futuros.
Além disso, eles são capazes de lidar com dados ruidosos e nenhuma suposição prévia é necessária para a construção do modelo. A única dependência está nas características dos dados que influenciam amplamente o modelo. O FFNN com uma camada oculta é um dos modelos mais utilizados para previsão de séries temporais com univariável (Zhang et al. 1998). Neste contexto, a rede é treinada com sequências anteriores de observações.
Eq. 15 mostra a relação de ***entrada(V(tÿ1),*** V(tÿ2), ..., V(tÿk)) para saída(V(t)) no caso de previsão de uma etapa com previsão de RNA modelo.

$$V(t) = w0 + \sum_{q=1}^{H} wq.f\left(w0.q + \sum_{r=1}^{k} wqr.V(t-r) \right) + \varepsilon(t) \quad (15)$$

onde, *k* representa o número de entradas ou os valores de atraso na sequência de entrada, *H* fornece a contagem de nós ocultos junto com wq(q = **0, 1, ...,** H), ***wqr(q*** = **0, 1, . ..,** **H,r** = **0, 1, ...,** k)

como os pesos de conexão que são parâmetros da ANN. Depois de atribuir pesos, as entradas ponderadas são adicionadas e passadas para uma função de ativação (geralmente uma função sigmóide) que, por sua vez, fornece a saída final. O trabalho mais importante na construção do modelo de RNA é escolher a dimensão da sequência de entrada que influencia as saídas de previsão de várias etapas. O modelo é treinado com o algoritmo backpropagation (BP) que é computacionalmente simples e amplamente utilizado para treinamento de RNAs e minimiza a função de erro na camada de saída.

- **DT:** DT (Chu et al. 2012) é uma estrutura em árvore de um algoritmo de tomada de decisão que se enquadra na categoria de modelos de ML. É amplamente utilizado para resolver problemas de regressão (não supervisionados) e de classificação (supervisionados). Além disso, a variável alvo pode ser contínua ou categórica. Para séries temporais ele prevê ocorrências do fenômeno alvo dependendo de regras de supervisão que são inferidas de observações passadas. Ela começa com a raiz da árvore, cresce dividindo-a em conjuntos menores contendo características e para quando nenhum aumento significativo na homogeneidade é encontrado. O desvio padrão é usado para calcular a homogeneidade. Os DTs são altamente sensíveis aos dados pelos quais são treinados. Portanto, é necessária uma amostragem aleatória dos dados de treinamento. Vários algoritmos são usados pelo DT para decidir como dividir um nó em subnós, no entanto, para dados de séries temporais, as árvores de decisão são modeladas com base no algoritmo de árvore de classificação e regressão (CART) (Choubin et al. 2018) onde cada nó faz uma decisão baseada em uma desigualdade. O algoritmo foi desenvolvido por Breiman et al. (1984) e inclui três etapas: (1) construção da árvore com tamanho máximo, (2) escolha do tamanho ideal da árvore, (3) previsão de valores da variável alvo usando a árvore construída. Um grande problema com os DTs é que seu desempenho depende dos dados e sofre de overfting dos dados.
- **kNN:** É o modelo de ML mais primitivo, simples e amplamente utilizado tanto para problemas de classificação quanto para problemas de regressão (Dragomir 2010). Fix e Hodges introduziram esse conceito em 1951 (Al-Qahtani e Crone 2013). O algoritmo de classificação kNN, aplicado à variável categórica, pode ser estendido com regressão para fins de previsão. A ideia subjacente deste método é encontrar k sequências passadas que são mais semelhantes à que está sendo prevista. A previsão de valores sucessivos na sequência é feita de maneira iterativa. A intuição por trás é que existe certo padrão de repetição nos dados. As etapas envolvidas na previsão de séries temporais univariadas com kNN são: (1) criação de vetores, (2) estimativa e previsão de similaridade. Em séries temporais univariadas, devido à ausência de variáveis explicativas, um conjunto de vetores autorregressivos é criado como feições. *Vÿÿÿÿÿÿm* = (v(tÿ1), v(tÿ2), ... v(tÿ(mÿ1))) denota um vetor para o evento m sendo um número natural **(t)** tal dentro do conjunto de vetor é predeterminado manualmente. Os vetores podem ser sobrepostos ou não sobrepostos. A próxima é a estimativa de similaridade para descobrir sequências recorrentes. As técnicas de medição de distância euclidiana ou minkowski são geralmente usadas para medição de similaridade. A etapa final é agregar o número k de vetores, mais semelhantes aos valores alvo. Isto se faz do seguinte modo

$$V(t+H) = \sum_{q=0}^{m\ddot{y}1} \left(1k \sum_{r=1}^{k} \ddot{y} \right) VR \quad (16)$$

(tÿq))

Embora a área de problemas de previsão de séries temporais tenha sido dominada por modelos lineares devido à sua eficácia e simplicidade, no entanto, a aplicação de abordagens de aprendizado profundo (Qiao et al. 2019; Das e Ghosh 2018) em séries temporais está surgindo à medida que aprendem automaticamente o relação complexa arbitrária entre entradas e saídas. Eles também suportam problemas com múltiplas entradas e saídas. Além disso, eles lidam automaticamente com componentes como tendências e sazonalidade. Aqui apresentamos a visão teórica do LSTM um modelo DL popularmente aplicado para previsão de séries temporais.

• **LSTM:** O conceito de LSTM veio com o aumento da rede neural recorrente (RNN) onde em vez de memória momentânea de RNN, memória de longo prazo é adicionada a RNNs. Ele lida com o problema do gradiente de fuga devido ao qual o aprendizado é encerrado nas redes neurais. Aqui o resultado do estado atual é fornecido para o próximo estado. Os blocos de construção básicos do LSTM são três tipos de portas, a saber, entrada (gravação), saída (leitura) e porta de esquecimento (redefinição). O portão de esquecimento especifica os valores precedentes que são citados no futuro. As portas estão contidas em blocos de memória que, por sua vez, são conectados por meio de camadas. As portas agem como interruptores que decidem quais dados manter e descartar e por meio de qual memória de longo prazo é incorporada ao modelo. Outra porta é a porta de modulação que atualiza as entradas de corrente. Portanto, todas essas portas lidam com a entrada presente ***v(t)*** e o estado passado s(tÿ1) (Haider et al. 2019).

Com *I, O, F, M* denotando as portas de entrada, saída, esquecimento e modulação respectivamente, C representando a célula e *W, U* representando os pesos de entrada e peso do estado anterior respectivamente, os resultados passo a passo para cada uma dessas portas são dadas abaixo

$$I(t) = F_{(\,W(q)vC(t) + U(q)sC(t\ddot{y}1) + \ddot{y}\ddot{y}(i)\,)} \tag{17}$$

$$M(t) = \tanh_{(\,W(C)vC(t) + U(C)sC(t\ddot{y}1) + \ddot{y}\ddot{y}(C)\,)} \tag{18}$$

$$M(t) = F_{(\,W(F)vC(t) + U(F)sC(t\ddot{y}1) + \ddot{y}\ddot{y}(f)\,)} \tag{19}$$

aqui, *ÿÿ* é o viés e os subscritos apontam para as portas correspondentes. O cálculo para atualizar o valor de uma célula ***C(t)*** é,

$$C(t) = F_{(\,W(O) + vC(t) + U(O)sC(t\ddot{y}1) + V(O)sC(t) + \ddot{y}\ddot{y}(O)\,)} \tag{20}$$

Finalmente, a porta de saída (controlada pelo estado anterior da célula, estado atual da célula e entrada atual) e o estado atualizado da célula são mostrados na Eq. 21 e Eq. 22 respectivamente.

$$o(t) = F_{(\,W(O)vC(t) + U(O)sC(t\ddot{y}1) + V(O)sC(t) + \ddot{y}\ddot{y}(O)\,)} \tag{21}$$

$$s(t) = o(t)\,\tanh(C(t)) \tag{22}$$

Como os dados de poluentes do ar das estações de monitoramento variam temporalmente e problemas de saúde ocorrem devido à exposição de longo prazo aos poluentes, os LSTMs são considerados uma boa previsão

modelo para esse tipo de dados, pois usa observações anteriores por um longo período de tempo e os erros são preservados em uma célula fechada.

3.2.4 Modelo baseado em cadeia de Markov

Cadeias de Markov são comumente usadas para determinar as probabilidades de sequências de variáveis aleatórias. Essas variáveis estocásticas, também chamadas de estados, obtêm seus valores de conjuntos. Esses conjuntos são a representação de coisas de interesse como o clima. Uma suposição principal por trás do uso de MC é que, ao prever a ocorrência de uma sequência de variáveis aleatórias, apenas o impacto do estado atual é importante. Os estados que aparecem antes do estado atual não têm efeito sobre o estado atual. De uma forma completa, um MCFO consiste no tripleto **({S}, ÿÿÿÿ, ÿÿ)** onde,

(i) *ÿ* = **S1...SK** é um conjunto de K estados ou a variável de estado.

(ii) **ÿÿÿÿ** representa uma matriz de dimensão ***K*** × K, onde **ÿÿÿÿ** ÿ **ÿÿÿÿ.** Cada elemento de **ÿÿÿÿ** nos diz a probabilidade de transição do estado i para j, onde **i, *j*** ÿ K. Portanto, a soma de cada linha da matriz é sempre 1.

(iii) **ÿÿ** é um vetor de comprimento K que mostra a distribuição de probabilidade inicial dos estados. Os elementos de **ÿÿ** nos dizem a probabilidade de estar em um determinado estado em um determinado momento. A soma dos elementos do vetor é sempre 1. Um 0 nos elementos significa que o estado correspondente não pode ser o estado inicial para iniciar a cadeia.

Com os componentes mencionados acima calculados para a Cadeia de Markov de uma variável de estado até o tempo atual (t), é possível estimar o vetor de estado para um passo futuro ***(t*** + 1) com a multiplicação da matriz de transição de estado, gerada a partir de o histórico da sequência de estados e o vetor de estado inicial. O novo vetor de estado informa o estado mais provável que pode ocorrer no tempo ***(t*** + 1). Para prever os próximos saltos ***(t*** + h) na série temporal, a matriz de transição é multiplicada pelo vetor resultante de cada etapa do total de h etapas. A cadeia se propaga com t. Novamente, a suposição de Markov afirma que as observações futuras ao longo da série são influenciadas apenas pelo estado presente. Estados ocorridos antes do estado presente não influenciam o futuro. Assim, a cadeia de Markov é formulada como,

$$P\{S(t+1) = s^{p}_{(t+1)} \mid S(t\ddot{y}q) = s^{o}_{(t\ddot{y}q)}, \ddot{y}q \ddot{y} [0, 1, ..., (t \ddot{y} 1)]\} = P\{S(t+1) = s^{p}_{(t+1)} \mid S(t) = s^{o}_{(t)}\} \quad (23)$$

onde, cada estado previsto ***sp*** e estado observado ***so*** ÿ *ÿ*.

Considera-se que a natureza da série temporal abordada por um modelo de cadeia de markov varia aleatoriamente com o tempo. Portanto, a formulação de uma matriz de transição **ÿÿÿÿ** = **{p11p12.....p1K.....pK1.....pKK}** é feita para explicar uma única instância desse processo aleatório. Cada elemento genérico ***pij*** da matriz de transição representa as chances de passar do estado *i* para j, com valores de **i, *j*** ÿ **[0, 1, ...K].** A estimativa de ***pij*** é dada como a razão cujo numerador é a contagem do número de vezes ***que*** o estado *i* foi seguido do estado ***assim*** nas séries. O denominador é o número total de vezes ***que*** atingiu qualquer estado no espaço de estados. Formalmente,

$$pij = P_{\{S(t+1) = s^{o}_{j} \mid S(t) = s^{o}_{eu}\}} \tag{24}$$

A estimativa de ***pij*** é dada como,

$$pij = \frac{n(\text{então}^{\ddot{y}\ddot{y}\ \text{então}}_{j})}{\sum_{j=1}^{\ddot{y}K} n(s^{o}_{-\ddot{y}\ddot{y}\ s}\ {}^{o}_{j})} \quad \ddot{y}i, j\ \ddot{y}\ [0, 1, 2, ..., K] \tag{25}$$

$$\sum_{j=1}^{\ddot{y}K} pij = 1\ \ddot{y}i$$

Com Eq. 25 a probabilidade máxima de probabilidades de transição é estimada, ou seja, ela informa qual estado é mais provável de ser alcançado a partir do estado atual, durante *t* para ***(t*** + 1). Se o numerador da equação resultar **0** ***ÿj*** = **(0, 1, 2, ...,** K), então ***pij*** = **1** quando ***i*** = ***j*** e 0 quando ***i*** ÿ j.

- **Formação e atualização da Matriz de Transição(TM) e diagrama de estado correspondente:** Aqui, foi dado um exemplo para mostrar a formação da matriz de transição a partir de uma sequência de estados. Além disso, foi descrito que dependendo do estado atual, como a matriz de transição é atualizada. Seja **{s1s2s3}** o espaço de estados e a sequência ocorrida é < **s1,s2,s1,s3,s3,s2,s2,s1,s3,s2,s3,s2,s2,s1,s2,s3,s1,s1** > . Assim, a Fig. 3 mostra a matriz de transição inicial formada a partir desta sequência. Agora, o estado atual ocorrido na sequência dada é **s1** e o estado **s2** foi alcançado a partir de **s1** no próximo passo de tempo. Assim, para incorporar esta nova transição nos dados do histórico, a matriz de transição é atualizada de forma que todos os denominadores da linha de s1 sejam incrementados em 1 e o valor da coluna de **s2** associado à linha **s1** seja incrementado em 1.
  A matriz de transição inicial e atualizada foi mostrada na Fig. 4 junto com seus diagramas de transição de estado correspondentes.

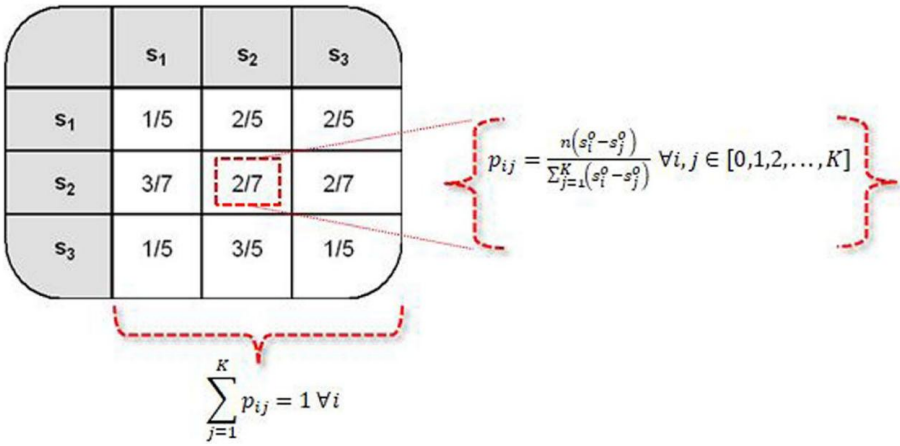


Fig. 3 Matriz de Transição **(ÿÿÿÿ)** para a sequência de estados < **s1,s2,s1,s3,s3,s2,s2,s1,s3,s2,s3,s2,s2,s1,s2,s3,s1,s1** >

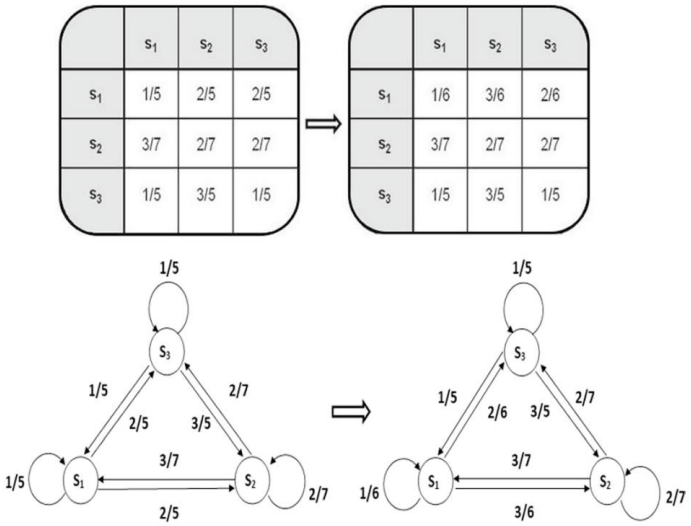


Fig. 4 Matriz de transição inicial e atualizada e diagramas de estado correspondentes

De acordo com o exemplo acima para K número de estados a matriz de transição será uma matriz ***(K*** × K) onde cada uma das linhas se refere ao estado presente na sequência e cada coluna se refere a um dos K estados possíveis que podem ser alcançado no próximo carimbo de hora. Portanto, é um processo aleatório conectado com a teoria da estimativa de máxima possibilidade. Os valores dos elementos de uma linha sempre 1, pois é a soma das probabilidades de transição do estado atual para qualquer estado possível no espaço de estados.

A matriz de transição é atualizada recursivamente com a chegada de novos dados.

Agora, para prever o estado no tempo (t+1) é importante conhecer a distribuição das probabilidades do estado no tempo t. É um vetor coluna para o passo de tempo t que contém o número K de linhas. O valor em cada linha do vetor coluna é a probabilidade de estar no k- ***ésimo*** estado no tempo t ***ÿk*** ÿ **[0, 1, ...,K].** Formalmente, o vetor é representado usando, ***V(t)*** = [v0(t) , v1(t) , ..., vK(t)], onde cada um dos elementos em geral é indicado por, vi(t)) e estimado como,

$$v_{i(t)} = \boldsymbol{P\{S(t) = s_i\}\ \ddot{y}i\ \ddot{y}\ [0, 1, ..., K]} \quad (26)$$

Embora o modelo de previsão baseado em MCFO aborde problemas complexos do mundo real, um desafio importante é escolher o intervalo apropriado para transformar um conjunto de valores em estados. É muito intuitivo que um intervalo maior resultará em uma previsão mais precisa, pois tanto o estado real quanto o estado de previsão podem estar dentro do mesmo intervalo, mas pode haver um erro significativo em relação ao valor real. Portanto, é desejável que a faixa de estados seja mínima. Novamente, como a cadeia de markov trabalha com estados, para compará-la com outras classes de modelos de séries temporais que lidam com valores reais, é necessário mapear os estados para valores reais para calcular as métricas de desempenho. Outro problema com esta classe de modelo é a ordem da cadeia. A ordem da cadeia depende do número de dados observados recentemente levados em consideração para influenciar a probabilidade de estar no estado atual. Aumentando o

ordem da cadeia pode aumentar a precisão, porém o cálculo do TM ficará mais complexo.

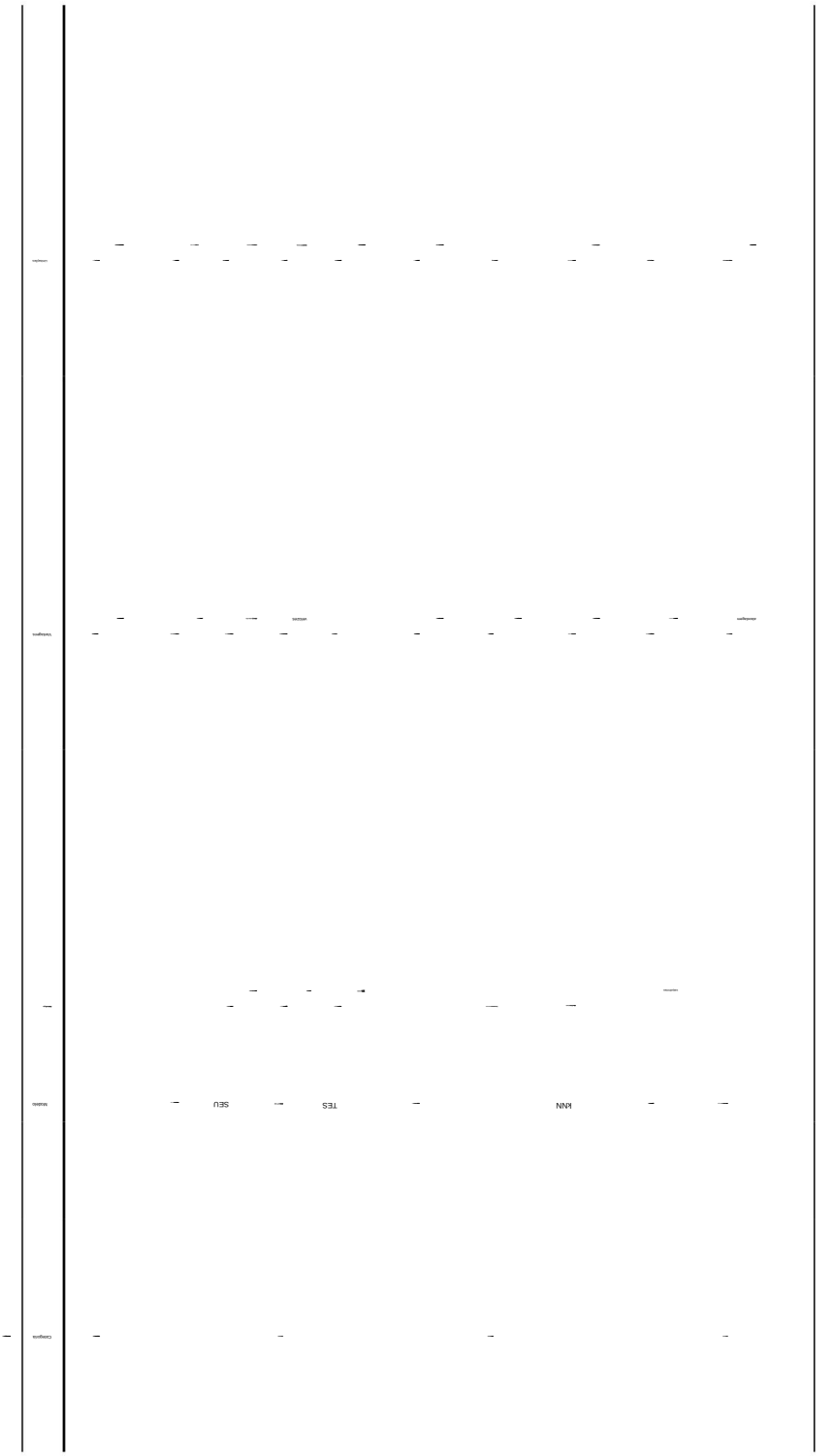
Um breve resumo dos modelos de previsão de séries temporais que discutimos nesta seção é fornecido na Tabela 2. O resumo dos modelos selecionados é fornecido em termos de sua estratégia de trabalho, vantagens e limitações. Embora um total de dez modelos sejam selecionados correspondendo às quatro categorias (ou seja, modelos baseados em regressão, modelos baseados em ES, modelos baseados em ML e DL e modelos baseados em MC), existem vários outros esquemas de previsão semelhantes disponíveis na literatura. Para modelos baseados em regressão, ARIMA e SARIMA são apenas implementados neste estudo, porque são adequados para séries temporais univariadas como dados **PM2.5 .** Na literatura, outros métodos semelhantes baseados em regressão, como ARIMAX, SARIMAX, etc, estão disponíveis para previsão de séries temporais. No entanto, ARIMAX e SARIMAX são mais adequados para previsão enquanto variáveis explicativas adicionais estão presentes (ou seja, séries temporais multivariadas), portanto, não as incluímos em nosso estudo. Neste trabalho usamos métodos baseados em deep learning como LSTM, embora vários outros métodos de deep learning como CNN, RNN, etc sejam amplamente utilizados para previsão de séries temporais. No entanto, o LSTM pode abordar melhor as dependências na sequência temporal dos dados. Da mesma forma, existem modelos de Cadeia de Markov de ordem superior para previsão de tempo, eles podem ter um desempenho melhor do que Cadeia de Marcov de primeira ordem. Mas, tem crescente complexidade de implementação com o aumento da ordem da cadeia.

4 Avaliação

Nesta seção, primeiro um total de cinco conjuntos de dados de séries temporais univariadas de **PM2.5** foram estudados em termos de algumas estatísticas importantes. Os dados foram coletados de cinco estações de monitoramento instaladas pelo governo na Índia. Em seguida, as unidades de medida de desempenho foram discutidas brevemente com base nas quais o estudo comparativo foi feito. Eventualmente, os modelos mencionados foram implementados ao longo dessas cinco séries temporais para fazer previsões. A discussão comparativa de suas exatidões de desempenho foi fornecida adiante. Além disso, gráficos de previsão e gráficos comparativos de métricas de desempenho foram mostrados com referência a um conjunto de dados de exemplo. Todos os experimentos necessários foram feitos em linguagem python utilizando bibliotecas e pacotes necessários. O sistema consiste no sistema operacional Windows 10 de 64 bits, processador de 2,00 GHz e 8,00 GB de RAM.

4.1 Dados

Os dados usados para este estudo foram coletados de cinco diferentes estações de monitoramento de poluição do ar instaladas pelo governo na Índia. Na Fig. 5, as localizações geográficas dessas estações de monitoramento foram fornecidas. Um total de cinco conjuntos de dados, nomeadamente **DS1, DS2, DS3, DS4,** e ***DS5*** são criados correspondendo às estações de monitoramento localizadas em Bidhannagar, Fort William, Rabindra Bharati University, Rabindra Sarobar e Victoria respectivamente. Entre eles, o conjunto de dados de Victoria **(DS5),** Índia, foi considerado como o conjunto de dados de referência ao longo deste estudo para descrever a análise estatística e os gráficos correspondentes. Esses conjuntos de dados são as séries temporais de concentração de um dos principais poluentes do ar **PM2.5** que são registrados em ÿÿgÿm3 . Todas essas cinco séries temporais contêm dados de um ano registrados com um intervalo de meia hora. A duração de um ano é de julho de 2018 a junho de 2019. Além disso, eles foram registrados em cenário do mundo real e possuem valores ausentes. Assim, eles também foram pré-processados pela imputação de valores ausentes usando o método de imputação média (Noor



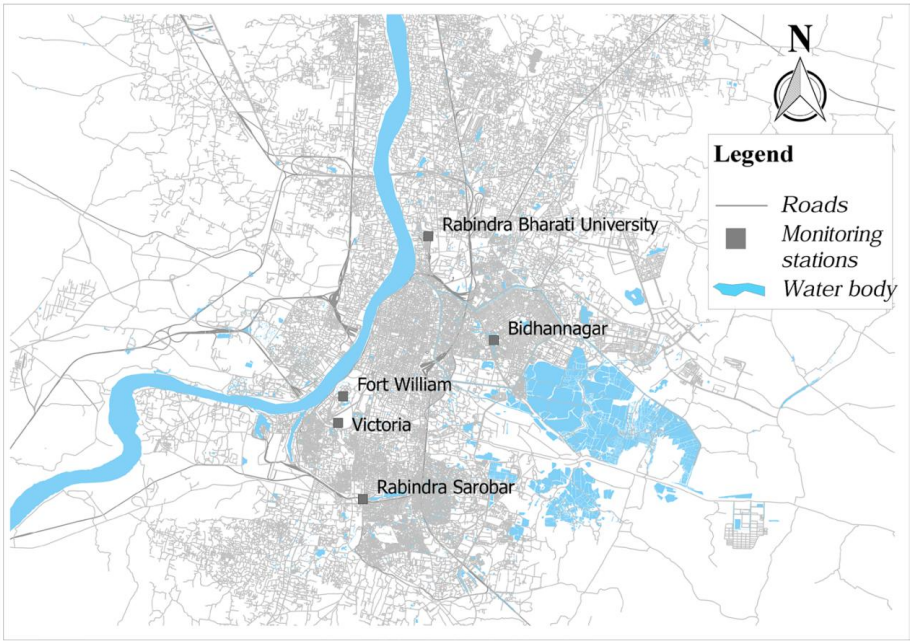


Fig. 5 Localizações das estações de monitoramento da poluição do ar

et ai. 2014; Gómez-Carracedo et al. 2014). A Fig. 6a, b) apresenta respectivamente o gráfico de dados **PM2.5** pré-processado e o gráfico de histograma correspondente para o conjunto de dados de referência. Em todos esses conjuntos de dados, picos incomuns repentinos foram encontrados durante o festival Diwali (um festival de luzes e fogos de artifício) na Índia. Além disso, a faixa de **PM2,5** foi considerada a mais alta durante o inverno e a mais baixa durante as monções. Esses efeitos do mundo real tornam os dados interessantes e têm um efeito mais sustentável nos resultados da previsão.

Para descrever a natureza e a variabilidade da série temporal **PM2.5** de cinco estações de monitoramento localizadas de forma diferente, suas localizações e resumo estatístico foram fornecidos na Tabela 3. As medidas estatísticas incluem valores mínimos e máximos, média, desvio padrão, intervalos interquartis, coeficientes de assimetria e curtose dos dados. Os coeficientes de assimetria são medidos para examinar a falta de simetria nos dados da série temporal. Isso é

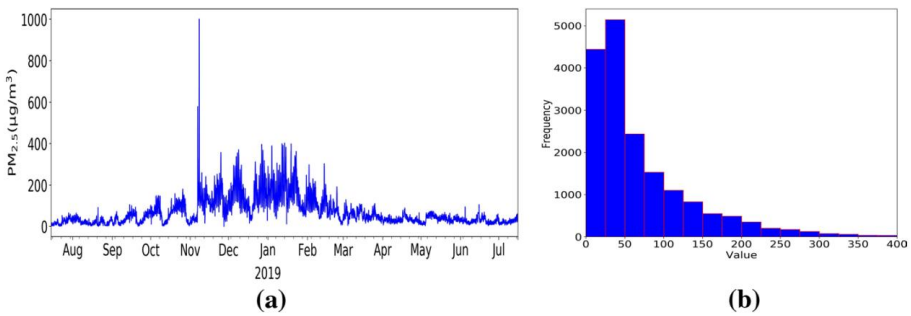


Fig. 6 a Gráfico de dados do conjunto de dados de referência de Victoria, Índia *(DS5)* para o período de julho de 2018 a junho de 2019 e **b** gráfico de histograma correspondente representando a concentração de **PM2,5**

[illegible]	DS1	DS2	DS3	DS4	DS5	Min.	0.25	0.3	0.8	1.02	1.02
						Max.	436.44	856.3	815	796.05	666.8
						—	70.1	43.27	56.54	53.68	42.08
						Std.Dev.	66.5	40.75	51.14	47.72	35.92
						25%	25	13.28	21	19.62	13.9
						50%	45.2	29.2	41.55	34.91	31.35
						75%	93.4	66.4	82.8	76.06	61.46
						Classe/pa	1.82	3.92	2.2	2.14	2.38
						Classe	3.59	54.78	12.1	12.33	21.53

sensível à flutuação aleatória, enquanto a curtose descreve se os dados da série temporal são fortemente caudados (alto valor de curtose) ou não. As medições estatísticas na Tabela 3 mostram que todos esses cinco conjuntos de dados possuem assimetria positiva e a curtose não é tão alta. Novamente, os valores de média e desvio padrão são quase iguais, o que indica que os conjuntos de dados não são sensíveis a incertezas.

Junto com isso, as variações dos conjuntos de dados durante os diferentes meses do ano também foram estudadas em termos de média, mediana, desvio padrão e comportamentos de pico dos dados. Os resultados representativos do conjunto de dados de referência ***(DS5)*** foram mostrados na Tabela 4. Os comportamentos dos picos mensais foram dados em termos de número médio de picos por dia, hora do dia em que o pico ocorre e altura e largura médias do pico. Altura e largura de um pico indicam o alcance máximo de **PM2,5** e duração do mesmo respectivamente. Na previsão de séries temporais univariadas, a relação dos valores com suas defasagens desempenha um papel muito importante, portanto, investigar tais variações nos dados com determinados intervalos pode ser interessante. A partir da observação pode-se observar que durante o mês de novembro a fevereiro, ou seja, durante a estação pós-monção e inverno na área de estudo, a altura média dos picos é muito alta o que por sua vez evidencia a má qualidade do ar. Além disso, as larguras dos picos indicam se esse pico é repentino ou não. Outra observação é que, durante o verão, ou seja, nos meses de abril e maio, a largura média do pico durante o amanhecer é muito baixa. Além disso, pode-se observar que a média e o desvio padrão não são fixos durante o período total de tempo dos dados. Esta é uma indicação clara de que há uma natureza não estacionária presente nos dados, o que é um fator importante a ser abordado ao lidar com séries temporais.

4.2 Métricas de desempenho

Neste estudo, os modelos foram comparados em termos de três critérios populares de medição de desempenho, a saber, raiz média do erro quadrado (RMSE), desvio da média

Tabela 4 Medidas de variabilidade da concentração de **PM2,5** durante julho de 2018 a junho de 2019 para o conjunto de dados de referência de Victoria, Índia *(DS5)*

Mês	Média	Mediana	Desv.Pad.	Nº de Picos/ Dia	Intervalos de Pico (Tempo)	Avg. Altura do Pico (ÿÿgÿm3)	Média Largura do pico (hora)
Julho	22.10	20.02	14.22	2	**01h00 - 4h30** **14h30 - 18h00**	**23.23** **25,63**	**6,47** **9.427**
agosto	21.43	19.08	12,63	1	11h00-17h00	31.12	20,48
setembro	38,70	38,04	22,36	1	00h00-04h30	51,59	7,89
Outubro	69,01	69,06	39,94	1	04.30-09.30	118,7	14,74
Novembro	122,64	112,60	96,33	1	00h00-04h30	172,88	9,69
Dezembro	152,91	138,50	74,07	1	00h00-04h30	233,59	7,88
janeiro	170,51	159,55	79,49	1	00h00-06h00	269,80	11.09
fevereiro	97,18	86,07	março 53,28	1	02.00-07.00	171,98	15,51
55,57	51,00		20,78	1	03.00-05.00	87,2	16,81
abril	32,04	30,02	12,58	2	**00h00 - 04h30** **15h30 - 20h00**	**28,53** **35.04**	**5,93** **16,93**
Maio	37,78	38,77	14,54	2	**00h00 - 04h00** **15h00 - 19h00**	**38,81** **50,36**	**4,47** **19.15**
Junho	27,87	25,00	13,85	1	13h30-19h30	33.03	19.115

porcentagem absoluta (MAPE) e média do desvio absoluto (MAE). RMSE é a medida da raiz da média do quadrado da diferença entre os valores reais e previstos. Ele basicamente mede a quantidade de concentração dos dados em torno da linha idealmente ajustada.
O MAPE descreve a precisão da previsão como um valor percentual. Funciona melhor quando não há zero ou extremos nos dados. Por outro lado, o MAE fornece a estimativa da média dos valores absolutos dos erros de previsão individuais para todos os casos no conjunto de teste. O objetivo é sempre minimizar essas métricas. Dentre esses três, o MAE é o mais fácil de entender, pois expressa diretamente o erro, enquanto que, para estimar o RMSE, são atribuídas altas penalidades a erros maiores, pois ele eleva ao quadrado os erros de previsão. Assim, o RMSE é útil nos casos em que se deseja evitar grandes erros de previsão.

A fórmula matemática dessas três métricas é dada abaixo, onde x_i^p denota o valor previsto considerando que, xo_i é o valor real para o i Tempo. , $\overline{Xo}$ é o valor médio cal Xp calculado para as séries previstas e observadas respectivamente e T é o número total de passos de tempo observados.

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i^p - xo_i)^2} \quad (27)$$

$$MAPE = \frac{1}{T}\sum_{i=1}^{T}\left|\frac{\overline{Xp} - \overline{Xo}}{\overline{Xo}}\right| \quad (28)$$

$$FÁCIL = \frac{1}{n}\sum_{i=1}^{n}|x_i^p - xo_i| \quad (29)$$

4.3 Resultados e discussão

Nesta seção são apresentados os resultados experimentais que foram obtidos pela aplicação das dez abordagens discutidas na Seção. 3.2 em cinco conjuntos de dados de séries temporais de **PM2.5** nomeadamente **DS1, DS2, DS3, DS4** e **DS5.** Agora, a previsão de séries temporais de várias etapas pode ser feita de duas maneiras. Uma é a estratégia recursiva onde os valores mais recentes são incorporados para prever o próximo valor. Outra é a estratégia de previsão direta, na qual um modelo separado é desenvolvido para prever em cada etapa de tempo. No trabalho atual foi utilizada a estratégia de previsão direta. Outra coisa importante durante a experimentação com modelos de previsão é escolher os parâmetros ideais. Para ARIMA e SARIMA, os parâmetros ótimos foram escolhidos usando o método auto.arima(). Ele determina automaticamente os melhores parâmetros por conta do menor valor de AIC. Os parâmetros ótimos para outros modelos foram encontrados através do método de busca em grade com base nos menores valores das métricas de erro. As configurações de parâmetros para os modelos para os quais o melhor desempenho foi alcançado em diferentes conjuntos de dados são fornecidos na Tabela 5. Observe que a literatura existente (Myung e Pitt 1997) ilustra que existem diferentes fatores que podem ser usados para analisar a complexidade de os modelos, como o número de parâmetros do modelo, o alcance do seu espaço de parâmetros, a forma funcional dos modelos, etc. Nesse contexto, a Tabela 5 também poderia ser usada para analisar a complexidade do modelo em termos do número total de parâmetros do modelo. Com base no número total de parâmetros pode-se concluir que a complexidade do modelo é maior para os modelos baseados em rede neural. Agora, na Tabela 6, o tempo de execução

(em segundos) dos modelos com tamanhos de dados variados (em termos de porcentagem e número de amostras de dados). A tabela reflete claramente a dependência dos modelos sobre a quantidade de dados que estão sendo usados. A partir dos resultados pode-se observar que os modelos baseados em suavização exponencial levam muito menos tempo enquanto que o SARIMA tem o maior tempo de execução. Também os tempos de execução para os modelos ARIMA, ANN e LSTM são signifcativamente altos. Nesse contexto, pode-se concluir que o tamanho dos dados é um fator muito importante para o tempo de execução dos modelos, principalmente para ARIMA, SARIMA, ANN, LSTM e MCFO.

Finalmente, a fim de comparar as capacidades dos modelos, eles foram implementados sobre os cinco conjuntos de dados de séries temporais. Os conjuntos de dados foram divididos em conjunto de treinamento composto por aproximadamente 90% dos dados e um conjunto de teste de aproximadamente 10% dos dados. É importante notar que o conjunto de teste geralmente contém de 25% a 30% dos dados, mas neste trabalho, como nosso problema se concentra no cenário de previsão de séries temporais de curto prazo, um pequeno conjunto de teste com 10% dos dados será perfeitamente servir ao nosso propósito. Agora, a validação walk-forward é utilizada neste trabalho. Na abordagem walk forward (WFA), primeiro uma janela abrangendo um determinado período de tempo no início é usada para treinar e otimizar o modelo. Outro segmento composto pelos dados logo após o final da janela é usado para validar o modelo. Em seguida, a janela é rolada e o processo é repetido até que o final dos dados de treinamento seja alcançado. A abordagem de validação walk forward acaba por gerar um resultado mais realista, particularmente no caso de dados de séries temporais, onde as informações são adicionadas continuamente ao longo do tempo. Depois que os modelos são treinados, o desempenho em termos de RMSE, MAE e MAPE é calculado no conjunto de teste separado. No caso do MCFO, por tratar de estados (constituídos por uma faixa de valores), os estados de previsão são primeiramente mapeados para valores reais. Assim, para o MCFO, uma gama de medições de precisão foi fornecida. A variação no desempenho é examinada para três casos diferentes de horizonte de previsão de curto prazo ou conjunto de teste consistindo nas últimas 48 amostras (1 dia), 336 amostras (1 semana) e 1440 amostras (1 mês). Também como foco principal do trabalho, as previsões foram realizadas com alta granularidade que foi definida para 30 minutos. A precisão da previsão em termos de RMSE, MAE e MAPE dos modelos referentes a cinco conjuntos de dados ***(DS1,* DS2, DS3, DS4** e **DS5)** e os três horizontes de previsão mencionados acima foram fornecidos na Tabela 7. Além disso, a discussão detalhada dos resultados é dado da seguinte forma:

- *Caso I (Horizonte de previsão de 1 dia)* Neste caso, a previsão de curto prazo granular alta consiste em 48 etapas fornecendo os valores esperados a cada 30 minutos por um dia. A precisão da previsão dos modelos para todos os cinco conjuntos de dados foi fornecida na Tabela 7.
  Os resultados mostram que todos os modelos atingem desempenhos aceitáveis. O LSTM produz o melhor desempenho para todos os conjuntos de dados em comparação com outros métodos em referência ao RMSE. No entanto, em termos de MAE e MAPE, os modelos como DT, kNN e MCFO (com valor mínimo possível) dominam ligeiramente sobre LSTM para alguns conjuntos de dados.
  Na Fig. 7, os erros de previsão (em termos de RMSE, MAE e MAPE) dos modelos foram descritos para este caso com base no conjunto de dados de referência **(DS5).**
- *Caso II (Horizonte de previsão de 1 semana)* Aqui a previsão de curto prazo consiste em 336 etapas que fornecem valores futuros com alta granularidade de 30 minutos para uma semana. Os resultados (ver Tabela 7) mostram que todos os modelos atingem uma precisão suficientemente boa e os valores de erro são próximos uns dos outros. No entanto, o kNN produz consistentemente os melhores resultados (em termos de valores de RMSE) em comparação com outros modelos com todos os cinco conjuntos de dados. Novamente, DT e MCFO (com valor mínimo possível) dominam para alguns dos conjuntos de dados em termos de MAE e MAPE. Os resultados contra o conjunto de dados de referência ***(DS5)*** para este caso (ou seja, Caso II) são mostrados na Fig. 8 para avaliação visual.

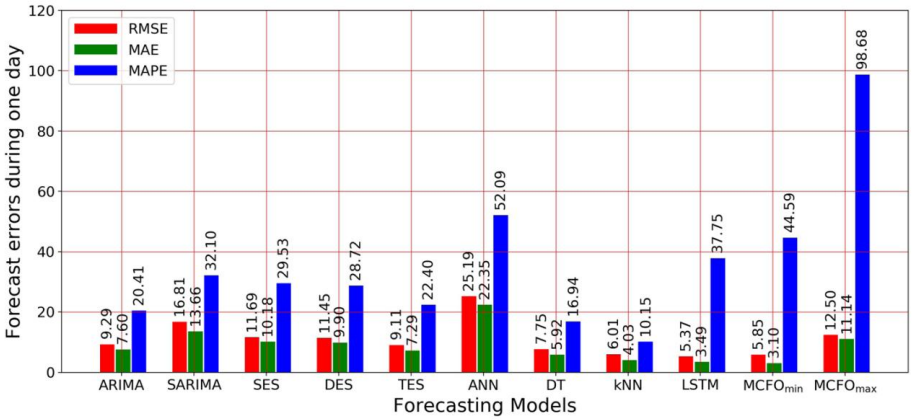


Fig. 7 Erros de previsão dos modelos de séries temporais encontrados em relação às séries temporais de referência de Victoria, Índia *(DS5)* para um dia

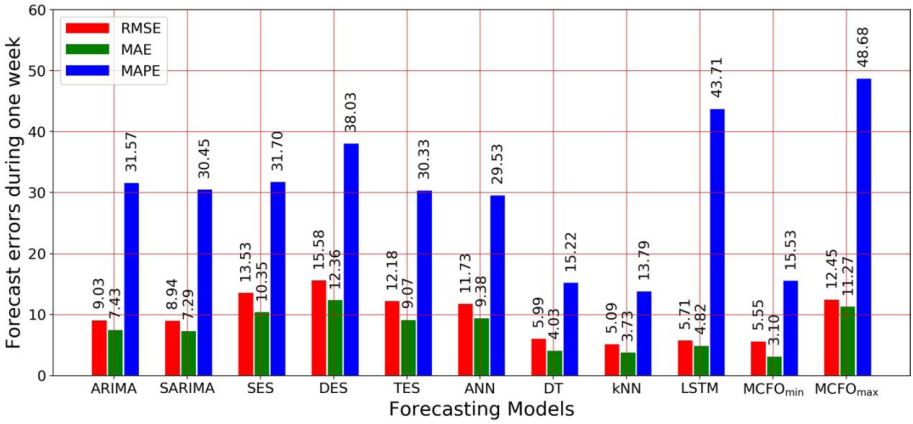


Fig. 8 Erros de previsão dos modelos de séries temporais encontrados em relação às séries temporais de referência de Victoria, Índia *(DS5)* por uma semana

• *Caso III (Horizonte de previsão de 1 mês)* Este é o caso com horizonte de previsão mais longo dando valores possíveis de cada 30 minutos de previsão granular alta para um mês.
Pode-se observar na Tabela 7 que os modelos ML e DL obtiveram bons resultados com o kNN alcançando a melhor precisão para todos os conjuntos de dados em termos de RMSE. Também os modelos como SARIMA, SES e MCFO atingem uma precisão bastante boa, enquanto os valores de erro são significativamente altos para outros modelos como ARIMA, DES e TES. O desempenho dos modelos em relação ao conjunto de dados de referência ***(DS5)*** para este caso (ou seja, Caso III) foi descrito na Fig. 9 para avaliação visual.

Na Fig. 10 , gráficos de saídas reais e de previsão dos modelos foram dados no caso de uma semana, ou seja, 336 etapas de previsão para o conjunto de dados de exemplo da localização de Victoria. Os resultados mostram que ARIMA e DES prevêem ao longo da inclinação, lidando com a tendência. O SARIMA mostra um pé médio, enquanto o TES mostra um bom pé ao lidar com sucesso com as variações sazonais. SES mostra uma previsão de gordura mostrando o valor do componente de nível. Por outro lado, o MCFO

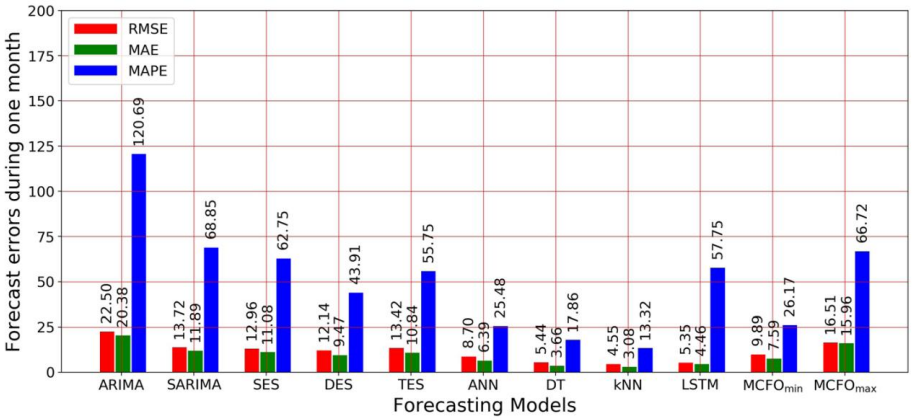


Fig. 9 Erros de previsão dos modelos de séries temporais encontrados em relação às séries temporais de referência de Victoria, Índia *(DS5)* por um mês

funciona definindo estados. Portanto, para compará-lo com outros modelos mapeamos os estados de previsão em valores reais. Ressalta-se que o MCFO obteve desempenho satisfatório. Os modelos ML e DL aprendem as dependências subjacentes da sequência. Portanto, para esses dados granulares de grande escala, a família de modelos de aprendizado de máquina ANN, DT, kNN mostra bons resultados que seguem claramente a forma dos dados reais para todo o horizonte de previsão. No entanto, o modelo de aprendizado profundo LSTM produz o melhor resultado. Este modelo prevê bastante próximo dos valores reais para todo o horizonte de previsão.

Uma coisa a ser observada é que, com o aumento do comprimento do horizonte de previsão, ARIMA e TES apresentam degradação de desempenho, enquanto o desempenho dos modelos ML e DL são mais consistentes em todos os casos de horizonte de previsão. Novamente, como MAPE é representado como uma razão, é sensível a zeros e valores extremos. Assim, com alto erro absoluto e valores de previsão mais altos, o valor de MAPE torna-se alto. Isso explica o alto aumento no MAPE de ARIMA, ANN e TES em alguns casos. Agora, se o desempenho por categoria dos modelos for investigado em relação aos cinco conjuntos de dados, sob os modelos baseados em regressão, o SARIMA domina sobre o ARIMA com o aumento do comprimento do horizonte de previsão. Sob os modelos baseados em ES (ou seja, SES, DES e TES), todas as três abordagens mostram valores de precisão próximos uns dos outros. Finalmente, a categoria de modelos ML e DL é a categoria de melhor desempenho sob a qual LSTM e kNN mostram os melhores resultados no caso de horizonte de previsão mais curto e mais longo, respectivamente. Portanto, a partir deste estudo, pode-se afirmar que os modelos ML e DL são mais aplicáveis para previsão de curto prazo altamente granular em relação à forma real das variações da série temporal.

5 Conclusão e escopo futuro

Este estudo investiga modelos de séries temporais de diferentes categorias e apresenta um estudo comparativo para previsão granular alta de **PM2.5.** O foco principal é determinar qual abordagem pode ser adotada para esse tipo de dado, como sua precisão é influenciada pelo comprimento do horizonte de previsão e quais são as desvantagens dessas abordagens. Para cumprir esse motivo, quatro categorias diferentes de modelos com um total de dez abordagens de previsão, incluindo ARIMA, SARIMA, SES, DES, TES, ANN, DT, kNN, LSTM e MCFO foram

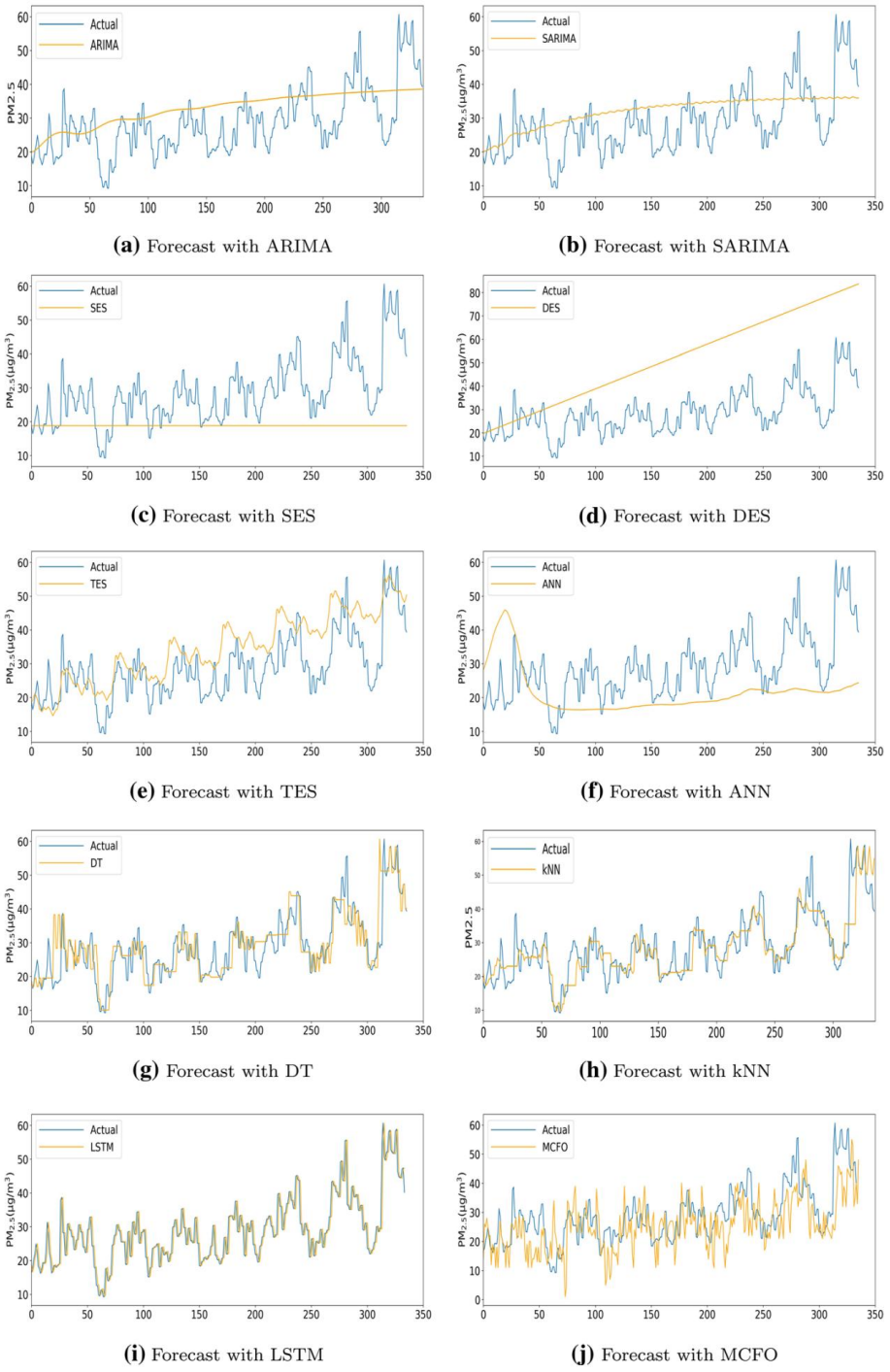


Fig. 10 Gráficos reais e previstos da concentração de **PM2,5** para 336 etapas de tempo feitas por modelos de previsão de séries temporais em relação às séries temporais de referência de Victoria, Índia *(DS5)*

Modelo	Parámetros	DS1	DS2	DS3	DS4	DS5
ARIMA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(10,0,4)
SARIMA						[illegible]
SEU	(ÿÿ)	0,7	0,2	0,8	0,3	0,2
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(0,15)
TES						[illegible]
ANN	[illegible]	18	16	15	18	20
TD	Profundidad [illegible]	3	4	2	2	2
kNN		4	5	5	3	3
LSTM	[illegible]	4	5	3	6	4
MCFO	[illegible]	10	10	10	10	10

Tabela 6 Tempo de execução (em segundos) dos modelos em relação ao tamanho variável do conjunto de dados (25%, 50%, 75%, 100%)

Nome do modelo	Tamanho do conjunto de dados (Número de tuplas)			
	25% (4380)	50% (8760)	75% (13140)	100% (17520)
ARIMA	120,69	200,51	275,54	499,88
SARIMA	2382,90	2551,21	2824,92	3050.11
SEU	0,41	0,54	0,65	1,87
A PARTIR DE	0,83	1,42	2,71	3,78
TES	1,71	4,47	5,93	6h30
ANN	196,76	716,40	1438,42	1752,20
TD	3,73	7,87	9,49	10,63
kNN	3,35	6,92	12,88	17,63
LSTM	81,27	276,50	314,26	325,19
MCFO	62,30	67,31	97,36	147,73

amplamente estudado e aplicado para prever séries temporais **PM2.5 .** Dados de cinco estações de monitoramento de poluição do ar instaladas pelo governo foram levados em consideração para a construção de modelos e avaliação de desempenho. Foram realizadas análises estatísticas desses conjuntos de dados do mundo real que mostram grande variação na concentração de **PM2,5** ao longo do ano, com maior faixa de pico encontrada durante o inverno. O estudo foi realizado com alta granularidade de trinta minutos para três diferentes comprimentos de horizontes de previsão (por exemplo, um dia, uma semana e um mês). Todos os modelos foram ajustados com uma ampla gama de conjuntos de parâmetros e os melhores modelos foram selecionados com base em medidas de precisão, como RMSE, MAE e MAPE. Também foi estudada a variação dos tempos de execução dos modelos em relação aos tamanhos de dados variados.

Os resultados experimentais mostram que todos os modelos atingem uma precisão bastante boa em dados do mundo real, embora as abordagens ML e DL dominem o restante dos métodos. Mais especificamente, para um horizonte de previsão mais curto de um dia, o LSTM oferece a melhor precisão em termos de RMSE para todos os conjuntos de dados. Por outro lado, com um horizonte de previsão mais longo (ou seja, uma semana e um mês), kNN foi considerado o modelo com melhor desempenho com todos os conjuntos de dados. O modelo MCFO mostra um desempenho bastante bom, cuja aplicação foi muito menor nesta área. Além disso, com um horizonte de previsão mais longo, a degradação acentuada do desempenho pode ser encontrada no caso do ARIMA. Por outro lado, embora SARIMA, ANN e LSTM apresentem boa acurácia com um horizonte de previsão maior, os tempos de execução desses modelos são bastante elevados. Modelos baseados em suavização exponencial (SES, DES, TES) e alguns outros modelos, como kNN e DT, levam muito menos tempo de execução em comparação com o restante dos métodos.

Os modelos deste estudo foram usados para prever o valor de **PM2,5** a cada 30 minutos. Foram testados três casos de horizonte de previsão que agregam valor a este trabalho. Além disso, a análise estatística indica grande variação ao longo do ano. Os resultados deste estudo podem ajudar os órgãos de monitoramento da poluição a escolher o modelo apropriado e utilizá-lo para detectar a concentração de **PM2,5** e tomar as medidas necessárias para reduzir seu efeito adverso na saúde humana. Eventualmente, pode ser recomendado a partir deste estudo considerar os modelos kNN, DT e LSTM como a primeira escolha ao trabalhar com dados de séries temporais de alta granularidade PM2.5. Por outro lado, como os modelos como SARIMA ou TES também alcançam resultados promissores para o horizonte de previsão maior, eles poderiam ser explorados para o

Horizonte	Métrica	MCPO	LSTM	KNN	TD	ANN	TES	ARIMA	SARIMA	SELI	ARIMA	SARIMA	ARIMA
DS3	RMSE	3,18	1,55	1,59	1,71	1,09	5,10	1,84	6,10	2,24	5,82	7,48	8,98
	MAPA	2,32	1,32	1,16	1,45	13,99	4,52	1,37	5,44	1,95	5,22	7,48	6,37
	RMSE	999,73	51,09	22,54	26,93	322,04	106,11	28,46	50,77	132,77		27,79	25,05
	MAPA	3,56	2,87	2,31	2,84	8,83	7,23	8,12	7,69	5,22	5,99	8,74	8,41
	RMSE	1,95	2,51	1,57	1,87	6,15	5,63	6,73	6,17	4,29	4,92	8,05	6,73
	MAPA	21,34	70,21	13,23	19,8	78,23	50,60	59,19	53,03	62,29		44,28	63,50
	RMSE	3,53	3,25	2,85	3,45	5,45	6,56	8,32	6,45	6,57	14,59	6,89	26,53
	MAPA	24,21	5,68	5,45	5,68	7,3	4,81	6,27	5,01	5,37		6,74	
DS4	MAPA	228,45	56,70	50,09	55,98	52,28	43,77	45,47	53,57	65,72	169,12	50,98	
	RMSE	999,43	82,42	16,91	19,18	42,97	55,69	45,47	53,57	65,72	169,12	55,98	
	MAPA	8,98	10,59	9,43	8,98	7,86	14,11	7,48	8,01	7,35	6,10	7,48	
	RMSE	6,37	7,48	6,66	6,24	4,94	11,91	4,42	4,79	2,93	5,44	8,32	
	MAPA	25,05	27,79	25,58	24,20	21,84	47,58	22,69	24,48	29,35	132,77		
	RMSE	6,73	4,73	6,34	8,05	5,53	8,86	2,76	2,34	2,51	4,92	2,83	
	MAPA	63,50	37,32	35,71	44,28	34,56	54,47	17,25	14,93	56,55			
	RMSE	6,41	6,46	8,74	10,55	7,74	22011	4,99	4,08	4,12	5,99	4,34	
	MAPA	24,21	5,68	5,45	5,68	7,3	4,81	2,33	1,9	3,36			
	RMSE	8,98	10,59	9,43	8,98	7,86	14,11	7,48	8,01	7,35	6,10	7,43	
	MAPA	6,37	7,48	6,66	6,24	4,94	11,91	4,42	4,79	2,93	5,44	8,32	
	RMSE	25,05	27,79	25,58	24,20	21,84	47,58	22,69	24,48	29,35	132,77		
	MAPA	8,41	6,46	8,74	10,55	7,74	22011	4,99	4,08	4,12	5,99	4,34	
	RMSE	6,73	4,73	6,34	8,05	5,53	8,86	2,76	2,34	2,51	4,92	2,83	
	MAPA	63,50	37,32	35,71	44,28	34,56	54,47	17,25	14,93	56,55			
	RMSE	26,53	6,89	6,74	6,89	9,23	6,56	4,06	3,13	4,2	14,59	4,09	
	MAPA	24,21	5,68	5,45	5,68	7,3	4,81	2,33	1,9	3,36			
	RMSE	228,45	56,70	50,09	55,98	52,28	43,77	17,22	13,66	59,73	169,12	16,74	
	MAPA	999,43	82,42	16,91	19,18	42,97	55,69	45,47	53,57	65,72	169,12	55,98	
	RMSE	8,98	10,59	9,43	8,98	7,86	14,11	7,48	8,01	7,35	6,10	7,43	
	MAPA	6,37	7,48	6,66	6,24	4,94	11,91	4,42	4,79	2,93	5,44	8,32	
	MAPA	25,05	27,79	25,58	24,20	21,84	47,58	22,69	24,48	29,35	132,77		
	RMSE	999,73	51,09	22,54	26,93	322,04	106,11	28,46	50,77	132,77			
	MAPA	8,41	6,46	8,74	10,55	7,74	22011	4,99	4,08	4,12	5,99	4,34	
	RMSE	6,73	4,73	6,34	8,05	5,53	8,86	2,76	2,34	2,51	4,92	2,83	
	MAPA	63,50	37,32	35,71	44,28	34,56	54,47	17,25	14,93	56,55			
	RMSE	26,53	6,89	6,74	6,89	9,23	6,56	4,06	3,13	4,2	14,59	4,09	
	MAPA	24,21	5,68	5,45	5,68	7,3	4,81	2,33	1,9	3,36			
	MAPA	228,45	56,70	50,09	55,98	52,28	43,77	17,22	13,66	59,73	169,12	16,74	

conjuntos de dados com sazonalidade clara. Além disso, para dados categóricos, o MCFO pode ser explorado. Portanto, tal estudo comparativo com resultados experimentais rigorosos em múltiplas séries temporais pode ser uma referência útil para os pesquisadores desta área de pesquisa.

Como nosso trabalho futuro, gostaríamos de explorar outros métodos de ML e DL nesses dados, pois essa categoria de modelos alcançou os melhores resultados. Séries temporais de outros poluentes também estão planejadas para serem coletadas e com base em suas correlações a área de previsão de séries temporais multivariadas pode ser explorada. Além disso, será investigado se um melhor desempenho pode ser alcançado para um horizonte de previsão mais longo. Agora, neste estudo, as séries temporais consistem principalmente em dados recentes de um a dois anos, porque se descobriu que são suficientes para capturar a variabilidade em **PM2,5.** No entanto, pode ser eficaz considerar um intervalo maior de séries temporais consistindo de vários anos no caso de tomada de decisão de longo prazo. Assim, um grande conjunto de dados com vários anos de dados históricos está planejado para ser coletado. Um plano adicional é aplicar os modelos de previsão de dados de séries temporais espaço-temporais de **PM2.5** que são coletados por meio de sensoriamento participativo, pois eles produzem informações mais refinadas da dinâmica da poluição do ar urbano.

Agradecimentos Este trabalho de pesquisa é apoiado pelo projeto intitulado - Participatory and Realtime Pollution Monitoring System For Smart City, financiado pelo Ensino Superior, Ciência e Tecnologia e Biotecnologia, Departamento de Ciência e Tecnologia, Governo de Bengala Ocidental, Índia.

Declarações

Conflito de interesse Os autores declaram não ter conflito de interesse.

Referências

Al-Qahtani FH, Crone SF (2013, agosto). Regressão multivariada de k-vizinhos mais próximos para dados de séries temporais– Um novo algoritmo para prever a demanda de eletricidade no Reino Unido. Na conferência conjunta internacional de 2013 sobre redes neurais (IJCNN), IEEE. pp.1-8

Amato A, Calabrese M, Di Lecce V (2008, maio) Árvores de decisão em problemas de reconstrução de séries temporais. Em 2008 IEEE Instrumentation and Measurement Technology Conference, IEEE. pp.895–899

Bose R, Dey RK, Roy S, Sarddar D (2020) Previsão de séries temporais usando suavização exponencial dupla para prever os principais poluentes do ar ambiente. Tecnologia da informação e comunicação para o desenvolvimento sustentável. Springer, Cingapura, pp 603-613

Breiman L, Friedman JH, Olshen R, Stone CJ (1984) Classificação e árvores de regressão. Wadsworth & Livros e software avançados Brooks/Cole, Pacifc Califórnia

Brunelli U, Piazza V, Pignato L, Sorbello F, Vitabile S (2007) Previsão de dois dias à frente das concentrações máximas diárias de **SO2, O3, PM10, NO2,** CO na área urbana de Palermo, Itália. Atmos Environ 41(14):2967–2995

Caillault ÉP, Bigand A (2018, setembro). Estudo comparativo de métodos de previsão univariada para séries temporais meteorológicas. Em 2018, 26ª Conferência Europeia de Processamento de Sinais (EUSIPCO), IEEE, pp.2380-2384

Chelani AB (2005) Prevendo séries temporais caóticas de concentração de PM10 usando rede neural artificial. Int J Environ Stud 62(2):181–191

Choubin B, Zehtabian G, Azareh A, Rafei-Sardooi E, Sajedi-Hosseini F, Kïÿi Ö (2018) Previsão de precipitação usando modelo de árvores de classificação e regressão (CART): um estudo comparativo de diferentes abordagens. Environ Earth Sci 77(8):314

Chu HJ, Lin CY, Liau CJ, Kuo YM (2012) Identificando fatores de controle dos níveis de ozônio troposférico no sudoeste de Taiwan usando uma árvore de decisão. Ambiente Atmos 60:142–152

Das M, Ghosh SK (2018) Abordagens orientadas por dados para previsão de séries temporais meteorológicas: um estudo comparativo das técnicas de inteligência computacional de última geração. Reconhecimento de Padrão Lett 105:155–164

Dastorani M, Mirzavand M, Dastorani MT, Sadatinejad SJ (2016) Estudo comparativo entre diferentes modelos de séries temporais aplicados à previsão mensal de chuva em condições de clima semiárido. Nat Hazards 81(3):1811–1827

Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, Moncada-Herrera JA (2008) Um modelo híbrido ARIMA e redes neurais artificiais para previsão de material particulado em áreas urbanas: o caso de Temuco, Chile. Atmos Environ 42(35):8331–8340

Divina F, García Torres M, Goméz Vela FA, Vázquez Noguera JL (2019) Um estudo comparativo de métodos de previsão de séries temporais para previsão de consumo de energia elétrica de curto prazo em edifícios inteligentes. Energias 12(10):1934

Domaÿska D, Wojtylak M (2012) Aplicação de modelos de séries temporais difusas para a previsão de concentrações de poluição. Aplicativo de sistema especialista 39(9):7673–7679

Dong M, Yang D, Kuang Y, He D, Erdal S, Kenski D (2009) Previsão de concentração **PM2.5** usando mineração de dados de série temporal baseada em modelo semi-Markov oculto. Aplicativo de sistema especialista 36(5):9046–9055

Dragomir EG (2010) Previsão do índice de qualidade do ar usando a técnica do vizinho K-mais próximo. Bull PG Univ Ploiesti Ser Math Inform Phys LXII 1(2010):103–108

Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S (2014) Análise de séries temporais complexas de **PM10** e **PM2.5** para um local costeiro usando modelagem de rede neural artificial e agrupamento k-means. Atmos Environ 94:106–116

Freeman BS, Taylor G, Gharabaghi B, Thé J (2018) Previsão de séries temporais de qualidade do ar usando aprendizado profundo. J Air Waste Manag Ass 68(8):866–886

Gocheva-Ilieva SG, Ivanov AV, Voynikova DS, Boyadzhiev DT (2014) Análise de séries temporais e previsão de poluição do ar em pequenas áreas urbanas: uma abordagem SARIMA e análise fatorial. Avaliação de Risco Stoch Environ Res 28(4):1045–1060

Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D (2014) Uma comparação prática de métodos de imputação simples e múltipla para lidar com dados complexos faltantes em conjuntos de dados de qualidade do ar. Sistema de Laboratório Químico Intel 134:23–33

Haider SA, Naqvi SR, Akram T, Umar GA, Shahzad A, Sial MR, Khaliq S, Kamran M (2019) Modelo de previsão baseado em rede neural LSTM para a produção de trigo no Paquistão. Agronomia 9(2):72

Huang SF, Cheng CH (2008, julho). Previsão da qualidade do ar usando modelo de série temporal baseado em OWA. Em 2008 Conferência Internacional sobre Aprendizado de Máquina e Cibernética, IEEE. Vol. 6, pp. 3254–3259

Khaldi R, Chiheb R, El Afa A (2018, maio). Redes neurais feedforward e recorrentes para previsão de séries temporais: estudo comparativo. In Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications pp. 1-6

Kumar U, Jain VK (2010) ARIMA previsão de poluentes do ar ambiente **(O3,** NO, **NO2** e CO). Estoque Envi ron Res Risk Assessment 24(5):751–760

Lin K, Jing L, Wang M, Qiu M (2017, agosto). Um novo algoritmo de previsão da qualidade do ar de longo prazo baseado em kNN e NARX. Em 2017 12th International Conference on Computer Science and Education,IEEE. (ICCSE) pp. 343-348

Li X, Peng L, Yao X, Cui S, Hu Y, You C, Chi T (2017) Rede neural de memória de longo prazo para previsões de concentração de poluentes atmosféricos: desenvolvimento e avaliação de métodos. Poluição ambiental 231:997–1004

Mahajan S, Chen LJ, Tsai TC, 2017, agosto. Um estudo empírico de previsão **PM2.5** usando rede neural. Em, (2017) IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications. Computação em Nuvem e Big Data, Internet das Pessoas e Inovação em Cidades Inteligentes (SmartWorld/ SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE, pp 1–7

Meryem O, Ismail J (2014, outubro). Um estudo comparativo de algoritmos preditivos para previsão de séries temporais. Em 2014, terceiro colóquio internacional IEEE em ciência e tecnologia da informação (CIST), IEEE, pp.68–73

Momani PENM, Naill PE (2009) Modelo de análise de séries temporais para dados de chuva na Jordânia: estudo de caso para uso de análise de séries temporais. Am J Environ Sci 5(5):599

Myung IJ, Pitt MA (1997) Aplicando a navalha de Occam na cognição de modelagem: uma abordagem Bayesiana. Psicônico Revisão do Touro 4(1):79–95

Noor MN, Yahaya AS, Ramli NA, Al Bakri AMM (2014) Técnicas de imputação média para preencher as observações faltantes no conjunto de dados de poluição do ar. Trans Tech Publications Ltd., Kapellweg

Nury AH, Hasan K, Alam MJB (2017) Estudo comparativo dos modelos wavelet-ARIMA e wavelet-ANN para dados de séries temporais de temperatura no nordeste de Bangladesh. J King Saud Univ Sci 29(1):47–61

Oprea M, Mihalache SF, Popescu M (2016, maio). Um estudo comparativo de técnicas de inteligência computacional aplicadas à previsão de poluição atmosférica **PM2.5 .** Em 2016 6th International Conference on Computers Communications and Control (ICCCC), IEEE. págs. 103–108

Qiao W, Tian W, Tian Y, Yang Q, Wang Y, Zhang J (2019) A previsão de **PM2.5** usando um modelo híbrido baseado em transformada wavelet e um algoritmo de aprendizado profundo aprimorado. Acesso IEEE 7:142814–142825

Qin D, Yu J, Zou G, Yong R, Zhao Q, Zhang B (2019) Um novo esquema de previsão combinado baseado na CNN e LSTM para concentração urbana de **PM2,5 .** Acesso IEEE 7:20050–20059

Reddy V, Yedavalli P, Mohanty S, Nakhat U (2018). Ar profundo: Previsão da poluição do ar em Pequim, China

Roy S, Biswas SP, Mahata S, Bose R (2018, outubro). Previsão de séries temporais usando suavização exponencial para prever os principais poluentes atmosféricos. Em 2018 Conferência Internacional sobre Avanços em Computação, Controle de Comunicação e Redes (ICACCCN), IEEE. pp.679-684

Unnikrishnan R, Madhu G (2019) Estudo comparativo sobre os efeitos de parâmetros meteorológicos e poluentes eters na modelagem de RNA para previsão de **SO2.** SN Appl Sci 1(11):1394

Ventura LMB, de Oliveira Pinto F, Soares LM, Luna AS, Gioda A (2019) Previsão das concentrações diárias de **PM2.5** aplicando redes neurais artificiais e modelos Holt-Winters. Air Qual Atmos Health 12(3):317–325

Wang P, Zhang H, Qin Z, Zhang G (2017) Um novo modelo híbrido-Garch baseado em ARIMA e SVM para previsão de concentrações de **PM2,5 .** Atmos Pollut Res 8(5):850–860

Wu L, Gao X, Xiao Y, Liu S, Yang Y (2017) Usando o modelo cinza Holt-Winters para prever o índice de qualidade do ar para cidades na China. Nat Hazards 88(2):1003–1012

Yang L, Li C, Tang X (2020) O impacto do **PM2.5** na defesa do sistema respiratório do hospedeiro. Desenvolvimento de célula frontal Biol, 8, p.91

Ye L, Yang G, Van Ranst E, Tang H (2013) Modelagem de séries temporais e previsão do absoluto mensal global temperatura para a tomada de decisão ambiental. Adv Atmos Sci 30(2):382–396

Zainuddin Z, Pauline O (2011) Rede neural wavelet modificada na aproximação de funções e sua aplicação na previsão de dados de poluição de séries temporais. Appl Soft Comp 11(8):4866–4874

Zakaria NN, Othman M, Sokkalingam R, Daud H, Abdullah L, Abdul Kadir E (2019) Desenvolvimento de modelo de cadeia de Markov para previsão do índice de poluição do ar de Miri, Sarawak. Sustentabilidade 11 (19): 5190

Zhang G, Patuwo BE, Hu MY (1998) Previsão com redes neurais artificiais: o estado da arte. Int J Previsão 14(1):35–62

Autores e Afiliações

Rituparna Das1 · Asif Iqbal Middya1 · Sarbani Roy1

Rituparna Das
ritu92.prf@gmail.com

Asif Iqbal Middya
asifm.rs@jadavpuruniversity.in

[1] Departamento de Ciência da Computação e Engenharia, Universidade de Jadavpur, Kolkata, Índia