Environmental Engineering and Sustainable Development Journal

Volume 1, Issue 1, 2025, 39-50

Hybrid Approach for Refining and Reconstructing Air Pollution Data in Tehran Megacity Using Machine Learning

Mohammad Hashemzadeh [1], Faezeh Borhani [2*]

1. Ph.D. Candidate, Department of Environmental Engineering, Faculty of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran.
2. Assistant Professor, Department of Environmental Engineering, Faculty of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran.

Abstract

Air pollution is a major challenge in megacities, and its management depends on high-quality data. In developing countries like Iran, accessing reliable ground-based data is difficult. Satellite data offers a promising solution, but incomplete and outlier data remain significant challenges. This study addresses the issue of incomplete air pollution data in Tehran by employing a hybrid approach for data refinement and reconstruction. The dataset includes NO_2, CO, and O_3 pollutants from the Sentinel-5p sensor and meteorological variables from ERA5-land, covering December 2018 to March 2025. Results indicate a high prevalence of incomplete data for all pollutants in December due to weather conditions, with CO showing the highest level of incompleteness. A two-stage process using univariate Robust Z-score and multidimensional Isolation Forest (IF) was applied to identify outliers. Analysis revealed that cold months had the highest number of outlier data for pollutants, with NO_2 exhibiting the most outliers compared to other pollutants. The LightGBM algorithm was used to reconstruct missing values, yielding (r^2) of 0.61, 0.50, and 0.38 for NO_2, O_3, and CO, respectively. Despite data limitations and the absence of complex spatio-temporal algorithms compared to previous studies, the results, particularly for NO_2 and O_3, are considered satisfactory. This research demonstrates the potential of integrating satellite and meteorological data with machine learning to enhance air quality monitoring in data-scarce urban environments.

Review History

Received: May 10, 2025

Accepted: May 25, 2025

Keywords

Air pollution

Gap filling

Outliers detection

Isolation Forest

LightGB

* Corresponding Author Email: f.borhani@modares.ac.ir - ORCID: 0000-0003-2686-6702

مجله علمی مهندسی محیط زیست و توسعه پایدار

دوره ۱، شماره ۱، سال ۱۴۰۴ – صفحات ۳۹ تا ۵۰

رویکرد ترکیبی جهت پالایش و بازسازی داده‌های آلایندگی هوا در کلان‌شهر تهران با استفاده از یادگیری ماشین

محمد هاشم زاده[۱]، فائزه برهانی[۲*]

۱. دانشجوی دکتری، گروه مهندسی محیط زیست، دانشکده مهندسی عمران و محیط‌زیست، دانشگاه تربیت مدرس، تهران، ایران.

۲. استادیار، گروه مهندسی محیط زیست، دانشکده مهندسی عمران و محیط‌زیست، دانشگاه تربیت مدرس، تهران، ایران.

چکیده

آلودگی هوا از معضلات اصلی کلانشهرهاست و مدیریت آن به داده‌های باکیفیت وابسته است. در کشورهای در حال توسعه مانند ایران، دسترسی به داده‌های زمینی باکیفیت چالش‌برانگیز است. داده‌های ماهواره‌ای می‌توانند راه‌حل مناسبی باشند، اما وجود داده‌های ناقص و پرت از چالش‌های اصلی آن‌هاست. این مطالعه با هدف رفع چالش داده‌های ناقص آلودگی هوای تهران، از رویکرد ترکیبی برای پالایش و بازسازی داده‌ها استفاده می‌کند. داده‌ها مورد استفاده شامل آلاینده‌ها NO_2، CO و O_3 از سنجنده Sentinel-5p و متغیرهای هواشناسی ERA5-land در بازه دسامبر ۲۰۱۸ تا مارس ۲۰۲۵ است. نتایج حاکی از فراوانی داده‌های ناقص آلاینده‌ها در دسامبر به دلیل شرایط جوی است که CO بیشترین نقص را داراست. در ادامه برای یافتن داده‌های پرت از یک فرآیند دو مرحله‌ای شامل Robust Z-score تک بعدی و Isolation Forest (IF) چندبعدی استفاده شد. تحلیل داده‌ها نشان داد که ماه‌های سرد سال بیش‌ترین تعداد داده‌های پرت برای آلاینده‌ها را دارا بوده و NO_2 بیشترین تعداد را نسبت به سایر آلاینده‌ها به خود اختصاص داد. برای بازسازی مقادیر گمشده از الگوریتم LightGBM استفاده شد که نتایج ضریب تعیین (r^2) به ترتیب ۰/۶۱، ۰/۵۰ و ۰/۳۸ برای NO_2، O_3 و CO به دست آمد. با توجه به محدودیت داده‌های مورد استفاده و عدم استفاده از الگوریتم‌های پیچیده زمانی-مکانی در قیاس با مطالعات پیشین، نتایج بدست آمده به خصوص برای آلاینده‌های NO_2 و O_3 قابل قبول ارزیابی می‌شود. نتایج این پژوهش ظرفیت تلفیق داده‌های ماهواره‌ای و هواشناسی با یادگیری ماشین برای بهبود پایش کیفیت هوا در محیط‌های شهری با کمبودِ داده را به نمایش گذاشت.

تاریخچه داوری

دریافت: ۱۴۰۴/۰۲/۲۰

پذیرش: ۱۴۰۴/۰۳/۰۵

کلمات کلیدی

آلودگی هوا

تکمیل داده‌های ناقص

شناسایی داده‌های پرت

Isolation Forest

LightGBM

۱- مقدمه

آلودگی هوا امروز به‌عنوان یکی از چالش‌های عمده زیست‌محیطی در کلانشهرها مطرح است و شهر تهران نیز از این قاعده مستثنی نیست (Borhani et al. 2022; Zali et al. 2018). غلظت بالای آلاینده‌ها در سال‌های اخیر خسارات جانی و مالی فراوانی برجای گذاشته است؛ به‌ویژه تأثیرات منفی بر گروه‌های آسیب‌پذیر نظیر سالمندان و کودکان که در روزهای با کیفیت هوای «ناسالم» باید از تردد در فضای باز پرهیز کنند (Bayat et al. 2019; Aithal et al. 2023). تعطیلی موقت مدارس و دستگاه‌های اداری نیز نشانگر ابعاد وسیع این معضل و لزوم مدیریت هوشمندانه آن است.

برای تصمیم‌سازی در مواقع بحرانی و کاهش پیامدهای

* رایانامه نویسنده مسئول: f.borhani@modares.ac.ir – ORCID: 0000-0003-2686-6702

آلودگی، وجود داده‌های باکیفیت و پیوسته از آلاینده‌های کلیدی ضروری است (Rollo et al. 2023). به عنوان نمونه یکی از راهکارهای مناسب و در دسترس، تغییر ساعات کاری مدارس و ادارات در روزهای پیک آلودگی است تا از این طریق تا حدی از حجم ترافیک و در نتیجه انتشار آلاینده‌ها کاست؛ که برای این منظور باید بتوان از اطلاعات دقیق و قابل‌اعتماد استفاده کرد (Sokhi et al. 2022; Ramírez et al. 2019).

در کشورهای در حال توسعه مانند ایران، ایستگاه‌های زمینی پایش هوا با محدودیت پراکندگی جغرافیایی و نارسایی در برداشت داده‌ها روبه‌رو هستند، بدین منظور استفاده از داده‌های سنجش از دور با پوشش گسترده و پیوستگی زمانی مناسب می‌توانند کمبود ایستگاه‌های زمینی را جبران کنند (Holloway et al. 2021). در این میان، سنجنده TROPOMI بر روی ماهواره Sentinel-5P با ارائه ستون غلظت آلاینده‌ها، به دلیل در دسترس‌بودن مداوم، توزیع مکانی خوب و امکان تبدیل فیزیکی ستون‌ها به غلظت سطح زمین، به‌عنوان منبع جایگزین قابل‌اتکایی برای ایستگاه‌های زمینی پایش آلاینده‌ها مطرح است (Borhani et al. 2023; Tabunshchik et al. 2024).

اگرچه این سنجنده دارای نقاط قوت مناسبی است اما دارای یک ایراد اساسی می‌باشد و آن وجود گپ در مجموعه داده‌های آن می‌باشد که علت وجود این داده‌های ناقص پوشش ابری و یا مشکلات مربوط به بازیابی داده‌هاست (Schneider et al. 2021). همچنین به جز وجود داده‌های ناقص اولیه، برخی از داده‌ها به علل مختلف مانند شرایط جوی، خطاهای ابزار و شرایط توپوگرافی را می‌توان به عنوان داده‌های غیر قابل اعتماد در نظر گرفت که کیفیت پایین دارند و نیاز است تا قبل از استفاده نهایی شناسایی شوند (Schneising et al. 2023).

وجود این گپ در سری زمانی داده‌ها امکان استفاده‌های بعدی مانند تولید داده‌های زمینی یا بهبود کیفیت مکانی تصاویر را با چالش‌هایی روبه‌رو کرده است (Zheng et al. 2019). اگرچه تحقیقات مختلفی بر روی این داده‌های ناقص و نحوه تکمیل آن‌ها صورت گرفته است اما در این تحقیقات تمرکز کمتری بر روی فرآیند پالایش اولیه، شناخت الگو داده‌های ناقص، شناسایی داده های نامعقول و در نهایت بازسازی داده‌ها صورت گرفته است (Rollo et al. 2023; Hua et al. 2024). لازم به ذکر است که روش شناسایی داده‌های غیر مطمئن در این حالت اکثرا مبتنی بر رویکردهای تک بعدی می‌باشد و این در حالیست که ممکن است یک داده به تنهایی معقول باشد ولی زمانیکه در کنار سایر داده‌ها هم دسته بررسی می‌شود، نامطئن به نظر آید (Dongre et al. 2025). بنابراین لزوم استفاده از یک رویکرد ترکیبی در این حالت احساس می‌شود (Li et al. 2022; Schneising et al. 2023). بعلاوه اگرچه طیف بسیار وسیعی از روش‌ها شامل مدل‌سازی آماری (Wang et al. 2012)، درونیابی مکانی-زمانی (Wang et al. 2012) و حتی روش‌های مبتنی بر یادگیری عمیق (Appel 2024) برای تکمیل داده‌های ناقص آلاینده‌ها مورد استفاده قرار گرفته است، اما در این بین مقالات بیشتر متمرکز بر روی داده‌های زمینی بوده‌اند و استفاده صرف از داده‌های ماهواره‌ای در یک پژوهش کمتر مورد توجه بوده‌است.

علاوه بر این لازم به ذکر است که تمرکز اصلی مدل‌های بازسازی داده‌های ناقص بر روی استفاده از ویژگی‌های (همبستگی) زمانی و مکانی است و اکثرا این تحقیقات از این ویژگی‌ها در قالب درون‌یابی برای تکمیل داده‌های ناقص استفاده می‌کنند (Appel 2024). اما به خصوص هنگام کار با داده‌های ماهواره‌ای امکان ایجاد یک گپ زمانی در داده‌ها به علت شرایط جوی بالاست، به نحویکه ممکن است برای یک یا چند ماه برای یک منطقه خاص داده‌های با کیفیت با تعداد بسیار کمتری از حالت نرمال موجود باشند. به همین علت در این بازه‌ها امکان استفاده از روش‌های مبتنی بر همبستگی زمان یا مکانی وجود ندارد و می‌بایست استفاده از روش‌های جایگزین و مدل‌سازی بر اساس سایر متغیرها (به عنوان نمونه هواشناسی) مورد توجه قرار گیرد.

در این مقاله، برای حذف داده‌های نامطمئن، یک رویکرد جدید و ترکیبی پیشنهاد شده است. در این ساختار ابتدا با استفاده از مدل تک‌بعدی Robust Z-score داده‌های پرت برای آلاینده‌ها شناسایی می‌شود و در ادامه با الگوریتم IF (Dongre et al. 2025)داده ها در حالت چند بعدی نیز مورد بررسی قرار میگیرند. پس از تشخیص و حذف نمونه‌های مشکوک از مدل ساده اما قدرتمند LightGBM که در مقالات مربوط به تخمین آلاینده‌های هواشناسی پیشنهاد شده است (X. Yu et al. 2024; de la Cruz Libardi et al. 2024)، استفاده می‌شود. داده‌های ورودی به مدل

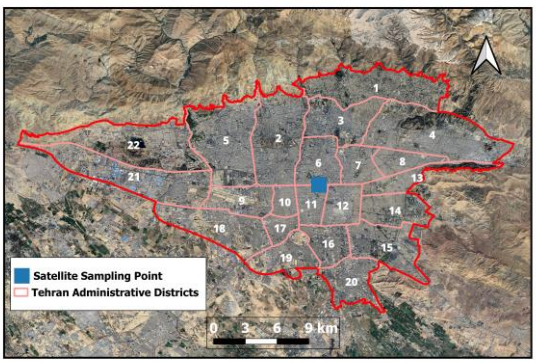
نیز با استفاده از متغیرهای هواشناسی بدست آمده از منبع ERA5-Land و اطلاعات زمانی (برای آموزش ویژگی‌های فصلی و ترافیکی به مدل) می‌باشد. در نهایت نیز به وسیله موارد فوق تخمینی از مقادیر جامانده سه آلاینده مهم NO_2، CO و O_3 انجام می‌پذیرد.

ساختار کلی مقاله در ادامه شامل ۳ بخش می‌باشد. در قسمت بعدی با عنوان مواد و روش‌ها، به بررسی محدوده مورد بررسی، الگوریتم‌های و داده‌های مورد استفاده پرداخته می‌شود. در بخش نتایج و بحث، نتایج بدست آمده تشریح می‌شود و از جنبه‌های مختلف مورد بررسی قرار می‌گیرد و در نهایت در بخش نتیجه‌گیری، خلاصه‌ای از نتایج و اقدامات آتی برای بهبود نتایج ارائه می‌شود.

۲- مواد و روش‌ها

۲-۱- محدوده مورد بررسی

تهران، پایتخت ایران، در عرض های جغرافیایی ۳۵ درجه و ۱۵ دقیقه تا ۳۵ درجه و ۴۸ دقیقه شمالی و طول های ۵۱ درجه و ۱۷ دقیقه تا ۵۱ درجه و ۳۳ دقیقه شرقی واقع شده است. این شهر با مساحتی تقریباً ۸۰۰ کیلومتر مربع شامل ۲۲ منطقه شهری است و از جنوب به دشت کویر، از شمال به رشته‌کوه البرز، از شرق به دره‌های جاجرود و از غرب به دره‌های کرج محدود می‌شود.آلودگی هوا در این حوزه عظیم شهری سال‌هاست که به یک مسئله جدی تبدیل شده و علی‌رغم ارائه راهکارهای گوناگون، بهبود قابل‌ملاحظه‌ای تجربه نکرده است (دلاور et al. 2020). نمایی از محدوده مورد مطالعه در شکل ۱ آورده شده‌است.



شکل ۱: محدوده مطالعاتی شهر تهران

۲-۲- داده‌ها

داده‌های مورد استفاده در این بخش شامل آلاینده‌ها و داده‌های کمکی هواشناسی می‌باشند. آلاینده‌های انتخابی در این مقاله شامل داده‌های NO_2، CO و O_3 می‌باشند که از محصول Sentinel-5P استخراج شده‌اند (Tabunshchik et al. 2024). داده‌های هواشناسی کمکی نیز شامل مولفه‌های باد شرقی/شمالی در ۱۰ متر، دمای نقطه شبنم، دمای متوسط، ماکزیمم و مینیمم روزانه، تابش دریافتی و فشار سطح از مجموعه ERA5-land می‌باشند (Jiménez-Navarro et al. 2024).

با توجه به داده‌های در دسترس برای مدل‌سازی بیش‌ترین بازه‌ی زمانی که می‌توان در آن مدل‌سازی را انجام داد، انتخاب شده است که این بازه از دسامبر ۲۰۱۸ تا ۳۱ مارس ۲۰۲۵ می‌باشد. همچنین داده‌های موجود توسط ابزار Google Earth Engine استخراج شده است (Al Saim and Aly 2024) و برای شهر تهران نمونه‌برداری از نقطه مرکزی شهر انجام گرفته است.

۲-۳- تعیین داده‌های پرت

شناسایی داده‌های پرت یکی از مراحل حیاتی در پیش‌پردازش داده‌ها است، به‌ویژه در تحلیل داده‌های محیطی مانند داده‌های آلودگی هوا که به دلیل تنوع منابع، شرایط جوی، و خطاهای ابزار اندازه‌گیری، ممکن است شامل مقادیر غیرعادی باشند. در این پژوهش، از یک رویکرد ترکیبی دو مرحله‌ای برای شناسایی و حذف داده‌های پرت استفاده شده است که مبتنی بر روش Robust Z-score برای تحلیل تک‌بعدی و الگوریتم IF برای تحلیل چندبعدی است. در ادامه، جزئیات هر مرحله به‌صورت جامع‌تری شرح داده می‌شود.

گام اول: روش Robust Z-score

یک ابزار آماری ساده اما مؤثر برای شناسایی داده‌های پرت در یک متغیر است. این روش به ویژه در مواردی که داده‌ها ممکن است دارای توزیع غیرنرمال یا شامل مقادیر غیرعادی باشند، عملکرد بهتری نسبت به روش استاندارد Z-score دارد (Junninen et al. 2004). فرمول محاسبه امتیاز یک مشاهده در ادامه آورده شده است (Dongre et al. 2025).

$$Z_{robust} = \frac{x - M}{MAD} \qquad (۱)$$

در این فرمول Z_{robust} نشان‌دهنده نمره مشاهده، x مقدار واقعی مشاهده، M میانه مشاهدات و MAD نیز نشان‌دهنده‌ی مقدار میانه انحراف می‌باشد.استفاده از میانه و MAD به‌جای میانگین و انحراف معیار باعث می‌شود این روش نسبت به داده‌های غیر معمول حساسیت کمتری داشته باشد (نسبت به روش پایه Z-Score)، زیرا میانگین و انحراف معیار به شدت تحت تأثیر مقادیر غیرعادی قرار می‌گیرند و ممکن است نتایج بدست آمده برای نمونه‌ها را با خطا روبه‌رو کنند. در مسائل آلودگی هوا، داده‌ها اغلب شامل مقادیر غیرعادی ناشی از خطاهای حسگرها، شرایط جوی خاص، یا حتی رویدادهای غیرمنتظره (مانند آتش‌سوزی‌ها یا انتشارات صنعتی ناگهانی) هستند که لزوم استفاده از یک روش مقاوم برای داده‌های غیر متعارف را چند برابر می‌کند (Dongre et al. 2025).

گام دوم: استفاده از الگوریتم IF برای تحلیل چندبعدی

پس از حذف داده‌های پرت در یک رویکرد تک‌بعدی، مرحله بعدی بررسی داده‌ها در فضای چندبعدی است؛ زیرا برخی داده‌های پرت ممکن است در تحلیل تک‌بعدی شناسایی نشوند، اما در ترکیب با سایر متغیرها ناهنجاری نشان دهند. برای این منظور، از الگوریتم IF استفاده شده است که یک روش یادگیری ماشین بدون نظارت برای شناسایی ناهنجاری‌ها است(Liu et al. 2008). الگوریتم IF یک رویکرد محبوب برای تعیین داده‌های ناهنجار است که علت محبوبیت این روش در محاسبات سریع، دقت مناسب و همچنین سیستم قابل فهم آن می‌باشد. (Arcudi et al. 2024).

مبنای عملکرد الگوریتم IF بر این اصل استوار است که داده‌های پرت معمولاً در مقایسه با داده‌های عادی، در یک ساختار درختی تصادفی با تعداد کمتری تقسیم‌بندی از بقیه داده‌ها جدا می‌شوند. مراحل اجرای این الگوریتم به شرح زیر می‌باشد (Liu et al. 2008).

تقسیم‌بندی تصادفی داده‌ها: در هر مرحله، یک ویژگی به‌صورت تصادفی انتخاب شده و یک مقدار آستانه تصادفی در بازه مقادیر آن ویژگی انتخاب می‌شود. داده‌ها بر اساس این آستانه به دو زیرمجموعه تقسیم می‌شوند.

ساخت درخت‌های تصادفی: این فرآیند به‌صورت بازگشتی تکرار می‌شود تا زمانی که هر داده در یک گره مجزا قرار گیرد یا تعداد تقسیم‌بندی‌ها به حد مشخصی برسد.

محاسبه طول مسیر: داده‌های پرت معمولاً با تعداد تقسیم‌بندی‌های کمتری (طول مسیر کوتاه‌تر) از داده‌های عادی جدا می‌شوند، زیرا از نظر آماری از مرکز توزیع داده‌ها فاصله بیشتری دارند.

امتیاز ناهنجاری هر داده بر اساس میانگین طول مسیر آن در مجموعه‌ای از درخت‌های تصادفی محاسبه می‌شود. داده‌هایی که امتیاز ناهنجاری بالایی دارند (یعنی طول مسیر کوتاه‌تری دارند) به‌عنوان داده‌های پرت شناسایی می‌شوند.

تنظیم ابرپارامترها

یکی از مهم‌ترین ابرپارامترهای الگوریتم IF، پارامتر contamination است که نشان‌دهنده نسبت داده‌های پرت مورد انتظار در مجموعه داده است. در این پژوهش، با توجه به تفاوت کیفیت داده‌های منابع مختلف، مقدار این پارامتر برای داده‌های هواشناسی برابر با ۰/۰۵ و برای داده‌های آلاینده‌ها برابر با ۰/۰۳ تنظیم شده است. این مقادیر بر اساس تحلیل اولیه داده‌ها و با در نظر گرفتن کیفیت بالاتر داده‌های ERA5-Land نسبت به داده‌های Sentinel-5p انتخاب شده‌اند.

۲-۴- مدل‌سازی نهایی

برای بازسازی مقادیر گمشده از الگوریتم LightGBM بهره برده شد. این مدل از یک رویکرد درختی برای مدل‌سازی استفاده می‌کند. در این روش از مزایا مدل قدرتمند Random Forest و همچنین الگوریتم یادگیری مبتنی بر گرادیان به صورت توامان بهره برده می‌شود و به همین علت است که این روش به دلیل سرعت بالا، مصرف حافظه کم و دقت مناسب در مسائل رگرسیونی و در بحث پیش‌بینی آلاینده‌های هواشناسی مورد توجه بوده است. یکی از مهم‌ترین چالش‌ها در استفاده از ابزارهای یادگیری ماشین مانند این الگوریتم بهینه‌سازی ابر پارامترهاست (Z. Yu et al. 2023). در تحقیقات مختلف نشان داده شده است که استفاده از روش جستجوی تصادفی برای یافتن ابر پارامترها دقت مناسبی ارائه می‌کند (Liashchynskyi and Liashchynskyi 2019)، در نتیجه در این پژوهش از جستجوی در این پژوهش از جستجوی تصادفی بر روی مجموعه اعتبارسنجی و انجام ۱۰۰ شبیه‌سازی برای تعیین بهترین ترکیبات ابرپارامترهای مدل LightGBM استفاده می‌شود.

۳- نتایج و بحث

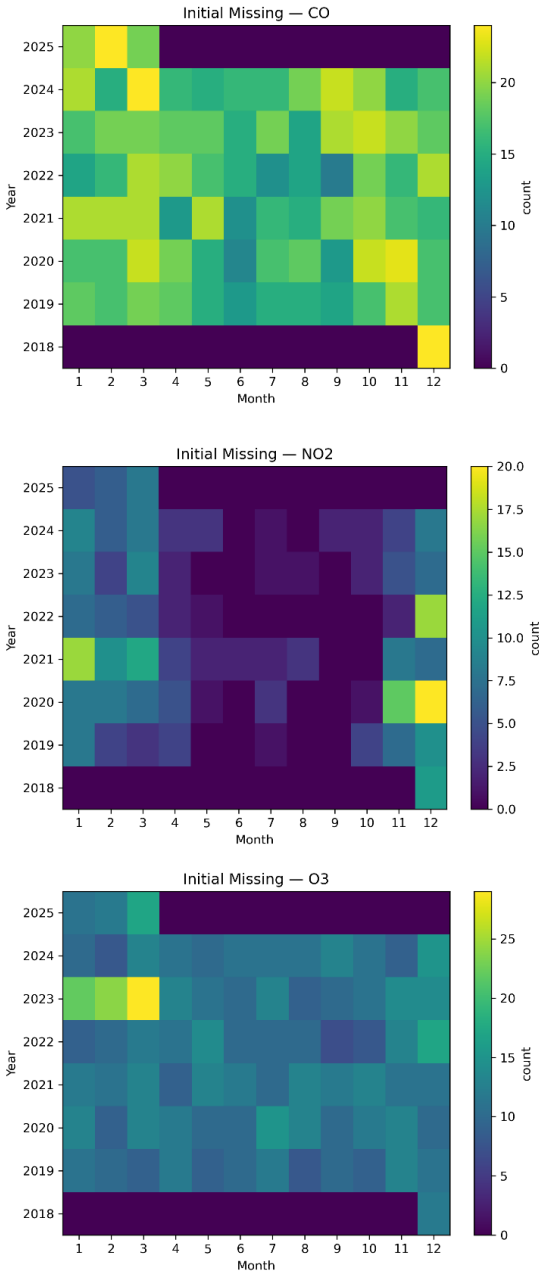
در این بخش نتایج بدست آمده آورده می‌شود و هر کدام مورد بررسی دقیق‌تر قرار می‌گیرد. در قسمت اولیه این بخش به بررسی داده‌های ناقص در مجموعه داده‌ها پرداخته می‌شود. در گام بعد، با استفاده از الگوریتم تک بعدی و سپس با استفاده از الگوریتم چند بعدی داده‌های پرت شناسایی و حذف می‌شوند. در گام بعد با توجه به داده‌های معتبر موجود، بازه‌ی زمانی که در آن کیفیت داده‌ها قابل قبول است، شناسایی می‌شود و در ادامه مدل‌سازی نهایی برای ۳ آلاینده مورد اشاره انجام می‌پذیرد.

۳-۱- توزیع داده‌های ناقص

همان‌طور که اشاره شد، داده‌های مورد مطالعه شامل دو مجموعه مجزا هستند. داده‌های هواشناسی ERA5-Land که از شبیه‌سازی‌های بازپراکنشی استخراج می‌شوند و تقریباً هیچ نمونه ناقصی ندارد. داده‌های آلایندگی Sentinel-5P که به‌صورت ستون‌های غلظت روزانه ارائه شده و به دلیل محدودیت‌های سنجش ماهواره‌ای و موارد مورد اشاره دارای مقادیر گمشده هستند.

در گام نخست، توزیع مکانی و زمانی نمونه‌های ناقص برای سه آلاینده NO_2، CO و O_3 ارزیابی شده‌است و در شکل ۲ ارائه می‌شود. نتایج نشان می‌دهد که:

- CO بیشترین نسبت داده‌های ناقص را دارا است.
- NO_2 کم‌ترین نسبت داده‌های ناقص را دارد.
- از نظر فصلی، ماه دسامبر بیشترین فراوانی نمونه‌های ناقص را در هر سه آلاینده تجربه می‌کند که ناشی از شرایط جوی خاص این ماه است.



شکل ۲: توزیع تعداد داده‌های ناقص ماهانه برای آلاینده‌های NO_2، CO و O_3 استخراج شده از Sentinel-5p در بازه‌ی زمانی دسامبر ۲۰۱۸ تا مارس۲۰۲۵

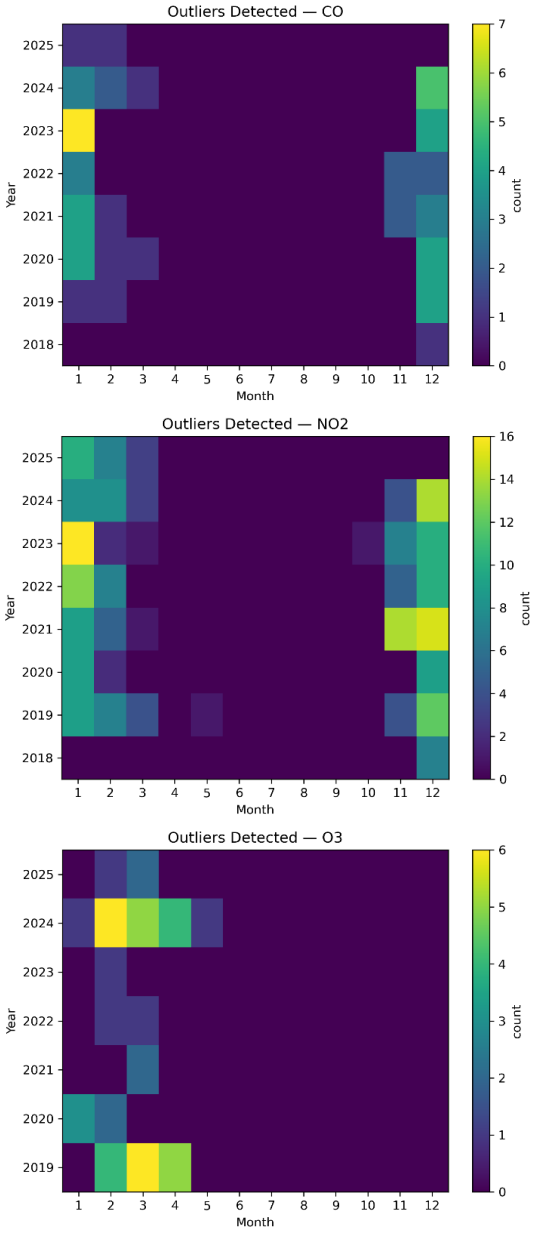
۳-۲- شناسایی داده‌های پرت – روش Robust Z-score

همان‌طور که اشاره شد در اولین مرحله از شناسایی داده‌های پرت از الگوریتم تک بعدی و مقاوم در برابر نویز Robust Z-score استفاده می‌شود. نتایج بدست آمده برای سری زمانی آلاینده‌ها و داده‌های هواشناسی به شرح زیر است (شکل ۳).

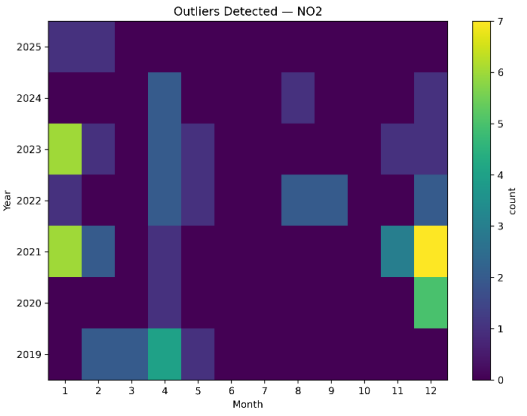
۳-۲-۱- آلاینده‌ها

- برای NO_2 تعداد ۲۳۷ نمونه پرت شناسایی شد که بیشترین تراکم آن‌ها در ماه‌های ژانویه و دسامبر قرار دارد و دلالت بر کیفیت پایین‌تر سنجش در این ماه‌هاست.
- دو آلاینده دیگر به‌طور میانگین حدود ۵۰ نمونه پرت داشتند که عدد معقول‌تری به نسبت NO_2 محسوب می‌شوند.
- توزیع فصلی نشان می‌دهد در ابتدای و انتهای سال میلادی، احتمال مشاهده داده‌های غیرمعمول بیشتر است.

با تحلیل مجدد سری زمانی NO_2 مشاهده شد که رفتار دوگانه‌ای در این سری زمانی این آلاینده وجود دارد. در فصل زمستان (اکتبر تا مارس) به‌دلیل پایداری جوی، مقادیر آلاینده به‌طور ناگهانی افزایش یافته و مدل در مقایسه با سایر داده‌ها آن‌ها



شکل ۳: توزیع تعداد نمونه‌های پرت یا غیرمعقول شناسایی شده برای برای ۳ آلاینده مورد بررسی با استفاده از روش Robust Z-score در مقیاس ماهانه



شکل ۴: توزیع تعداد نمونه‌های پرت NO_2 شناسایی شده در مقیاس ماهانه به وسیله Robust Z-score پس از تفکیک فصلی

را به عنوان مقادیر غیرمعمول شناسایی کرده است. عملا در واقعیت با دو خانواده متفاوت در یک آلاینده روبه‌رو هستیم که امکان بررسی همزمان آن‌ها وجود ندارد. بنابراین، شناسایی مجدد داده‌های غیرمعمول برای این آلاینده در دو بازه «زمستان» و «غیر زمستان» به‌صورت مجزا انجام شد تا تمرکز فصلی برطرف گردد و در نهایت نتایج بدست آمده تجمیع می‌شوند. پس از اجرای مدل برای دو خانواده به صورت جداگانه، نتایج تجمیعی نشان از کاهش چشمگیر تمرکز داده‌های پرت‌ها در یک ماه خاص داشت که نتایج آن در شکل ۴ ارائه شده است.

۳-۲-۲- داده‌های ERA5-land

مسیر قبلی برای آلاینده‌ها این بار برای داده‌های هواشناسی دنبال می‌شود. در این حالت مشخص شد که فقط مولفه شرقی باد (U) دارای ۴ عدد داده‌ی پرت می‌باشد و سایر داده‌ها بدون مشکل هستند. این نتیجه از قبل نیز قابل انتظار بود، زیرا داده‌های ERA5-land نتیجه مدل‌سازی می‌باشند و بعد از چند مرحله آنالیز کیفیت تولید شده‌اند و این درحالیست که آلاینده‌ها مستقیما از سنجش ماهواره‌ای بدست آمده بودند. با توجه به تنها ۴ عدد داده پرت یافت شده بنابراین در این حالت تصویر مربوط به توزیع آن‌ها ارایه نمی‌شود.

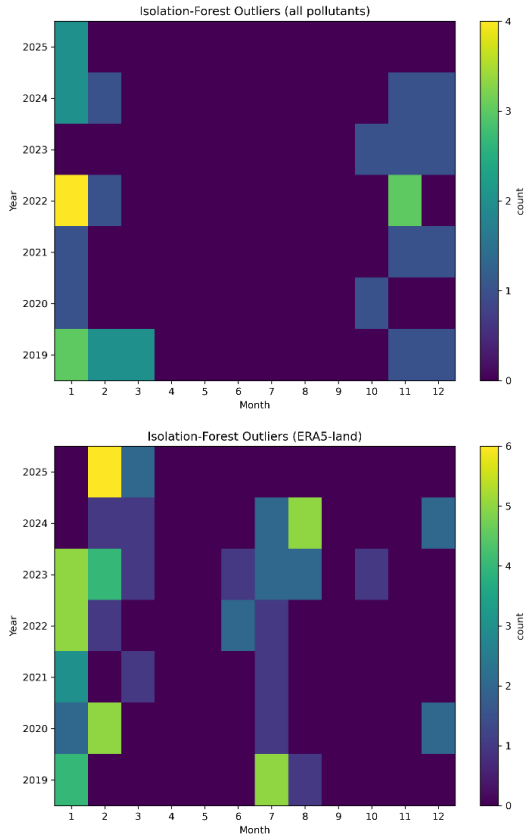
۳-۲-۳- شناسایی داده‌های پرت – الگوریتم IF

برای حذف مشاهدات نامعقول چندبعدی، ابتدا سطرهای حاوی داده‌های ناقص حذف و سپس تمام متغیرها نرمال‌سازی شدند تا مقیاس‌بندی اثری بر نتیجه نداشته باشد. الگوریتم IF برای هر دسته از داده‌ها اجرا شد تا مواردی که در بررسی تک‌بعدی شناسایی نشدند اما در الگوی گروهی نامعقول هستند، حذف شوند. نتایج این بخش به شرح زیر است (شکل ۵).

- آلاینده‌ها: از میان مشاهدات کامل، مجموعاً ۳۲ نمونه پرت در ماه‌های ابتدایی و انتهایی سال شناسایی و حذف شدند.
- داده‌های هواشناسی: تعداد ۶۹ نمونه ناهنجار تشخیص داده شد که علاوه بر فصول شروع و پایان سال، در برخی ماه‌های میانی نیز پراکندگی داشتند.

۳-۳- انتخاب بازه‌ی مناسب برای مدل‌سازی

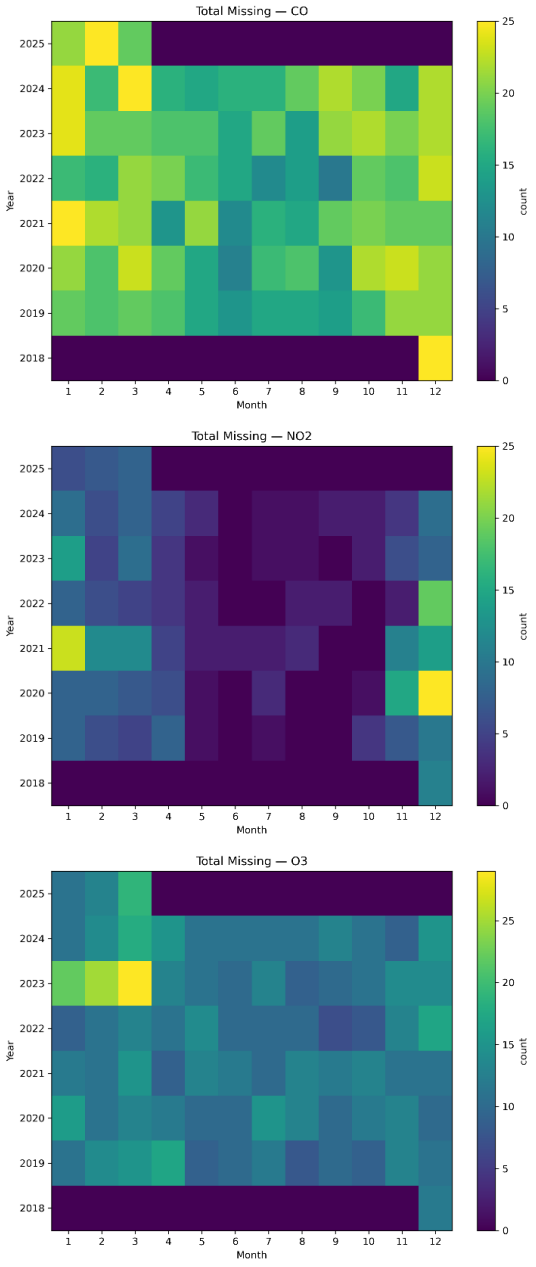
از آنجایی که هدف اصلی این مقاله یافتن مدلی برای جایگزینی داده‌های ناقص ماهواره‌ای با قطعیت بالاست، باید ماه‌ها یا سال‌هایی که تعداد داده‌های کافی برای شبیه سازی در آن وجود



شکل ۵: توزیع ماهانه تعداد نمونه‌های پرت حذف‌شده توسط IF برای مجموعه داده‌های آلاینده‌ها و متغیرهای هواشناسی

ندارد از محاسبات کنار گذاشته شود. داده‌های ERA5-land تقریبا کامل هستند و بنابراین باید بر مبنای داده‌های آلاینده‌ها تصمیم گیری انجام گیرد. در ابتدا باید نمودار مجموع داده‌های ناقص (داده‌های ناقص بعلاوه داده‌های پرت شناسایی شده) برای ۳ آلاینده ارائه شود (شکل ۶). با توجه به این تصاویر بازه‌ی نهایی بازسازی مدل برای آلاینده‌های مختلف به شرح زیر می‌باشد.

- CO: برای این آلاینده همانطور که مشخص است بیشترین تعداد داده‌ی ناقص نسبت به سایر آلاینده‌ها مشاهده می‌شود. از طرفی با توجه به داده‌های موجود مشخص است که برای ماه‌های اول تا سوم و همچنین ماه ۱۲ تعداد داده‌های موجود در اکثر سال‌ها کم تر از سایر ماه‌هاست که به طور منطقی می‌توان از مدل‌سازی در این ماه‌ها خودداری کرد. اما در صورت حذف این ماه‌ها، یک مشکل اساسی دیگر پیش می‌آید. مقدار داده‌های CO در این ماه‌ها بیشتر از سایر زمان‌هاست و اگر این داده‌ها حذف شود، مدل نمیتوان الگو مقادیر حدی را آموزش ببیند. بنابراین اگرچه می دانیم که دقت مدل برای این ماه‌ها کم است، اما ناچارا داده‌های این ماه را در فرآیند مدل‌سازی حفظ می‌کنیم.



شکل ۶: توزیع مجموع تعداد نمونه‌های ناقص و پرت شناسایی‌شده ماهانه برای آلاینده‌های Sentinel-5p

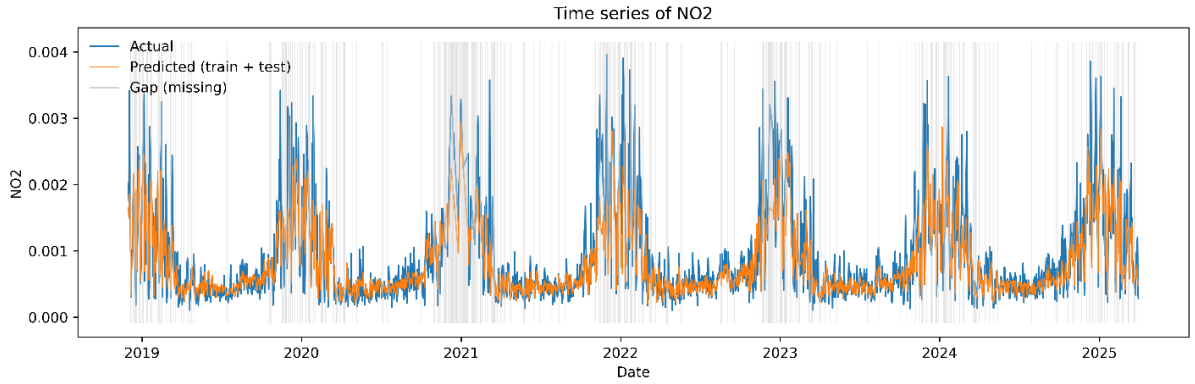
- NO_2: اگر چه در ماه‌های اول و دوم سال تعداد داده‌های ناقص زیاد می‌باشد اما از انجاییکه این موضوع برای تمام سال‌ها اتفاق نیافتاده است در نتیجه می‌توان فرآیند بازسازی داده‌ها را برای تمامی ماه‌ها انجام داد.
- O_3: بیشترین تمرکز داده‌های ناقص در اوایل سال ۲۰۲۳ می‌باشد ولی از آنجایی که این تمرکز فقط در این سال اتفاق افتاده است بنابراین فرآیند تکمیل داده‌های ناقص آن می‌تواند انجام گیرد و مدل توانایی یادگیری رفتار این فصل را از فصل‌های مشابه در سال‌های قبل نیز دارد. از اینرو برای O_3 نیز

در تمامی ماه‌های سال فرآیند بازسازی صورت می‌گیرد.

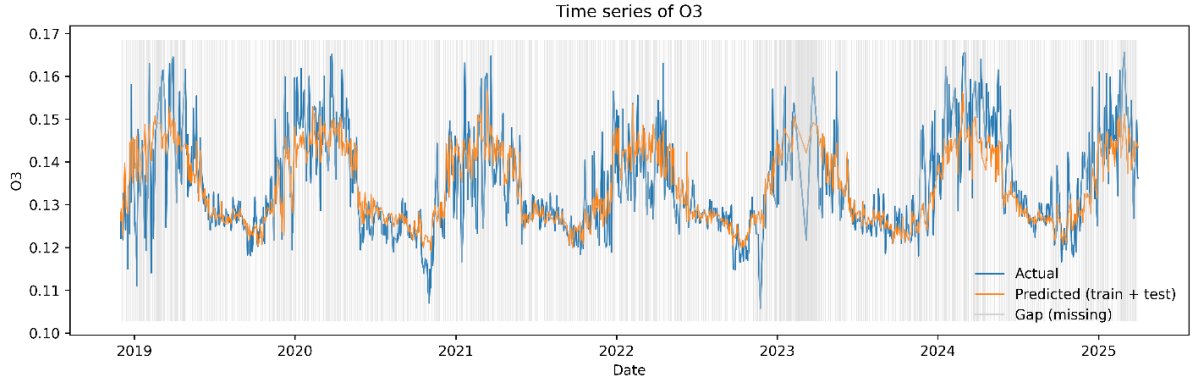
۳-۴- نتایج مدل‌سازی

برای تعیین مقادیر گمشده هر کدام از آلاینده‌ها، از متغیرهای ERA5-land به عنوان متغیر اصلی استفاده می‌شود. دو متغیر کمکی نیز به این داده‌ها اضافه می‌شوند تا اثر زمان را در فرآیند مدل‌سازی نهایی در برگیرند. متغیر اول شماره روز در هفته است که به اثر ترافیک در مدل سازی اشاره میکند. متغیر دوم نیز شماره روز در سال است که اثر تغییرات فصلی را لحاظ می‌کند. با استفاده از این داده‌ها فرآیند مدل‌سازی برای ۳ آلاینده مختلف انجام می‌شود که نتیجه آن‌ها در ادامه ارائه شده است.

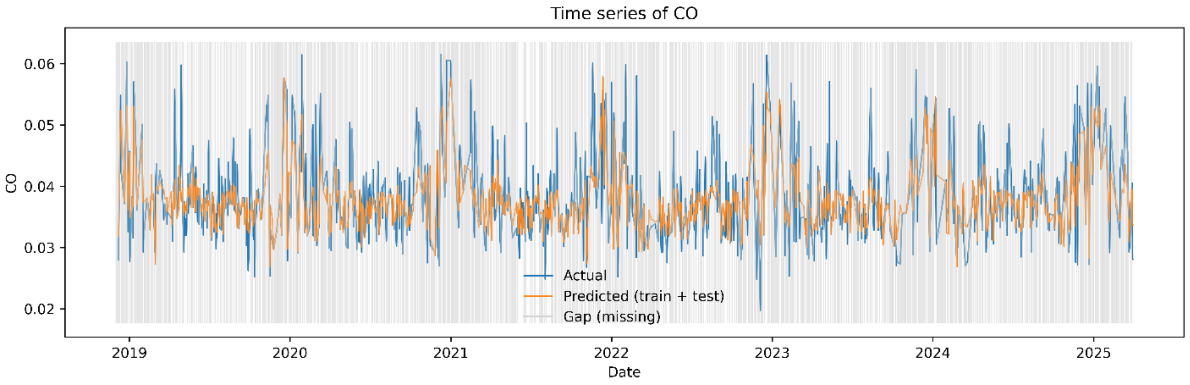
- NO_2: نتایج بدست آمده برای این آلاینده در شکل ۷ نمایش داده شده است. عملکرد نهایی مدل بر روی داده های تست مقدار r^2 را برابر با ۰/۶۱ و خطای MAE را برابر با ۰/۰۰۰۲۶۹ نشان می‌دهد. همانطور که در تصویر مشخص است و قبلا مورد اشاره قرار گرفت، مدل عملا با دو رفتار متفاوت در داده‌ها روبه‌رو بوده است؛ مقادیر زمستانه‌ای که اعداد بالایی دارند و سایر فصول که اعداد پایین‌تر دارند. نتایج نشان می‌دهد که مدل توانسته است بین این دو رفتار مختلف یک بالانس مناسبی را ایجاد کند و روندهای اصلی را تقریبا دنبال کند.
- O_3: نمودار پیش‌بینی مقادیر این آلاینده در شکل ۸ آورده شده است. اولین تفاوت ظاهری که با آلاینده NO_2 وجود دارد، تفاوت در گپ داده‌ها می‌باشد. در حالتیکه برای NO_2 تمرکز این داده‌ها بر روی ماه‌های سرد سال بود، در این حالت داده‌های ناقص در طول کل سال پراکنده شده‌اند. دقت مدل نهایی استفاده شده به نسبت NO_2 کمتر است؛ در این حالت برای داده‌های تست r^2 برابر با ۰/۵ و مقدار MAE برابر با ۰/۰۰۵۶ بدست آمده‌اند. مدل در این حالت نیز توانسته است رفتار داده‌های اصلی را به خوبی دنبال کند.
- CO: به لحاظ پراکندگی داده‌های مفقود این آلاینده مشابه O_3 می‌باشد و در تمام طول دوره دارای داده ناقص می‌باشد. دقت



شکل ۷: نمودار سری زمانی آلاینده NO_2 در شهر تهران (مقادیر مشاهده شده با رنگ آبی و مقادیر مدل‌سازی شده به وسیله مدل LightGBM با رنگ نارنجی- بازه‌های خاکستری نشان دهنده داده‌های مفقود می‌باشد)



شکل ۸: نمودار سری زمانی آلاینده O_3 در شهر تهران (مقادیر مشاهده شده با رنگ آبی و مقادیر مدل‌سازی شده به وسیله مدل LightGBM با رنگ نارنجی- بازه‌های خاکستری نشان دهنده داده‌های مفقود می‌باشد)



شکل ۹: نمودار سری زمانی آلاینده CO در شهر تهران (مقادیر مشاهده شده با رنگ آبی و مقادیر مدل‌سازی شده به وسیله مدل LightGBM با رنگ نارنجی- بازه‌های خاکستری نشان دهنده داده‌های مفقود می‌باشد)

مدل با r^2 برابر با ۰/۳۸ و MAE برابر با ۰/۰۰۴ بر روی داده‌های تست از دو آلاینده دیگر ضعیفتر می‌باشد. یکی از مهم‌ترین دلایل در تعداد زیاد داده‌های ناقص و کمبود داده برای آموزش مدل نهفته است. اما اگرچه دقت نهایی مناسب بدست نیامده است اما کماکان مدل توانسته است روند اصلی آلاینده را دنبال کند (شکل ۹).

۴- نتیجه‌گیری

آلودگی هوا طی سال‌های اخیر به مهم‌ترین چالش زیست‌محیطی و بهداشتی کلان‌شهر تهران بدل شده است؛ چالشی که مدیریت کارآمد آن مستلزم دسترسی به سری‌های زمانی کامل و قابل‌اعتماد از غلظت آلاینده‌هاست. مطالعه حاضر با اتکا به داده‌های ماهواره‌ای Sentinel-5P و متغیرهای هواشناسی ERA5-Land، مجموعه‌ای پیوسته برای سه آلاینده کلیدی NO_2، O_3 و CO را در بازه دسامبر ۲۰۱۸ تا مارس ۲۰۲۵ تولید کرد.

بررسی کیفیت اولیه داده‌ها حاکی از آن بود که CO بیشترین و NO_2 کمترین سهم داده ناقص را دارا هستند. از نظر زمانی، فصول سرد (دی و بهمن) بیشترین خلأ را نشان دادند؛ پایداری جوی، وارونگی حرارتی و پوشش ابری زمستان دلیل اصلی این الگوست. این یافته همسو با گزارش‌های جهانی مشابه است که الگوی «نقص پراکنده + حفره‌های متوالی طولانی» را برای داده‌های ماهواره‌ای مشاهده کرده‌اند (Čampulová et al. 2019).

به‌منظور بهبود صحت سری زمانی، یک رویکرد دو مرحله‌ای برای تشخیص داده‌های پرت به کار گرفته شد: Robust Z-score تک‌متغیره برای حذف داده‌های غیرمعمول آشکار و IF چندبعدی برای شناسایی داده‌های پیچیده با وابستگی متقابل متغیرها. مقایسه نشان داد که هر کدام از این روش‌ها به‌تنهایی بخشی از داده‌ها را شناسایی می‌کنند. نتایج همچنین نشان دادند که پس از پالایش تک‌بعدی، داده‌های پرت باقی‌مانده عمدتاً در متغیرهای هواشناسی نهفته‌اند؛ موضوعی که لزوم استفاده از رویکرد چندبعدی را تقویت می‌کند که این نتیجه مطابق با تحقیقات پیشین می‌باشد که استفاد از رویکردهای ترکیبی مختلف را پیشنهاد کرده بودند (Čampulová et al. 2019).

برای برآورد مقادیر مفقود، الگوریتم LightGBM با جست‌وجوی تصادفی ابرپارامترها آموزش داده شد. شاخص دقت r^2 مدل به ترتیب ۰/۶۱ برای NO_2، ۰/۵۰ برای O_3 و ۰/۳۸ برای CO به دست آمد. کارایی پایین‌تر CO به دو عامل نسبت داده می‌شود. اول از همه نرخ بالای داده‌های ناقص که باعث می‌شود آموزش مدل با محدودیت مواجه شود؛ و ثانیا رفتار غیرخطی شدید CO در پیک‌های زمستانی که به دلیل شرایط وارونگی حرارتی رخ می‌دهد. با این حال، با توجه به ساده بودن مدل و محدودیت داده، عملکرد برای NO_2 و حتی O_3 تقریبا در دامنه مقالات اخیر (۰/۶ تا ۰/۸) قرار می‌گیرد (Shao et al. 2023; de la Cruz Libardi et al. 2024; Zhang and Zhou 2024). این نتیجه نشان می‌دهد که حتی الگوریتم‌های نسبتاً سبک نیز، اگر داده‌های ورودی به درستی پالایش شوند، می‌توانند منجر به نتایج معقولی شوند.

پیشنهاد می‌شود در مطالعات آتی از آنجاییکه تقریبا دو رفتار متفاوت در سری زمانی آلاینده‌ها مشاهده می‌شود، امکان استفاده از چند مدل مختلف به ازای رفتارهای مختلف مورد بررسی قرار گیرد. علاوه بر این برای بهبود فرآیند مدل‌سازی می‌توان اثر متغیرهای مختلف در پیش‌بینی را هم بررسی کرد.

References

Aithal, Sathya Swarup, Ishaan Sachdeva, and Om P Kurmi. 2023. "Air Quality and Respiratory Health in Children." *Breathe* 19 (2).

Appel, Marius. 2024. "Efficient Data-Driven Gap Filling of Satellite Image Time Series Using Deep Neural Networks with Partial Convolutions." *Artificial Intelligence for the Earth Systems* 3 (2): 220055.

Arcudi, Alessio, Davide Frizzo, Chiara Masiero, and Gian Antonio Susto. 2024. "Enhancing Interpretability and Generalizability in Extended Isolation Forests." *Engineering Applications of Artificial Intelligence* 138: 109409.

Bayat, Reza, Khosro Ashrafi, Majid Shafiepour Motlagh, Mohammad Sadegh Hassanvand, Rajabali Daroudi, Günther Fink, and Nino Künzli. 2019. "Health Impact and Related Cost of Ambient Air Pollution in Tehran." *Environmental Research* 176: 108547.

Borhani, Faezeh, Majid Shafiepour Motlagh, Amir Houshang Ehsani, Yousef Rashidi, Masoud Ghahremanloo, Meisam Amani, and Armin Moghimi. 2023. "Current Status and Future Forecast of Short-Lived Climate-Forced Ozone in Tehran, Iran, Derived from Ground-Based and Satellite Observations." *Water, Air, & Soil Pollution* 234 (2): 134.

Borhani, Faezeh, Majid Shafiepour Motlagh, Andreas Stohl, Yousef Rashidi, and Amir Houshang Ehsani. 2022. "Tropospheric Ozone in Tehran, Iran, during the Last 20 Years." *Environmental Geochemistry and Health*, 1–23.

Čampulová, Martina, Jaroslav Michalek, and Jiří Moučka. 2019. "Generalised Linear Model-Based Algorithm for Detection of Outliers in Environmental Data and Comparison with Semi-Parametric Outlier Detection Methods." *Atmospheric Pollution Research* 10 (4): 1015–23.

Dongre, Pradeep Kumar, Viral Patel, Upendra Bhoi, and Nilesh N Maltare. 2025. "An Outlier Detection Framework for Air Quality Index Prediction Using Linear and Ensemble Models." *Decision Analytics Journal* 14: 100546.

Holloway, Tracey, Daegan Miller, Susan Anenberg, Minghui Diao, Bryan Duncan, Arlene M Fiore, Daven K Henze, Jeremy Hess, Patrick L Kinney, and Yang Liu. 2021. "Satellite Monitoring for Air Quality and Health." *Annual Review of Biomedical Data Science* 4 (1): 417–47.

Hua, Van, Thu Nguyen, Minh-Son Dao, Hien D Nguyen, and Binh T Nguyen. 2024. "The Impact of Data Imputation on Air Quality Prediction Problem." *Plos One* 19 (9): e0306303.

Jiménez-Navarro, Manuel J, Mario Lovrić, Simonas Kecorius, Emmanuel Karlo Nyarko, and María Martínez-Ballesteros. 2024. "Explainable Deep Learning on Multi-Target Time Series Forecasting: An Air Pollution Use Case." *Results in Engineering* 24: 103290.

Junninen, Heikki, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen. 2004. "Methods for Imputation of Missing Values in Air Quality Data Sets." *Atmospheric Environment* 38 (18): 2895–2907.

la Cruz Libardi, Arturo de, Pierre Masselot, Rochelle Schneider, Emily Nightingale, Ai Milojevic, Jacopo Vanoli, Malcolm N Mistry, and Antonio Gasparrini. 2024. "High Resolution Mapping of Nitrogen Dioxide and Particulate Matter in Great Britain (2003–2021) with Multi-Stage Data Reconstruction and Ensemble Machine Learning Methods." *Atmospheric Pollution Research* 15 (11): 102284.

Li, Meixin, Ying Wu, Yansong Bao, Bofan Liu, and George P Petropoulos. 2022. "Near-Surface NO2 Concentration Estimation by Random Forest Modeling and Sentinel-5P and Ancillary Data." *Remote Sensing* 14 (15): 3612.

Liashchynskyi, Petro, and Pavlo Liashchynskyi. 2019. "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS." *ArXiv Preprint ArXiv:1912.06059*.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. 2008. "Isolation Forest." In *2008 Eighth Ieee International Conference on Data Mining*, 413–22. IEEE.

Ramírez, A Susana, Steven Ramondt, Karina Van Bogart, and Raquel Perez-Zuniga. 2019. "Public Awareness of Air Pollution and Health Threats: Challenges and Opportunities for Communication Strategies to Improve Environmental Health Literacy." *Journal of Health Communication* 24 (1): 75–83.

Rollo, Federica, Chiara Bachechi, and Laura Po. 2023. "Anomaly Detection and Repairing for Improving Air Quality Monitoring." *Sensors* 23 (2): 640.

Saim, Abdullah Al, and Mohamed H Aly. 2024. "Big Data Analyses for Determining the Spatio-Temporal Trends of Air Pollution Due to Wildfires in California Using Google Earth Engine." *Atmospheric Pollution Research* 15 (9): 102226.

Schneider, Philipp, Paul D Hamer, Arve Kylling, Shobitha Shetty, and Kerstin Stebel. 2021. "Spatiotemporal Patterns in Data Availability of the Sentinel-5p No2 Product over Urban Areas in Norway." *Remote Sensing* 13 (11): 2095.

Schneising, Oliver, Michael Buchwitz, Jonas Hachmeister, Steffen Vanselow, Maximilian Reuter, Matthias Buschmann, Heinrich Bovensmann, and John P Burrows. 2023. "Advances in Retrieving XCH 4 and XCO from Sentinel-5 Precursor: Improvements in the Scientific TROPOMI/WFMD Algorithm." *Atmospheric Measurement Techniques* 16 (3): 669–94.

Shao, Yanchuan, Wei Zhao, Riyang Liu, Jianxun Yang, Miaomiao Liu, Wen Fang, Litiao Hu, Matthew

Adams, Jun Bi, and Zongwei Ma. 2023. "Estimation of Daily NO2 with Explainable Machine Learning Model in China, 2007–2020." *Atmospheric Environment* 314: 120111.

Sokhi, Ranjeet S, Nicolas Moussiopoulos, Alexander Baklanov, John Bartzis, Isabelle Coll, Sandro Finardi, Rainer Friedrich, Camilla Geels, Tiia Grönholm, and Tomas Halenka. 2022. "Advances in Air Quality Research–Current and Emerging Challenges." *Atmospheric Chemistry and Physics* 22 (7): 4615–4703.

Tabunshchik, Vladimir, Aleksandra Nikiforova, Nastasia Lineva, Polina Drygval, Roman Gorbunov, Tatiana Gorbunova, Ibragim Kerimov, Cam Nhung Pham, Nikolai Bratanov, and Mariia Kiseleva. 2024. "The Dynamics of Air Pollution in the Southwestern Part of the Caspian Sea Basin (Based on the Analysis of Sentinel-5 Satellite Data Utilizing the Google Earth Engine Cloud-Computing Platform)." *Atmosphere* 15 (11): 1371.

Wang, Guojie, Damien Garcia, Yi Liu, Richard De Jeu, and A Johannes Dolman. 2012. "A Three-Dimensional Gap Filling Method for Large Geophysical Datasets: Application to Global Satellite Soil Moisture Observations." *Environmental Modelling & Software* 30: 139–42.

Yu, Xinyu, Man Sing Wong, Majid Nazeer, Zhengqiang Li, and Coco Yin Tung Kwok. 2024. "A Novel Algorithm for Full-Coverage Daily Aerosol Optical Depth Retrievals Using Machine Learning-Based Reconstruction Technique." *Atmospheric Environment* 318: 120216.

Yu, Zhongqi, Jinghui Ma, Yuanhao Qu, Liang Pan, and Shiquan Wan. 2023. "PM2. 5 Extended-Range Forecast Based on MJO and S2S Using LightGBM." *Science of The Total Environment* 880: 163358.

Zali, Nader, Masoud Zamanipoor, Hassan Ahmadi, and Mehrdad Karami. 2018. "Analysis of Key Factors Influencing Air Pollution of Metropolises in Developing Countries by Year 2025 (Case Study: Tehran Metropolis, Iran)." *Anu. Do Inst. De Geocienc* 41: 548–59.

Zhang, Xiaoxia, and Pengcheng Zhou. 2024. "A Transferred Spatio-Temporal Deep Model Based on Multi-LSTM Auto-Encoder for Air Pollution Time Series Missing Value Imputation." *Future Generation Computer Systems* 156: 325–38.

Zheng, Zihao, Zhiwei Yang, Zhifeng Wu, and Francesco Marinello. 2019. "Spatial Variation of NO2 and Its Impact Factors in China: An Application of Sentinel-5P Products." *Remote Sensing* 11 (16): 1939.

دلاور، محمودرضا، غلامی، امین، شیران، غلامرضا، رشیدی، یوسف، نخعی زاده، غلام رضا، فدرا، کرت، و هاتفی افشار، اسماعیل. ۱۳۹۹. "بهبود برآورد میزان آلودگی هوای شهر تهران." *مجله علمی رایانش نرم و فناوری اطلاعات* ۹ (۲): ۸۷–۹۹.