مدیریت مخاطرات محیطی (دانش مخاطرات سابق)/ دورهٔ ۷، شمارهٔ ۳، پاییز ۱۳۹۹/ ص ۳۱۲-۲۹۹
DOI: 10.22059/jhsci.2020.309653.595

مقابله با مخاطرات ناشی از غلظت آلایندهٔ $PM_{2.5}$ با به‌کارگیری روش‌های رگرسیونی و شباهت مکانی- زمانی و تخمین مقادیر گم‌شده در سری زمانی آنها (مطالعهٔ موردی: شهر تهران)

مرجان فرجی
دانشجوی دکتری مهندسی نقشه‌برداری گرایش سنجش از دور، دانشکدهٔ مهندسی عمران و حمل‌ونقل، دانشگاه اصفهان
سعید نادی*
استادیار گروه مهندسی نقشه‌برداری، دانشکدهٔ مهندسی عمران و حمل‌ونقل، دانشگاه اصفهان

(تاریخ دریافت ۱۳۹۹/۶/۱۸ – تاریخ پذیرش ۱۳۹۹/۹/۹)

چکیده

با توجه به تأثیر نامطلوب آلاینده‌ها بر محیط زیست و سلامت انسان، تجزیه‌وتحلیل داده‌های کیفیت هوا اهمیت زیادی در حفاظت از محیط زیست و رویارویی با مشکلات آلودگی هوا دارد. داده‌های گم‌شده در سری‌های زمانی به‌خصوص داده‌های مربوط به آلودگی هوا موجب بروز چالشی ویژه در برابر آنالیز این داده‌ها می‌شود که ضرورت استفاده از روش‌هایی با عنوان جانهی را برای مقابله با این پدیده نمایان می‌کند. مقادیر گم‌شده، موجب کاهش حجم داده و تغییر الگوهای زمانی موجود در داده‌ها و نتیجه‌گیری اشتباه در تجزیه‌وتحلیل داده‌ها می‌شود. در این پژوهش به‌منظور جانهی مقادیر ازدست‌رفته در داده‌های سری زمانی غلظت آلایندهٔ $PM_{2.5}$ از ۱۲ ایستگاه سنجش آلودگی شهر تهران، روشی ترکیبی برمبنای رگرسیون جانهی با در نظر گرفتن وابستگی و شباهت‌های مکانی و زمانی بین ایستگاه‌ها توسط الگوریتم پیچش زمانی پویا معرفی شده است. داده‌هایی با مقادیر گم‌شده با الگویی مشابه با داده‌های اصلی در دامنهٔ ۱۰، ۱۵ و ۲۰ درصد گم‌شدگی در داده‌ها با هدف ارزیابی عملکرد مدل‌های جانهی شبیه‌سازی شدند. سپس روش پیشنهادی در ترکیب با روش‌های مختلف جانهی چندگانه همانند روش طبقه‌بندی و رگرسیون درختی، نمونهٔ تصادفی و میانگین تطابق پیش‌بینی کننده، اجرا و نتایج با روش‌های جانهی منفرد مقایسه شد. نتایج بیانگر برتری روش معرفی‌شده در ترکیب با رگرسیون درختی در مقایسه با دیگر روش‌های جانهی چندگانه و منفرد است.

واژه‌های کلیدی: آلایندهٔ $PM_{2.5}$، جانهی منفرد و چندگانه، داده‌های گم‌شده، مخاطرات، معیار شباهت DTW.

* نویسندهٔ مسئول، تلفن: ۰۹۱۳۳۱۷۱۷۸۱
Email: Snadi@eng.ui.ac.ir

مقدمه

با افزایش صنعتی شدن شهرها، آلودگی هوا به یکی از مخاطرات جدی زیست‌محیطی کلان‌شهرهای جهان از جمله تهران مبدل شده است که گذشته از ضررهای جبران‌ناپذیر برای سلامت شهروندان، تأثیرات اجتماعی و اقتصادی فراوانی بر شهر تحمیل می‌کند [۱،۴]. با توجه به آثار نامطلوب آلاینده‌ها بر محیط زیست و سلامت انسان، تجزیه‌وتحلیل داده‌های کیفیت هوا اهمیت زیادی در حفاظت از محیط زیست و مخاطرات آن و رویارویی با مشکلات آلودگی هوا دارد [۱۲]. در طول دههٔ اخیر شمار زیادی از داده‌های کنترل کیفیت هوا که دربردارندهٔ غلظت آلاینده‌های موجود در جو هستند، توسط ایستگاه‌های سنجش آلودگی در شهرهای مختلف کشور جمع‌آوری‌شده‌اند که به دلایل مختلف مانند تعمیرات و نگهداری سنجنده‌ها، خطاهای دستگاهی و خطاهای پردازشی دارای گم‌شدگی با فواصل مختلف‌اند [۲۱]. این مقادیر گم‌شده، سبب اشکالاتی در تجزیه‌وتحلیل داده‌ها می‌شود و تصمیم‌گیری براساس این داده‌ها را با چالش روبه‌رو می‌کند. برای مثال بسیاری از الگوریتم‌های یادگیری ماشین همانند شبکه‌های عصبی و بردار پشتیبان با استفاده از داده‌های کامل به‌کار گرفته می‌شوند [۷].

داده‌های گم‌شده، پدیدهٔ شایعی در مسائل سری زمانی است و معرفی مدل‌ها و روش‌های کارامد برای مدیریت این خلأ در داده‌ها، گام مؤثری برای کاهش بایاس و افزایش قدرت مدل است. روش‌های رایج که قادر به جانهی مقادیر گم‌شده در سری زمانی هستند، به‌طور معمول به سه دستهٔ حذف داده‌های گم‌شده، جانهی منفرد و جانهی چندگانه تقسیم می‌شوند [۲]. حذف داده‌های گم‌شده از سری زمانی با عنوان حذف کامل[1] یکی از روش‌های قدیمی و متداول برای مقابله با این مشکل است [۲۲]. این روش سبب از بین رفتن اطلاعات باارزش و کاهش حجم داده می‌شود. تحلیل نتایج در این روش نیز به‌دلیل تغییر در توزیع داده‌ها و برآورد اریب پارامترها، با خطا مواجه می‌شود. روش جانهی منفرد یکی از رویکردهای رایج برای پر کردن داده‌های جاافتاده است. ساده‌ترین شیوه جانهی منفرد، برآورد مقادیر با روش‌هایی نظیر میانگین، میانه، مد، جانهی تصادفی و میانگین شرط دار برای نقطه‌ای است که در آن مقادیر گم‌شده است [۱۵]. بدین منظور مقداری ثابت برای همهٔ مقادیر ازدست‌رفته جایگزین می‌شود و این جانهی، به الگو و ماهیت سری زمانی وابستگی ندارد. این روش‌ها سریع و آسان‌اند، اما موجب تولید پاسخ‌های نامطمئن و بایاس می‌شوند. در روش‌های جانهی چندگانه، مقادیر

1. Listwise deletion

گم‌شده با چند مقدار ممکن جایگذاری می‌شوند [۱۶]. از روش‌های به‌نسبت سادهٔ جانهی چندگانه می‌توان به روش‌های رگرسیون خطی و غیرخطی [۱۴]، جانهی چندمتغیرهٔ معادلات زنجیره‌ای[1] شامل روش‌های تطابق میانگین پیش‌بینی‌کننده[2] [۲۰]، رگرسیون درختی[3] [۵] و رگرسیون لجستیک [۱۹]، همچنین روش‌های پیچیده و زمان‌بر مانند بیشینه‌سازی انتظار[4] [۲۰] و بیزین [۸] اشاره کرد. افزون‌بر روش‌های جانهی چندگانه، روش‌های نوین همانند روش‌های شبکهٔ عصبی یادگیری عمیق [۹،۱۷] نیز امروزه پرکاربرد است.

با توجه به اینکه در بسیاری از موارد الگوی گم‌شدگی در سری زمانی ایستگاه‌ها مشابه هم نیست، در دوره‌های زمانی مواجهه با گم‌شدگی مقادیر در یک ایستگاه، ممکن است در ایستگاه‌های دیگر این گم‌شدگی وجود نداشته باشد. بر این اساس، نوآوری اصلی این مقاله استفاده از الگوی موجود در داده‌های سری زمانی ایستگاه‌های مشابه برای پر کردن داده‌های جاافتاده در سری زمانی یک ایستگاه است. بدین منظور در این پژوهش با استفاده از الگوریتم پیچش زمانی پویا، ایستگاه‌هایی با شباهت بیشینه شناسایی و در مدل‌های جانهی هم‌زمان پردازش شد تا از الگوی موجود در سری زمانی ایستگاه‌های مشابه برای پر کردن مقادیر جاافتاده استفاده شود. نتایج نشان داد که استفاده از روش‌های جانهی چندگانه برمبنای روش بیان‌شده در این مقاله، در بازه‌های گم‌شدگی طولانی در مقایسه با روش‌های منفرد برتری دارد.

منطقهٔ تحقیق و داده‌های استفاده‌شده

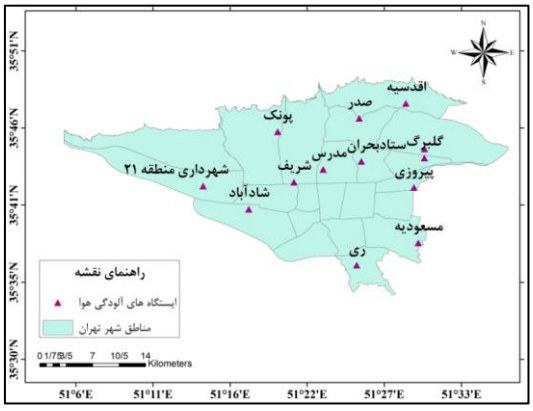
تهران با جمعیتی بالغ بر ۸/۵ میلیون نفر، از شمال به رشته‌کوه البرز و از جنوب به دشت کویر منتهی است. امروزه آلودگی هوای ناشی از ترافیک سنگین، افزایش جمعیت، صنایع و گردوغبار به یکی از مشکلات جدی تهران تبدیل شده است [۲۱]. در این تحقیق، از داده‌های غلظت آلایندهٔ $PM_{2.5}$ ثبت‌شده در ۱۲ ایستگاه نظارت بر کیفیت هوا استفاده شده است. داده‌ها به‌صورت ساعتی در دامنهٔ زمانی ۱۷ آذر ۱۳۹۵ تا ۸ اسفند ۱۳۹۷ از سایت کنترل کیفیت هوا جمع‌آوری شد. اطلاعات آماری ۱۲ ایستگاه سنجش آلودگی همانند میانگین، میانه، انحراف معیار، تعداد مقادیر گم‌شده، و بیشترین و کمترین مقدار داده‌های ساعتی در دامنهٔ زمانی مورد نظر، در جدول ۱ نشان داده شده است.

1. Multivariate Imputation by Chained Equations (MICE)
2. Predictive mean matching (PMM)
3. Classification and regression trees (Cart)
4. Maximum Exception Maximum Exception

بیشترین و کمترین مقدار گم‌شدگی در داده‌های ایستگاه‌ها متعلق به ایستگاه شادآباد و ایستگاه شهرری به‌ترتیب با مقدار عددی ۲۳۵۲ و ۵۲۰ دادهٔ گم‌شده است. ایستگاه‌های شادآباد، پیروزی، شهرداری منطقهٔ ۲۱ و مسعودیه بیشترین مقادیر غلظت آلاینده را داشتند. موقعیت مکانی این ایستگاه‌ها در شکل ۱ مشاهده می‌شود.

جدول ۱. مشخصات آماری غلظت آلایندهٔ $PM_{2.5}$ ثبت‌شده در ایستگاه‌های سنجش آلودگی

نام ایستگاه	انحراف معیار	مقادیر گم‌شده	میانگین	میانه	بیشترین	کمترین
اقدسیه (S1)	۱۳	۱۲۰۵	۲۲	۲۰	۱۸۹	۰
مسعودیه (S2)	۱۳	۱۵۸۷	۳۹	۳۶	۳۸۴	۱
پیروزی (S3)	۲۳	۲۰۲۷	۳۳	۳۸	۲۸۵	۱
ری (S4)	۲۴	۵۲۰	۳۹	۳۴	۳۱۹	۱
شهرداری منطقهٔ ۲۱ (S5)	۲۰	۱۹۰۴	۳۵	۳۰	۳۵۹	۱
شادآباد (S6)	۲۴	۲۳۵۲	۴۰	۳۵	۳۹۳	۱
شریف (S7)	۲۲	۱۸۵۶	۳۸	۳۴	۲۷۸	۱
گلبرگ (S8)	۱۵	۱۴۲۶	۲۰	۲۲	۱۱۷	۰
صدر (S9)	۲۳	۱۰۴۱	۳۸	۳۲	۳۰۸	۱
پونک (S10)	۱۶	۱۹۷۱	۲۴	۲۱	۱۹۲	۱
مدرس (S11)	۲۳	۶۴۸	۳۵	۲۹	۲۰۴	۱
ستاد بحران (S12)	۱۸	۸۲۷	۳۰	۲۷	۱۵۸	۱



شکل ۱. موقعیت مکانی ایستگاه‌های سنجش آلودگی شهر تهران

روش تحقیق

هدف این مقاله، معرفی روشی نوآورانه برمبنای شامل کردن وابستگی‌های مکانی بین سری‌های زمانی مربوط به ایستگاه‌های مشابه از دیدگاه رفتار سری زمانی در جانهی اطلاعات گم‌شده مربوط به هر ایستگاه سنجش آلودگی است. در گام اول از طریق الگوریتم پیچش زمانی پویا، شباهت مکانی- زمانی میان سری‌های زمانی غلظت آلایندهٔ $PM_{2.5}$ ایستگاه‌ها به‌صورت دوبه‌دو محاسبه شد. سپس برای جانهی در هر ایستگاه هدف، از وابستگی ایستگاه‌هایی که بیشترین شباهت را با ایستگاه مورد نظر داشتند استفاده می‌شود. در گام دوم، دادهٔ کامل اولیه با حذف مقادیر گم‌شده در هر ایستگاه تشکیل می‌شود. در گام بعدی با الگوی مشابه با داده‌های گم‌شدهٔ اصلی، داده‌های گم‌شدهٔ جدیدی با ۱۰، ۱۵ و ۲۰ درصد گم‌شدگی در داده‌ها حاصل می‌شود. گام چهارم، دربرگیرندهٔ اجرا و مقایسهٔ الگوریتم‌های مختلف جانهی چندگانه و منفرد برای پر کردن داده‌های ازدست‌رفته است. در نهایت عملکرد روش‌های مختلف جانهی از طریق شاخص‌های معرفی‌شده بررسی می‌شود.

استفاده از الگوریتم پیچش زمانی پویا (DTW)

ایدهٔ اصلی در این مقاله این است که با توجه به این موضوع که بسیاری از روش‌های جانهی چندگانه رگرسیون‌مبنا هستند با کنار هم قرار دادن ایستگاه‌های مشابه از لحاظ مکانی- زمانی برای جانهی مقادیر گم‌شده در یک ایستگاه به دقت بیشتری می‌توان دست یافت. برای این منظور در این پژوهش از الگوریتم پیچش پویای زمانی[1] برای تعیین وابستگی و شباهت مکانی- زمانی بین سری زمانی ثبت‌شده در ایستگاه‌های مختلف استفاده شد. این الگوریتم به‌منظور اندازه‌گیری شباهت بین دو دنباله از داده‌های سری زمانی استفاده می‌شود. به بیان دیگر این الگوریتم روش مناسبی برای یافتن مسیر بهینه در نقاط بهینهٔ متناظر بین دو سری زمانی است به‌نحوی که کمترین فاصلهٔ مکانی یا بیشترین تشابه بین دو سری زمانی حاصل شود [۱۸].

استفاده از روش‌های جانهی چندگانه

در این پژوهش از روش‌های جانهی چندگانه در ترکیب با معادلات زنجیره‌ای[2] استفاده شده است. در الگوریتم معادلات زنجیره‌ای، هر مقدار گم‌شده با دو یا چند مقدار جانهی از طریق روش‌های رگرسیونی پر می‌شود و هر مجموعه از داده‌های به‌دست‌آمده را با استفاده از

1. Dynamic Time Wrapping (DTW)
2. Multiple Imputation by Chained Equations

روش‌های مربوط به داده‌های کامل تحلیل می‌کنند. داده‌های گم‌شده با داده‌های به‌دست‌آمده در مدل جانهی جایگزین می‌شود. با تکرار m بار این عمل، m مجموعه دادهٔ کامل به‌دست می‌آید. هر یک از این مجموعه‌ها جداگانه تحلیل شده و برآوردی از پارامترهای مدل حاصل می‌شود. مقادیر برآوردشده از mامین مجموعه دادهٔ جانهی‌شده به‌صورت رابطهٔ ۱ نشان داده می‌شود [۳].

$$\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta} \qquad (۱)$$

که در آن $\hat{\theta}$ برآوردگر θ از مجموعه دادهٔ کامل است. برای به‌دست آوردن ضرایب رگرسیون از قوانین ادغام روبین استفاده می‌شود و برآورد پارامترها با میانگین‌گیری از m برآورد به‌صورت رابطهٔ ۲ حاصل می‌شود [۳].

$$\hat{\beta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}^{(k)} \qquad (۲)$$

که در آن $\hat{\beta}$ ضرایب رگرسیونی حاصل از مدل‌های جانهی است. در پایان m نتیجه با محاسبهٔ میانگین، واریانس و دامنهٔ اطمینان متغیر مورد نظر در یک نتیجه تلفیق می‌شود. از مزایای معادلات زنجیره‌ای، امکان چند بار محاسبه و برآورد خطای معیار با استفاده از روشی یکسان است [۳].

جانهی منفرد

برخی الگوریتم‌های جانهی منفرد نیز مبتنی بر ساختار و رفتار سری زمانی هستند. از الگوریتم‌های سادهٔ این دسته می‌توان به روش آخرین مشاهدهٔ رو به جلو[1]، روش میانگین وزن‌دار متحرک[2] و روش‌های درون‌یابی[3] اشاره کرد [۶]. افزون‌بر آن، در این روش‌ها دیگر یک مقدار ثابت برای همهٔ مقادیر جاافتاده جایگزین نمی‌شود. بسته‌های آماری متعددی به‌منظور پیاده‌سازی این الگوریتم‌ها در زبان برنامه‌نویسی R تعبیه شده است که در مقالات به‌کار رفته است [۲۳، ۱۰]. پیاده‌سازی این الگوریتم‌ها نیز سریع و ساده است، اما در مواردی که دامنهٔ اعداد گم‌شده بزرگ باشد، موجب ناپایداری الگوریتم و عدم قطعیت در جواب‌ها می‌شود. [۱۱].

1. Last observation carried forward (LOCF)
2. Weighted moving average
3. Interpolation

داده‌ها و بحث

به‌منظور اعتبارسنجی و بررسی نتایج روش‌های مختلف جانهی باید در داده‌های خام اصلی شبیه‌سازی صورت گیرد. واضح است که عملکرد جانهی افزون‌بر تعداد داده‌های ازدست‌رفته به الگوی داده‌های گم‌شده نیز وابسته است در این پژوهش فرضیهٔ کاملاً تصادفی بودن داده‌ها آزمایش شده است. همچنین نرمال و همگن بودن توزیع داده‌ها نیز بررسی شده است. نتایج حاکی از این است که فرضیهٔ قبول "کاملاً تصادفی" بودن داده‌ها رد شده و بر این اساس سازوکار داده‌های مورد نظر "تصادفی" شناخته شده است [۱۴].

در گام بعدی به بررسی رابطه و شباهت مکانی زمانی سری زمانی آلایندهٔ $PM_{2.5}$ ایستگاه‌های سنجش آلودگی هوا براساس الگوریتم DTW پرداخته شده است. نتایج اعمال الگوریتم DTW به‌صورت جدول ۲ مشاهده می‌شود. در این جدول فواصل DTW دوبه‌دوی همهٔ ایستگاه‌ها محاسبه شده است. بدین ترتیب فاصلهٔ هر دو سری زمانی تعیین می‌شود. ایستگاه‌هایی با فواصل DTW کمتر، شباهت بیشتری از منظر رفتار سری زمانی آلودگی ثبت‌شده در آنها دارند.

جدول ۲. الگوریتم DTW روی آلایندهٔ $PM_{2.5}$ در ۱۲ ایستگاه سنجش آلودگی

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
(S1)	۰	۰/۹۵	۱/۴۴	۱/۷۸	۱/۳۶	۱/۵۹	۱/۳۹	۰/۷	۱/۵۷	۰/۷۵	۱/۴۷	۰/۹۵
(S2)	۰/۹۵	۰	۱/۳۷	۱/۷	۱/۲۹	۱/۵۲	۱/۲۹	۰/۶۸	۱/۴۷	۰/۶۹	۱/۳۴	۰/۸۶
(S3)	۱/۴۴	۱/۳۷	۰	۱/۱۹	۱	۰/۵۹	۰/۸۷	۱/۲۷	۰/۹	۱/۰۶	۰/۹۸	۰/۹۴
(S4)	۱/۷۸	۱/۷	۱/۱۹	۰	۱/۰۵	۰/۵۹	۰/۹۷	۱/۶۴	۱	۱/۵۵	۰/۹۹	۱/۲
(S5)	۱/۳۶	۱/۲۹	۱/۰۲	۱/۰۵	۰	۰/۸	۰/۸۲	۱/۲۶	۰/۸۳	۱/۲	۰/۸۳	۰/۸۹
(S6)	۱/۵۹	۱/۵۲	۰/۹۵	۰/۹۵	۰/۸۲	۰	۰/۷۱	۱/۴۴	۰/۷۸	۱/۳۹	۰/۷۵	۱/۰۱
(S7)	۱/۳۹	۱/۲۹	۰/۸۷	۰/۹۸	۰/۸۱	۰/۷۱	۰	۱/۲۶	۰/۶۹	۱/۱۳	۰/۶	۰/۷۸
(S8)	۰/۷	۰/۶۸	۱/۲۷	۱/۶۴	۱/۲۶	۱/۴۴	۱/۲۶	۰	۱/۳۹	۰/۷۲	۱/۲۹	۰/۸۹
(S9)	۱/۵۷	۱/۴۷	۰/۹	۱	۰/۸۳	۰/۷۸	۰/۶۹	۱/۳۹	۰	۱/۳	۰/۶۶	۰/۸۹
(S10)	۰/۷۵	۰/۶۹	۱/۰۶	۱/۵۵	۱/۲	۱/۳۹	۱/۱۳	۰/۷۲	۱/۳	۰	۱/۱۸	۰/۷۴
(S11)	۱/۴۷	۱/۳۴	۰/۹۸	۰/۹۹	۰/۸۳	۰/۷۵	۰/۶	۱/۲۹	۰/۶۶	۱/۱۸	۰	۰/۷۳
(S12)	۰/۹۵	۰/۸۶	۰/۹۴	۱/۲	۰/۸۹	۱/۰۱	۰/۷۴	۰/۸۹	۰/۸۹	۰/۷۸	۰/۷۳	۰

مطابق با جدول ۲، ایستگاه‌های ری (S4) و اقدسیه (S1)، بیشترین فاصلهٔ DTW را با مقدار عددی ۱/۷۸ دارند. روی نقشه نیز این دو ایستگاه بیشترین فاصله و در نتیجه کمترین شباهت را با هم دارند. در مقابل ایستگاه‌های مدرس (S11) و شریف (S7) با فاصلهٔ DTW با مقدار ۰/۶ کمترین فاصله و بیشترین شباهت را دارند.

پس از تعیین ایستگاه‌های دارای بیشترین شباهت مکانی- زمانی، شبیه‌سازی داده‌های گم‌شده و تولید دادهٔ جدید از داده‌های کامل انجام گرفت. به‌منظور شبیه‌سازی الگوی تصادفی داده‌ها به‌صورت تصادفی بخشی از داده‌ها حذف شده است. با توجه به اینکه سری زمانی داده‌های ساعتی آلودگی دارای فواصلی از داده‌های گم‌شده در بیش از یک ماه نیز هستند، الگوهای فواصل مختلف باید در نظر گرفته شود. بدین منظور بخشی دیگر از داده‌ها به‌صورت فواصل چندروزه و چندساعته به‌صورت تصادفی از داده‌های کامل حذف شد تا الگویی مشابه با الگوی دادهٔ ازدست‌رفتهٔ اصلی ایجاد شود. در نهایت داده‌ها با الگوهای داده‌های ازدست‌رفته ۱۰، ۱۵ و ۲۰ درصد شبیه‌سازی شده‌اند. از طریق نتایج به‌دست‌آمده از جدول ۲ که بیانگر فواصل DTW دوبه‌دوی ایستگاه‌هاست، داده‌های سری زمانی ایستگاه هدف در کنار داده‌های سری زمانی سه ایستگاه مشابه آن به‌ترتیب با کمترین فاصلهٔ DTW قرار گرفته و ورودی مدل‌های جانهی چندگانه را تشکیل داده است.

ارزیابی دقت روش‌های جانهی فرایندی چالش‌انگیز است، زیرا استفاده از روش‌های جانهی به‌معنای وجود داده‌های ناقص است و شبیه‌سازی این داده‌های گم‌شده با هر دو وابستگی زمانی و مکانی با مشکل مواجه است و حتی بهترین مدل نیز نمی‌تواند به قطعیت دینامیک فرایند تصادفی داده‌ها را به تصویر بکشد [۱۳]. در این مقاله به‌منظور ارزیابی دقت و مقایسهٔ روش‌های مختلف جانهی چندگانه و منفرد با در دست داشتن دادهٔ کامل اولیه و دادهٔ جانهی شده با میزان گم‌شدگی ۱۰، ۱۵ و ۲۰ درصد، از شاخص‌های جذر میانگین مربعات خطا[1]، ضریب تعیین[2] و ضریب همبستگی پیرسون[3] استفاده شد [۲۱].

در این پژوهش برای پیاده‌سازی الگوریتم‌های جانهی چندگانه همانند روش تطابق میانگین پیش‌بینی‌کننده، طبقه‌بندی و رگرسیون درختی، نمونهٔ تصادفی و همچنین پیاده‌سازی الگوریتم‌های مختلف جانهی منفرد همانند روش‌های درون‌یابی و آخرین مشاهدهٔ رو به جلو از زبان برنامه‌نویسی R استفاده شد. عملکرد روش‌های جانهی چندگانه طبقه‌بندی رگرسیون درختی، تطابق میانگین پیش‌بینی‌کننده و نمونهٔ تصادفی با الگوی ۱۰، ۱۵ و ۲۰ درصد گم‌شدگی در داده‌ها در جدول ۳ نشان داده شده است. شایان ذکر است که معیارهای ارزیابی داده‌های هر ایستگاه با میانگین سه بار تکرار، جداگانه محاسبه شده و سپس میانگین همهٔ ایستگاه‌ها در جداول مذکور ارائه شده است.

1. RMSE
2. R Squared
3. R coefficient

جدول ۳. عملکرد روش‌های جانهی چندگانه با الگوی دادهٔ گم‌شده ۱۰، ۱۵ و ۲۰ درصد

روش‌های جانهی چندگانه	۲۰ درصد گم‌شدگی			۱۵ درصد گم‌شدگی			۱۰ درصد گم‌شدگی		
	R^2	r	RMSE	R^2	r	RMSE	R^2	r	RMSE
رگرسیون درختی	۰/۵۸	۰/۷۵	۱۱/۷۵	۰/۶	۰/۷۶	۱۰/۸۳	۰/۶۶	۰/۸	۱۱/۷۸
تطابق میانگین پیش‌بینی‌کننده	۰/۵۶	۰/۷۵	۱۱/۹۳	۰/۵۸	۰/۷۵	۱۱/۰۱	۰/۶۴	۰/۷۸	۱۱/۰۲
نمونهٔ تصادفی	۰/۵۵	۰/۷۳	۱۱/۹۶	۰/۵۶	۰/۷۴	۱۱/۲	۰/۶۲	۰/۷۷	۱۱/۳۷

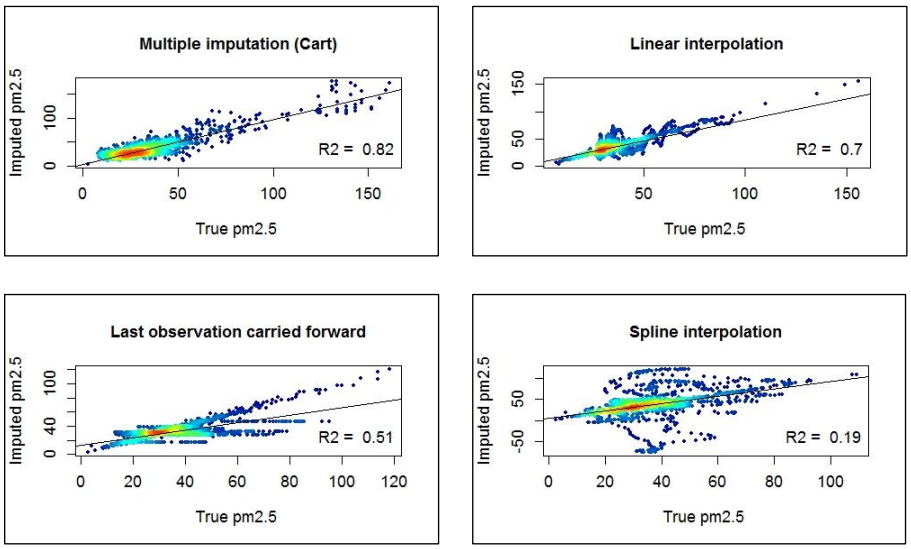
همان‌طور که مشاهده می‌شود، روش جانهی چندگانهٔ رگرسیون درختی با معیار ضریب تعیین ۰/۶۶ و ضریب همبستگی ۰/۸ در ۱۰ درصد گم‌شدگی داده، ضریب تعیین ۰/۶ و ضریب همبستگی ۰/۷۶ در ۱۵ درصد گم‌شدگی داده‌ها، ضریب تعیین ۰/۵۸ و ضریب همبستگی ۰/۷۵ در ۲۰ درصد گم‌شدگی داده‌ها، در بین روش‌های جانهی چندگانه بهترین عملکرد را نشان داده است. واضح است که با افزایش درصد گم‌شدگی داده‌ها دقت معیارهای ارزیابی کاهش می‌یابد. در بین ۱۲ ایستگاه استفاده‌شده در این پژوهش، بهترین نتیجه از ترکیب روش جانهی رگرسیون درختی مطابق با معیار ضریب تعیین، مربوط به ایستگاه مدرس با مقدار عددی ۰/۹۲، ۰/۸۴، ۰/۸۲ و بدترین عملکرد روش ذکرشده، مربوط به ایستگاه مسعودیه با مقدار عددی ضریب تعیین ۰/۳۲، ۰/۲۵، ۰/۲۳ به‌ترتیب با مقدار ۱۰، ۱۵ و ۲۰ درصد گم‌شدگی در داده‌هاست.

مطابق با نتایج ارائه‌شده، روش میانگین تطابق پیش‌بینی‌کننده و روش تصادفی، عملکرد مشابهی نشان داده‌اند و به نسبت روش رگرسیون درختی ضعیف‌تر عمل کرده‌اند. جدول ۴ عملکرد روش‌های جانهی منفرد با روش‌های درون‌یابی خطی، اسپیلاین و روش آخرین مشاهدهٔ رو به جلو را نشان می‌دهد.

جدول ۴. خلاصهٔ عملکرد جانهی منفرد در سه دادهٔ گم‌شده ۱۰، ۱۵ و ۲۰ درصد

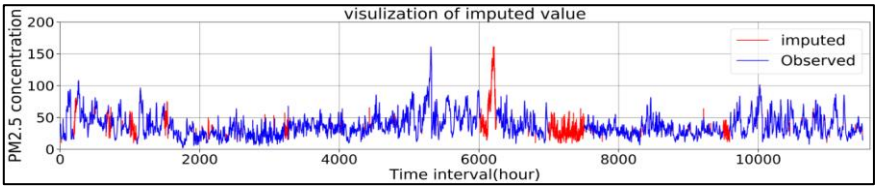
روش‌های جانهی خطی	۲۰ درصد گم‌شدگی			۱۵ درصد گم‌شدگی			۱۰ درصد گم‌شدگی		
	R^2	r	RMSE	R^2	r	RMSE	R^2	r	RMSE
درون‌یابی خطی	۰/۴۵	۰/۶۶	۱۳/۹۴	۰/۵۹	۰/۷۵	۱۰/۹	۰/۶۸	۰/۸۲	۸/۵۳
درون‌یابی اسپیلاین	۰/۰۳	۰/۱۷	۶۷/۰۶	۰/۰۶	۰/۱۹	۵۹/۶۳	۰/۱۴	۰/۳۴	۳۱/۴۲
آخرین مشاهده رو به جلو	۰/۳۷	۰/۶	۱۶/۳۴	۰/۴۲	۰/۶۳	۱۲/۴۵	۰/۵۵	۰/۷۳	۱۰/۸۱

مطابق با جدول ۴، براساس هر سه معیار ارزیابی، روش درون‌یابی خطی از دیگر روش‌های ارائه‌شده بهتر بوده است. از این‌رو در میان روش‌های منفرد برای داده‌های ارائه‌شده، این روش مناسب‌تر است. همچنین روش درون‌یابی اسپیلاین ضعیف‌ترین عملکرد را در بین همهٔ روش‌های جانهی چندگانه و منفرد از خود نشان داده است. اگرچه در داده‌هایی با ۱۰ درصد گم‌شدگی، روش درون‌یابی خطی در مقایسه با روش رگرسیون درختی در شاخص‌های ارزیابی بیشترین ضریب تعیین و همبستگی و کمترین خطا را کسب کرده است، اما باید به این نکته توجه داشت که روش درون‌یابی خطی برای مقادیر گم‌شده با بازهٔ کم، عملکرد بسیار مناسبی نشان می‌دهد، اما هنگامی‌که بازهٔ گم‌شدگی داده‌ها بیشتر شود، برای مثال در بازهٔ مربوط به گم‌شدگی ۲۰ درصد این روش‌ها قادر به جانهی مناسب برای داده‌های گم‌شده نیستند و یک مقدار ثابت یا با تغییرات کم را برای تمام مقادیر گم‌شده در هر بازه در نظر می‌گیرد. شکل ۲ نمودار پراکندگی مقادیر مشاهداتی در مقابل مقادیر جانهی شده با روش‌های رگرسیون درختی، درون‌یابی خطی درون‌یابی اسپیلاین و آخرین مشاهدهٔ رو به جلو در ۱۰ درصد گم‌شدگی در ایستگاه منطقهٔ ۲۱ را نشان می‌دهد. این نمودارها ارتباط بین مقادیر مشاهداتی و مقادیر جانهی شده را به‌صورت یک معادله خطی درجه یک نشان می‌دهد.

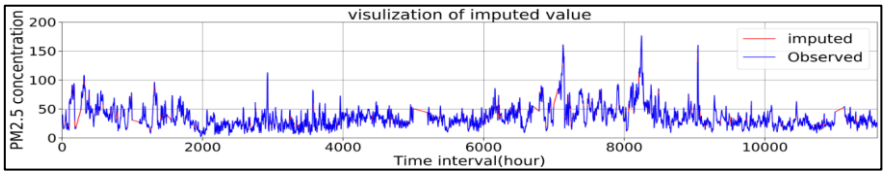


شکل ۲. نمودار مقدار مشاهداتی در مقابل مقدار جانهی‌شدهٔ ایستگاه شهرداری منطقهٔ ۲۱

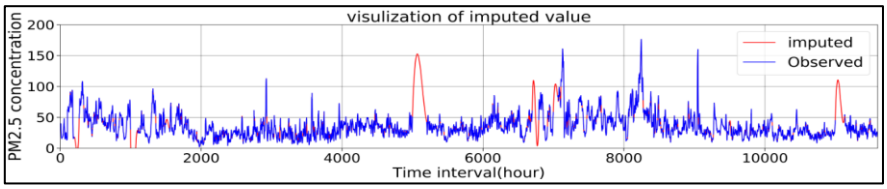
در شبیه‌سازی داده‌های گم‌شده به‌دلیل اینکه بخشی از داده‌ها به‌صورت تصادفی با بازهٔ یک تا سه رکورد حذف شده‌اند، پر کردن این داده‌ها توسط روش درون‌یابی خطی با دقت زیادی انجام گرفته است از این‌رو در مجموع دقت روش درون‌یابی خطی در ۱۰ درصد گم‌شدگی در داده‌ها بیشتر از روش‌های جانهی چندگانه برآورد شده است. شکل ۳ نمایش بصری الگوی حاصل از اعمال روش‌های مختلف جانهی برای کامل کردن گم‌شدگی‌های ۱۰ درصد را نشان می‌دهد.



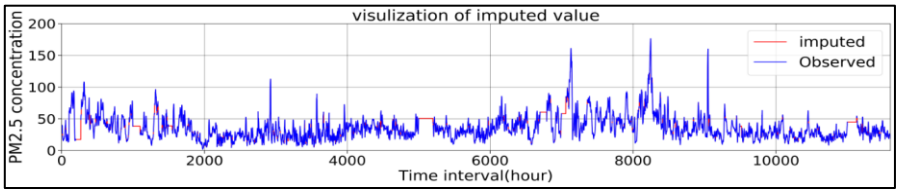
(الف)



(ب)



(پ)



(ج)

شکل ۳. جانهی چندگانه و منفرد مقادیر $PM_{2.5}$ ایستگاه شهرداری منطقهٔ ۲۱ با گم‌شدگی ۱۰ درصد داده‌ها: (الف) جانهی چندگانهٔ روش رگرسیون درختی؛ (ب) درون‌یابی خطی؛ (پ) درون‌یابی اسپیلاین؛ (ج) جانهی منفرد روش آخرین مشاهدهٔ رو به جلو

همان‌طور که در شکل ۳ نیز دیده می‌شود، الگوی حاصل از روش رگرسیون درختی بیش از روش‌های دیگر با الگوی موجود در سری زمانی تطابق دارد.

نتیجه‌گیری

وجود داده‌های گم‌شده در سری زمانی غلظت آلاینده‌ها بر عملکرد تجزیه‌وتحلیل داده‌ها در الگوریتم‌های یادگیری ماشین تأثیر منفی می‌گذارد و موجب بایاس می‌شود. روش‌های متعددی برای جانهی داده‌های گم‌شده ارائه‌شده است که انتخاب بهترین روش نیازمند مطالعه است. نتایج این پژوهش نشان داده است که تعیین شباهت مکانی- زمانی ایستگاه‌ها و استفاده از الگوی ایستگاه‌های مشابه با استفاده از الگوریتم پیچش زمانی پویا در ترکیب با روش‌های رگرسیون‌مبنا، موجب بهبود عملکرد مدل با فواصل زیاد گم‌شدگی می‌شود و مدل رگرسیون درختی مناسب‌ترین روش برای جانهی چندگانه به‌شمار می‌رود. روش‌های جانهی منفرد، با وجود سرعت و سادگی، وابسته به طول بازهٔ گم‌شدگی در زمان هستند و عملکرد آنها به متغیر بررسی‌شده بستگی دارد. بنابراین استفاده از روش‌های منفرد در داده‌های آلودگی هوا که فواصل گم‌شدگی زیاد دارند توصیه نمی‌شود. شایان ذکر است که در هر منطقهٔ مطالعاتی که با استفاده از الگوریتم پیچش زمانی پویا بتوان شباهت بین ایستگاه‌های مختلف را تشخیص داد، مدل قابلیت استفاده برای جانهی مقادیر ازدست‌رفتهٔ غلظت آلاینده‌ها در آن منطقهٔ مطالعاتی را خواهد داشت. با توجه به تأثیر دیگر عوامل مانند پارامترهای هواشناسی بر آلودگی هوا، در پژوهش‌های آینده می‌توان با افزودن این پارامترها دقت مدل را افزایش داد.

منابع

[۱]. بازگیر، سعید؛ قدیری معصوم، مجتبی؛ شمسی‌پور، علی‌اکبر؛ و سیدی سرنجیانه، شیوا (۲۰۱۵). «تحلیل رابطهٔ آلودگی هوای تهران با ترافیک و شرایط جو برای کاهش مخاطرات، *مدیریت مخاطرات محیطی*، دوره ۲، شمارهٔ ۱. ص ۴۹-۳۵.

[۲]. باقی یزدل، رقیه؛ جمالی، احسان؛ خدایی، ابراهیم؛ و حبیبی مجتبی (۱۳۹۵). «روش‌های برخورد با داده‌های گمشده: مزایا، معایب، رویکردهای نظری و معرفی نرم‌افزارها». *نامهٔ آموزش عالی*، دورهٔ ۹، شمارهٔ ۳۳، ص ۳۷-۱۱.

[۳]. عثمانی، فرشته؛ و راسخی، علی‌اکبر (۱۳۹۷). «روش‌های وزن‌دهی احتمال معکوس و جانهی چندگانه برای تحلیل پاسخ در حالت گم‌شدگی». *علوم آماری*، دورهٔ ۱۲ شمارهٔ ۲، ص ۴۸۳-۴۶۹.

[۴].کرمانی، آذر؛ اکبری، مهری؛ علیجانی، بهلول؛ و مفاخری، امید (۲۰۱۵). «تحلیل آماری-همدیدی غلظت آلایندهٔ مونواکسیدکربن براساس سمت و سرعت باد و مخاطرهٔ آن در شهر تهران». *مدیریت مخاطرات محیطی*، دورهٔ ۲ شمارهٔ ۴، ص ۴۵۰-۴۳۹.

[5]. Burgette, L.F.; & Reiter, J.P. (2010). "Multiple imputation for missing data via sequential regression trees", *American journal of epidemiology*, 172(9), pp: 1070-1076. doi: https://doi.org/10.1093/aje/kwq260.

[6]. Caillault, É.P.; Lefebvre, A.; & Bigand, A. (2017). "Dynamic time warping-based imputation for univariate time series data", *Pattern Recognition Letters*. doi:https://doi.org/10.1016/j.patrec.2017.08.019.

[7]. Chen, X.; & Xiao, Y. (2018). "A novel method for air quality data imputation by nuclear norm minimization", *Journal of Sensors*. doi:https://doi.org/10.1155/2018/7465026.

[8]. Erler, N.S.; Rizopoulos, D.; Jaddoe, V.W.; Franco, O.H.; & Lesaffre, E.M. (2019). "Bayesian imputation of time-varying covariates in linear mixed models", *Statistical methods in medical research*, 28(2), pp: 555-568. doi:https://doi.org/10.1177/0962280217730851.

[9]. Fortuin, V.; Rätsch, G.; & Mandt, S. (2019). "Multivariate time series imputation with variational autoencoders", *arXiv preprint arXiv:1907.04155*. doi: https://arxiv.org/abs/1907.04155..

[10]. Ghazali, S.M.; Shaadan, N.; & Idrus, Z. (2020). "Missing data exploration in air quality data set using R-package data visualisation tools", *Bulletin of Electrical Engineering and Informatics*, 9(2), pp: 755-763. doi:https://doi.org/10.11591/eei.v9i2.2088.

[11]. Gómez-Carracedo, M.; Andrade, J.; López-Mahía, P.; Muniategui, S.; & Prada, D. (2014). "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets", *Chemometrics and Intelligent Laboratory Systems*, 134, pp: 23-33. doi:https://doi.org/10.1016/j.chemolab.2014.02.007.

[12]. Hadeed, S.J.; O'Rourke, M.K.; Burgess, J.L.; Harris, R.B.; & Canales, R.A. (2020). "Imputation methods for addressing missing data in short-term monitoring of air pollutants", *Science of The Total Environment*, pp: 139140. doi:https://doi.org/10.1016/j.scitotenv.2020.139140.

[13]. Junger, W.; & De Leon, A.P. (2015). "Imputation of missing data in time series for air pollutants", *Atmospheric Environment*, 102, pp: 96-104. doi:https://doi.org/10.1016/j.atmosenv.2014.11.049.

[14]. Junninen, H.; Niska, H.; Tuppurainen, K.; Ruuskanen, J.; & Kolehmainen, M. (2004). "Methods for imputation of missing values in air quality data sets", *Atmospheric Environment*, 38(18), pp: 2895-2907. doi:https://doi.org/10.1016/j.atmosenv.2004.02.026.

[15]. Lin, J.; Li, N.; Alam, M.A.; & Ma, Y. (2020). "Data-driven missing data imputation in cluster monitoring system based on deep neural network", *Applied Intelligence*, 50(3), pp: 860-877. doi:https://doi.org/10.1007/s10489-019-01560-y.

[16]. Liu, X.; Wang, X.; Zou, L.; Xia, J.; & Pang, W. (2020). "Spatial imputation for air pollutants data sets via low rank matrix completion algorithm", *Environment International*, 139, pp: 105713. doi:https://doi.org/10.1016/j.envint.2020.105713.

[17]. Ma, J.; Cheng, J.C.; Jiang, F.; Chen, W.; Wang, M.; & Zhai, C. (2020). "A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data", *Energy and Buildings*, pp: 109941. doi:https://doi.org/10.1016/j.enbuild.2020.109941.

[18]. Mishra, S.; Dwivedi, V.; Sarvanan, C.; & Pathak, K. (2013). "Pattern discovery in hydrological time series data mining during the monsoon period of the high flood years in Brahmaputra River basin", *International Journal of Computer Applications*, 67(6).

[19]. Raghunathan, T.E.; Lepkowski, J.M.; Van Hoewyk, J.; & Solenberger, P. (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey methodology*, 27(1), pp: 85-96.

[20]. Rombach, I.; Gray, A.M.; Jenkinson, C.; Murray, D.W.; & Rivero-Arias, O. (2018). "Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level", *BMC medical research methodology*, 18(1), pp: 87. doi:https://doi.org/10.1186/s12874-018-0542-6.

[21]. Shahbazi, H.; Karimi, S.; Hosseini, V.; Yazgi, D.; & Torbatian, S. (2018). "A novel regression imputation framework for Tehran air pollution monitoring network using outputs from WRF and CAMx models", *Atmospheric Environment*, 187, pp: 24-33. doi:https://doi.org/10.1016/j.atmosenv.2018.05.055.

[22]. Stead, A.D.; & Wheat, P. (2020). "The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using English local highway data", *European Journal of Operational Research*, 280(1), pp: 59-77. doi:https://doi.org/10.1016/j.ejor.2019.06.042.

[23]. Zeileis, A.; Grothendieck, G.; Ryan, J.A.; Andrews, F.; & Zeileis, M.A. (2019). "Package "zoo"".