مجله ژئوفیزیک ایران، جلد ۱۶، شماره ۱، ۱۴۰۱، صفحه ۱۴۱ –۱۵۰ (مقاله مروری) DOI: 10.30499/IJG.2021.304252.1364

کنترل کیفی داده های اوزن سطحی اندازه گیری شده در ایستگاه های شهر تهران با کمک نرم افزار آماری جدید

نجمه کفاش زاده [۱*] و عباسعلی علی اکبری بیدختی [۲]

[۱] پژوهشگر پسادکتری، گروه فیزیک فضا، مؤسسه ژئوفیزیک دانشگاه تهران، تهران، ایران
[۲] استاد، گروه فیزیک فضا، مؤسسه ژئوفیزیک دانشگاه تهران، تهران، ایران

(دریافت: ۱۴۰۰/۰۶/۰۲۸، پذیرش: ۱۴۰۰/۰۹/۲۵)

چکیده

بی‌توجهی به وجود خطاهای متعدد شامل خطای فاحش، اعداد ثابت و غیره در داده می تواند به نتایج نادرست در تحلیل داده ها منجر شود؛ ازاین رو کنترل کیفی داده گامی ضروری جهت حصول اطمینان از صحت داده است. در دسترس نبودن واقعیت، سبب پیچیدگی در تشخیص خطا و انجام دادن کنترل کیفی داده می شود. روش ها و آزمایش های آماری گوناگونی برای کنترل کیفی داده وجود دارد، ولی هیچ یک یافتن تمامی خطاها را در داده ضمانت نمی کنند. اجرای هرچه بیشتر آزمایش ها سبب افزایش اطمینان نسبی از کیفیت داده می شود. در این مطالعه به دلیل اهمیت و ضرورت مطالعه آلاینده ازن سطحی، کیفیت این داده ها در سطح شهر تهران بررسی شد. کنترل کیفی داده ها با استفاده از ابزار AutoQA4Env انجام شد. این ابزار متشکل از مجموعه آزمایش های آماری گروه‌بندی شده در دو حالت پایه و پیشرفته است. از ویژگی های خاص این ابزار، تنظیمات کاربری، تکرارپذیری و گسترش پذیری آن است. نتایج اجرای این ابزار در حالت پایه، حاکی از وجود خطای فاحش در برخی از داده ها بود که این موضوع به‌منزله لزوم بررسی کنترل کیفی داده پیش از به‌کارگیری آن است. از طرف دیگر، در برخی موارد نشان داده شد اجرای ابزار در حالت پایه کافی نیست و کاربست ابزار در حالت پیشرفته مناسب تر است.

واژه‌های کلیدی: کنترل کیفی، داده های ازن سطحی، خطای فاحش، ابزار AutoQA4Env

* نگارنده رابط: n.kaffashzadeh@ut.ac.ir

۱ مقدمه

ازن سطحی از آلاینده‌های خطرناکی است که سالانه به خسارات زیاد ازجمله مرگ‌ومیر هزاران جاندار و از بین رفتن محصولات کشاورزی منجر می‌شود (مونکس و همکاران، ۲۰۱۵). سری زمانی ازن سطحی اندازه‌گیری‌شده با ابزار سنجش زمینی، یکی از منابع مهم تحقیقاتی برای بررسی این آلاینده به‌شمار می‌رود (سفن و همکاران، ۲۰۱۶). این داده‌ها حاوی اطلاعات باارزشی شامل چگونگی تغییرپذیری و روند این آلاینده هستند که از آنها در برنامه‌ریزی کنترل آلودگی هوا استفاده می‌شود. یکی از مشکلات این داده‌ها وجود خطا در آنهاست. وقوع خطا در داده موضوعی انکارناپذیر است که ناشی از عوامل طبیعی و غیرطبیعی است. از جمله عوامل طبیعی می‌توان به سیل، آتش‌سوزی و وزش بادِ شدید اشاره کرد (کمپل، ۲۰۱۳). عوامل غیرطبیعی که می‌تواند عمدی یا غیرعمدی باشد، شامل خطای انسانی حین جمع‌آوری، ثبت یا اندازه‌گیری داده است.

خطا به معنای انحراف از واقعیت است. هر مقدار اندازه‌گیری‌شده، برآوردی از واقعیت است که از مجموعه‌ای از نمونه داده به‌دست‌آمده‌است (لورنتس، ۱۹۸۱). خطاهای موجود در داده از انحراف‌های متعددی ناشی می‌شوند که از آن جمله خطای تصادفی است که به مجموعه خطاهای ذاتی داده به دلیل نبود امکان اندازه‌گیری متغیر با سامانه مشاهداتی اشاره دارد. نوع دیگر خطا، خطای غیرِمعرف (non-representative) است که به دلیل پدیده‌های خردمقیاس نظیر تلاطم و اغتشاشات کوچک جوّی به‌وجودمی‌آید. نبود کالیبراسیون یا نبود کارکرد و رانش طولانی‌مدت دستگاه اندازه‌گیری سبب به‌وجودآمدن نوع سوم خطا یعنی خطای نظام‌مند (systematic error) می‌شود (زوربنکو و همکاران، ۱۹۹۶). خطای فاحش (gross error) خطایی ناشی از خرابی دستگاه یا اشتباه در پردازش و انتقال داده است. خطاهای دسته اول و دوم توزیع گوسی و تقریباً نظام‌مند دارند، اما در زمان پایدار نیستند (اشتانِکر و همکاران، ۲۰۱۱)؛ بنابراین تشخیص و حذف این خطاها در داده به دلیل رفتارهای آشفته و غیرنظام‌مند آنها ساده نیست. از طرف دیگر، خطاهای نوع سوم و چهارم مهم‌ترین انواع خطا هستند. برخلاف خطای اول و دوم، خطای نظام‌مند در زمان پایدار است پس می‌توان آن را در مجموعه داده تشخیص داد (گاندین، ۱۹۸۸ و زوربنکو و همکاران، ۱۹۹۶). دسته چهارم خطا به‌ندرت رخ می‌دهد و تشخیص آن به نظر دشوار است، ولی ازآنجاکه رفتار این نوع خطا با سایر داده‌ها کاملاً متفاوت است، تشخیص آنها امکان‌پذیر است؛ برای مثال این اعداد خطادار، مقادیر بسیار زیاد یا بسیار کم دارند. در این مطالعه فقط به بررسی خطای سوم و چهارم پرداخته می‌شود.

بی‌توجهی به خطا در داده می‌تواند بر تمام بررسی‌ها حتی بر یک محاسبه آماری ساده نظیر میانگین‌گیری اثرگذار باشد و به اطلاعات گمراه‌کننده منجر شود (عابدینی و همکاران، ۱۳۸۲ و اُزبرنه و اورِ بِی، ۲۰۰۴)؛ ازاین‌رو، کاربر داده پیش از استفاده از داده‌ها باید از صحت و کنترل کیفی آنها مطمئن شود (زاهومنسکی، ۲۰۱۶). اغلب روش‌های کنترل کیفی بر مشاهده داده‌ها و حذف دستی مقادیر اشتباه استوار هستند. این روش می‌تواند بسیاری از خطاها را در داده پیدا و حذف کند، ولی در مجموعه (پایگاه) داده‌های بزرگ، ناکارآمد است و حتی سبب بروز خطا می‌شود. پس استفاده از یک ابزار (نرم‌افزار) برای کنترل کیفی داده ضروری است. یکی از ابزارهای جدید در این راستا AutoQA4Env است که کفاش‌زاده و شولتز (۲۰۲۰ الف، ب) در مرکز تحقیقاتی یولیش کشور آلمان جهت کنترل کیفی خودکار داده توسعه داده‌اند. این ابزار متشکل از مجموعه آزمایش‌های آماری است که به چندین گروه یعنی g0، g1 و غیره

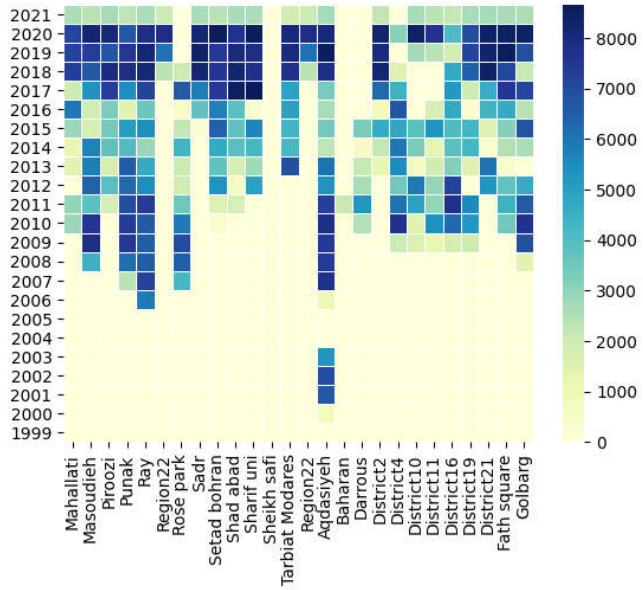
طبقه‌بندی شده است. گروه‌بندی این آزمایش‌ها بر اساس نیاز و هدف مطالعه است که کاربرمی‌تواند آن را تنظیم کند. افزایش تعداد آزمایش‌های آماری، به دشوار کردن وضعیت آزمایش و افزایش احتمالی کیفیت داده منجر می‌شود. این ابزار منبع‌باز در دو حالت پایه و پیشرفته در دسترس است (کفاش‌زاده و شولتز، ۲۰۲۰ الف، ب). گفتنی است روش‌های زیادی برای شناسایی و حتی تصحیح داده‌های اشتباه وجود دارد (تن هوآ و همکاران، ۲۰۱۰ و اسکولی- آلیسن و همکاران، ۲۰۱۸)، اما هیچ روشی نمی‌تواند تشخیص کامل خطا را تضمین کند. کمترین مزایای استفاده از این ابزار، توانایی نگهداری، انعطاف‌پذیری، تکرارپذیری و اطمینان نسبی از کیفیت داده‌ها است. این ابزار در مخزن گیت (Git) نگهداری می‌شود و کاربر می‌تواند راحت به آن دسترسی یابد و حتی در توسعه آن همکاری کند. از این ابزار در چرخه پایگاه داده توآر (TOAR) نیز استفاده شده است. این پایگاه یکی از بزرگ‌ترین پایگاه‌های جهانی داده اندازه‌گیری کیفیت هوا (شولتز و همکاران، ۲۰۱۷) است.

در این تحقیق به بررسی کنترل کیفی داده‌های ازن سطحی اندازه‌گیری‌شده در ایستگاه‌های سنجش آلودگی شهر تهران پرداخته شده است. ابزار AutoQA4Env در دو حالت پایه و پیشرفته برای مجموعه‌ای از این داده‌ها جهت نمونه اجرا و نتایج آن نشان داده شده است. در بخش ۲ مروری بر داده‌ها و در بخش ۳ نتایج کنترل کیفی آنها ارائه شده است. یافته‌ها و نتیجه‌گیری‌های این مطالعه در بخش ۴ خلاصه شده است.

۲ مروری بر داده‌ها

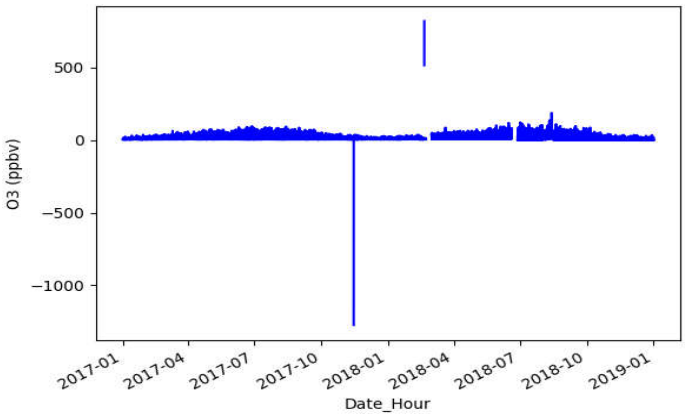
داده‌های سری زمانی ازن سطحی در تمامی ایستگاه‌های شهر تهران از سامانه کنترل کیفی هوای شهر تهران (http://airnow.tehran.ir/home/OnlineAQI.aspx) دریافت شد. این داده‌ها از لحاظ ساختار فایل شامل رسم‌الخط، تاریخ و زمان و ... هماهنگ بودند. تعداد این ایستگاه‌ها ۲۳ تا است و پراکندگی جغرافیایی آنها در شکل ۱ نشان داده شده است. پراکندگی زمانی داده‌ها با یکدیگر متفاوت است؛ ایستگاه اقدسیه، بیشترین و ایستگاه بهاران، کمترین تعداد داده ازن سطحی را دارند (شکل ۲).

شکل ۱. پراکندگی جغرافیایی ایستگاه‌های اندازه‌گیری ازن سطحی در شهر تهران (برگرفته از ویگاه شرکت کنترل کیفی هوا).

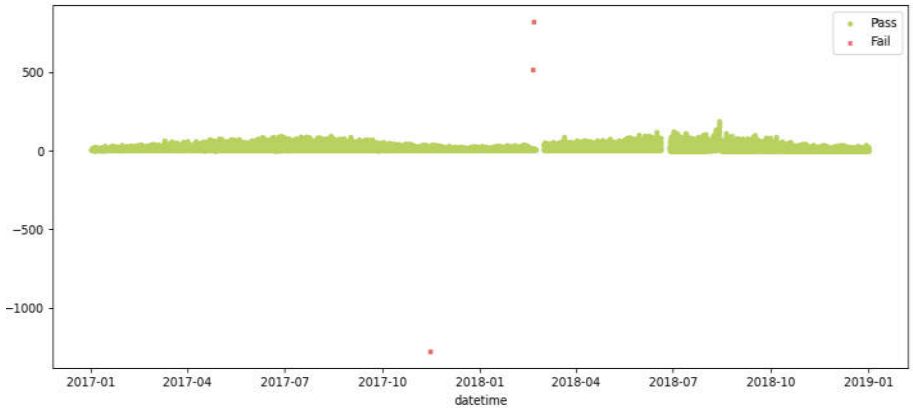


شکل ۲. پراکندگی زمانی داده‌های ازن سطحی در ایستگاه‌های سنجش آلودگی هوای شهر تهران.

در شکل ۳ سری زمانی ازن سطحی در یکی از این ایستگاه‌ها برای مثال نشان داده شده است. در این شکل به‌وضوح دو عدد مثبت بسیار بزرگ یعنی (ppbv) ۷۸/۵۱۵ و (ppbv) ۸۲۱ به‌ترتیب در تاریخ ۱۹ فوریه ۲۰۱۸ ساعت ۱۴ و ۱۵ به وقت محلی دیده می‌شود. در این شکل در تاریخ ۱۴ نوامبر ۲۰۱۷ ساعت ۱۳:۰۰ به وقت محلی، عدد بسیار منفی (ppbv) ۱۲۷۳- نیز به‌چشم‌می‌خورد. هر سه عدد خارج از محدوده نرمال مقادیر ازن سطحی در این ایستگاه هستند؛ ازاین‌رو این اعداد، خطای فاحش محسوب می‌شوند و نادیده گرفتن آنها به بروز اشتباه در نتایج و محاسبات حتی برای یک میانگین‌گیری ساده منجر می‌شود. این نوع خطا و دیگر خطاها در این داده‌ها کم‌وبیش وجود دارند که در بخش بعد شرح داده می‌شود.



شکل۳. سری زمانی ازن سطحی اندازه‌گیری‌شده در یکی از ایستگاه‌های شهر تهران. گفتنی است این ایستگاه به‌طور تصادفی انتخاب شده است؛ لذا از ذکر نام ایستگاه خودداری شده است.



	the number of pass (%)	the number of fail (%)
g0_range_test	[16391, ' (99.98)']	[3, ' (0.02)']
g1_constant_value_test	[16394, ' (100.0)']	[0, ' (0.0)']
g1_negative_value_test	[16394, ' (100.0)']	[0, ' (0.0)']

شکل ٤. نتایج کنترل کیفی سری زمانی ازن سطحی (شکل ۳). این خروجی حاصل اجرای ابزار AutoQA4Env است و شامل یک نمودار (بالا) و یک جدول (پایین) است. آزمایش‌های آماری اجراشده و نتایج آنها در قالب تعداد و درصد در جدول ذکر شده است.

۳ نتایج کنترل کیفی داده‌ها

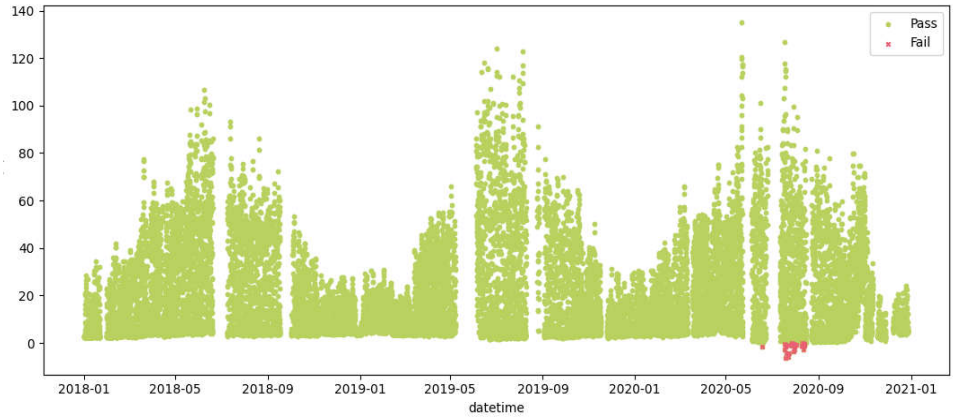
یکی از مراحل مهم در زنجیره جمع‌آوری داده، کنترل‌کیفی است که در آن خطا در داده پیدا و گزارش می‌شود. همان‌طور که در بخش‌های پیشین اشاره شد، خطاها به دلایل متفاوت در داده به‌وجود می‌آیند. شایان ذکر است خطاهای موجود در داده همیشه به‌راحتی و به‌وضوح در داده‌ها نمایان نمی‌شوند؛ برای مثال در برخی موارد، خطا به‌صورت یک سری اعداد ثابت پشت‌سرهم در داده‌ها وجود دارد که با مشاهده نمی‌توان تشخیص داد. خوشبختانه امروزه با پیشرفت فناوری و در دسترس بودن رایانه‌ها ابزارهای ویژه‌ای جهت کنترل کیفی داده ت وسعه یافته‌اند. استفاده از این ابزارها نه تنها سبب سهولت و تسریع کنترل کیفی داده می‌شود، بلکه موجب افزایش دقت در تشخیص این خطاها نیز می‌شود. در این بخش به معرفی و استفاده از ابزار AutoQA4Env پرداخته می‌شود که از آن برای کنترل کیفی داده‌های مورد مطالعه استفاده شده است.

۳-۱ حالت پایه

شکل ۴ خروجی حاصل از اجرای AutoQA4Env را برای نمودار سری زمانی شکل ۳ نشان می‌دهد. در این حالت، خروجی به‌صورت یک سامانه نشان‌گذاری دوتایی (باینری) یعنی قبول (Pass) و قبول‌نشدنی (Fail) است. در این شکل سه عدد که در بخش ۲ از آنها با نام خطا یاد شد، داده قبول‌نشدنی تشخیص داده شده‌اند (علامت ضربدر قرمزرنگ). علاوه‌بر نمودار سری زمانی، این خروجی حاوی یک جدول است که در آن آزمایش‌های آماری اجرا و نتایج گزارش شده است. مطابق این جدول، سه آزمایش آماری شامل محدوده، اعداد ثابت و اعداد منفی اجرا شده است. در آزمایش اول (محدوده)، سه داده (۰/۰۲ درصد) قبول‌نشدنی تشخیص داده شدند. دیگر اعداد از آزمایش‌های آماری دوم و سوم عبورکرده‌اند و قبول هستند.

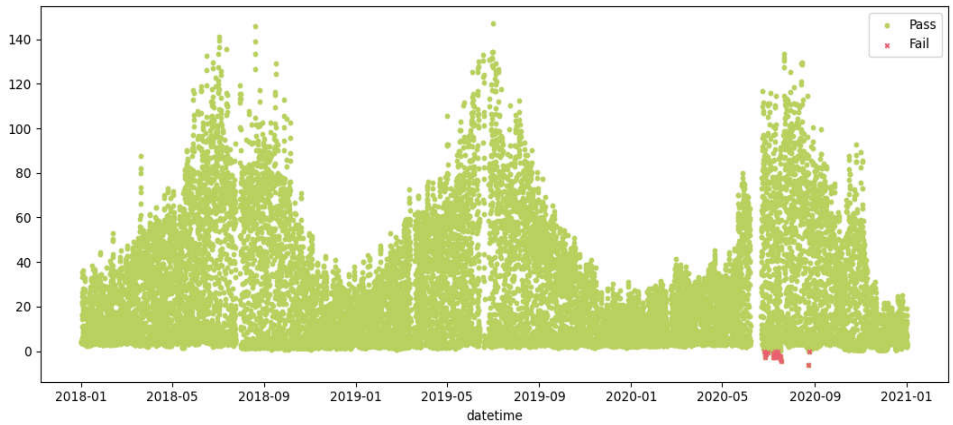
این ابزار برای سری زمانی داده‌های ازن سطحی در دیگر ایستگاه‌ها نیز اجرا شد. نتایج برای دو ایستگاه در شکل‌های ۵ و ۶ نشان داده شده است. همان‌گونه‌که این

نمودارها نشان می‌دهند، تعداد ۲۷ و ۲۳ عدد در آزمایش آماری اول داده قبول‌نشدنی محسوب می‌شوند. بیشتر این اعداد منفی هستند و نکته درخور توجه، وقوع آنها در یک بازه زمانی خاص یعنی ماه ژولای ۲۰۲۰ است. گفتنی است در برخی از ایستگاه‌ها وضعیت مشابهی مشاهده می‌شود؛ بنابراین نشان‌گذاری این ارقام با نام داده قبول‌نشدنی منطقی به نظر نمی‌رسد و اجرای AutoQA4Env در حالت پیشرفته برای این داده‌ها مناسب‌تر است.



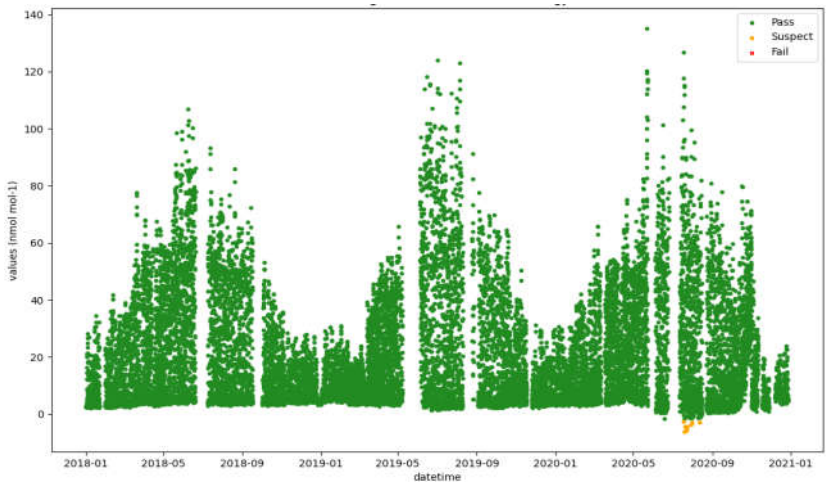
	the number of pass (%)	the number of fail (%)
g0_range_test	[21535, ' (99.87)']	[27, ' (0.13)']
g1_constant_value_test	[21562, ' (100.0)']	[0, ' (0.0)']
g1_negative_value_test	[21562, ' (100.0)']	[0, ' (0.0)']

شکل۵. نتایج کنترل کیفی سری زمانی ازن سطحی در یکی از ایستگاه‌ها. این خروجی حاصل اجرای ابزار AutoQA4Env در حالت پایه است.

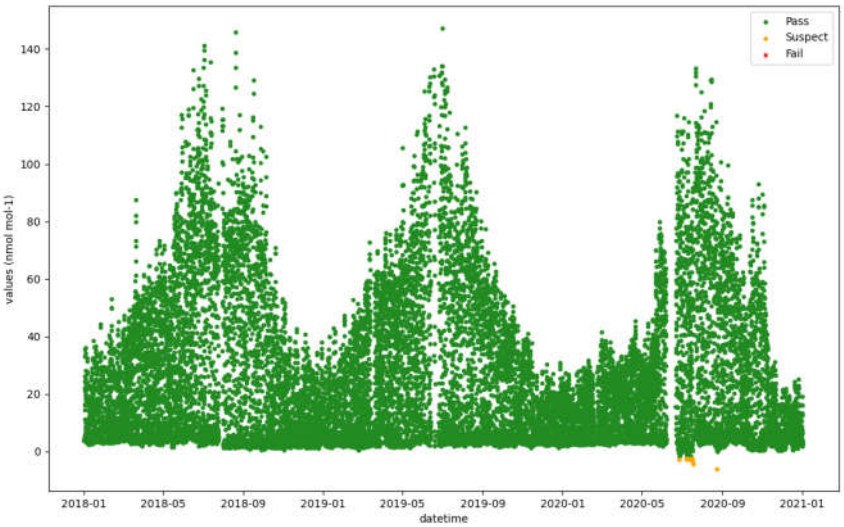


	the number of pass (%)	the number of fail (%)
g0_range_test	[23727, ' (99.9)']	[23, ' (0.1)']
g1_constant_value_test	[23750, ' (100.0)']	[0, ' (0.0)']
g1_negative_value_test	[23750, ' (100.0)']	[0, ' (0.0)']

شکل٦. نتایج کنترل کیفی ازن سطحی در سال‌های ۲۰۱۸، ۲۰۱۹ و ۲۰۲۰ که در یکی از ایستگاه‌های سنجش آلودگی هوای شهر تهران اندازه‌گیری شده است. این خروجی حاصل اجرای ابزار AutoQA4Env در حالت پایه است.



شکل۷. خروجی اجرای ابزار AutoQA4Env در حالت پیشرفته برای سری زمانی ازن سطحی (شکل ۵).



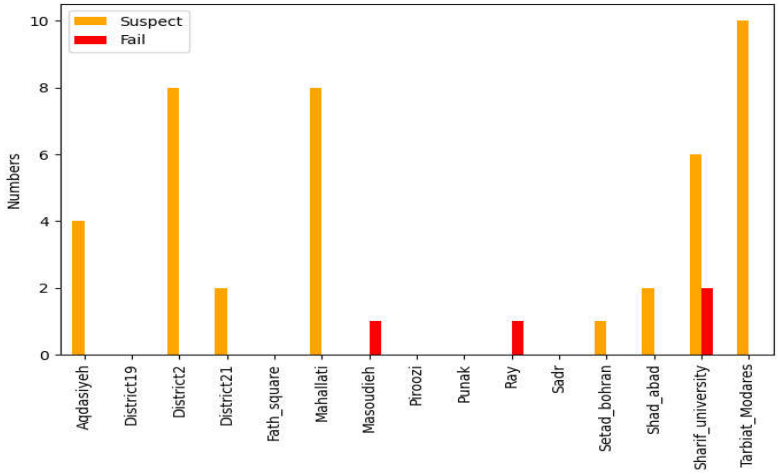
شکل۸. خروجی اجرای ابزار AutoQA4Env در حالت پیشرفته برای سری زمانی ازن سطحی (شکل ٦).

۳-۲ حالت پیشرفته

در حالت پیشرفته، سامانه نشان‌گذاری ابزار AutoQA4Env شامل سه حالت قبول، قبول‌نشدنی و مشکوک است. این ابزار در حال حاضر فقط شامل یک آزمایش آماری محدود برای تشخیص خطای فاحش است که تنظیمات کاربری آن بر اساس پیشنهادهای دیگر مطالعات (سامانه یکپارچه مشاهدات اقیانوسی آمریکا، ۲۰۱۴) توسعه یافته است؛ برای مثال در این حالت کاربر ملزم به وارد کردن دو آستانه علمی و تجربی در بخش تنظیمات است. این ابزار برای دو سری زمانی نشان‌داده‌شده در شکل‌های ۵ و ۶ اجرا شد که نتایج آن به‌ترتیب در شکل‌های ۷ و ۸ ارائه شده است. در این دو شکل مشاهده می‌شود برخی از اعداد قبول‌نشدنی در حالت پایه، قبول و برخی دیگر مشکوک تشخیص داده شده‌اند. در این حالت تمامی اعداد منفی که قدر مطلق آنها زیاد نیست، قبول‌نشدنی تشخیص داده نشده‌اند؛ برای

نمونه، در برخی ایستگاه‌ها به‌ویژه در مکان‌هایی با آمدوشد زیاد، اعداد منفی اندازه‌گیری‌شده برای متغیر ازن سطحی که بین صفر و ppbv ۲- هستند، به دلیل دقت اندازه‌گیری، قبول هستند (شولتز و همکاران، ۲۰۱۷).

نتایج کنترل کیفی اجرای ابزار AutoQA4Env برای داده‌های پانزده ایستگاه که سه سال داده متوالی دارند، خلاصه‌وار در نمودار میله‌ای شکل ۹ نشان داده شده است. همان‌گونه که این شکل نشان می‌دهد، در پنج ایستگاه هیچ‌گونه خطایی در داده‌ها یافت نشده است و خطاها اغلب به‌صورت داده مشکوک در ده ایستگاه نمایان شده‌اند. تعداد داده‌های قبول‌نشدنی در این حالت، ۶/۵ درصد تعداد داده‌های قبول‌نشدنی در حالت پایه است. ۷/۳۶ و ۷/۵۷ درصد داده‌های قبول‌نشدنی در حالت پایه، به‌ترتیب داده قبول و مشکوک تشخیص داده شده است؛ بنابراین عملکرد این ابزار در کنترل کیفی داده‌ها بسیار موفق به‌نظرمی‌رسد.



شکل ۹. خلاصه کنترل کیفی حاصل از اجرای ابزار AutoQA4Env برای داده‌های ازن سطحی اندازه‌گیری‌شده در ایستگاه‌های شهر تهران طی سال‌های ۲۰۱۸، ۲۰۱۹ و ۲۰۲۰. کلمات Suspect و Fail به‌ترتیب به داده‌های مشکوک و قبول‌نشدنی اشاره دارند.

۴ نتیجه‌گیری

سری زمانی ازن سطحی اندازه‌گیری‌شده، منبعی ارزشمند در مطالعه کنترل آلودگی هوا و گرمایش هواکره زمین به‌شمارمی‌رود. متأسفانه این داده‌ها در معرض انواع خطاهای ناشی از عوامل طبیعی و غیرطبیعی هستند که در مراحل مختلف چرخه داده اعم از اندازه‌گیری، جمع‌آوری، پردازش و ... به‌وجود می‌آیند. با توجه به اثرگذاری خطا بر نتایج، کنترل کیفی داده حائز اهمیت فراوان است و گامی ضروری در چرخه داده محسوب می‌شود. در این راستا ابزارها و روش‌های متعددی برای کنترل کیفی داده توسعه یافته است که هریک مزایا و کاربردهای خاص خود را دارد. در این مطالعه از یک ابزار جدید به نام AutoQA4Env استفاده شد که توانایی بررسی کنترل کیفی طیف وسیعی از داده‌ها را دارد.

ابزار AutoQA4Env جهت کنترل کیفی خودکار داده‌های محیطی بر اساس برنامه‌نویسی پیشرفته و با استفاده از فناوری‌های روز در مرکز تحقیقاتی یولیش کشور آلمان توسعه یافته است. این ابزار در دو حالت پایه و پیشرفته به‌ترتیب با سامانه‌های نشان‌گذاری دوحالته و سه‌حالته به‌صورت رایگان در دسترس است. در این پژوهش از این ابزار در بررسی کنترل کیفی داده‌های ازن سطحی اندازه‌گیری‌شده در ایستگاه‌های سنجش آلودگی

هوای شهر تهران استفاده شد. نتایج نشان داد به‌کارگیری این ابزار در حالت پایه می‌تواند به تشخیص تعداد زیادی داده قبول‌نشدنی منجر شود. اغلب داده‌های قبول‌نشدنی، خطای فاحش بودند که از اثر آنها بر نتایج نمی‌توان چشم‌پوشی کرد. در برخی از ایستگاه‌ها، اجرای ابزار در حالت پایه به تشخیص نادرست منجر شد که در آن بعضی از داده‌های قبول با نام داده قبول‌نشدنی نشان‌گذاری شده بودند. اجرای AutoQA4Env در حالت پیشرفته سبب تعدیل نتایج شد و تعداد زیادی از داده‌های قبول‌نشدنی در حالت پایه (۷/۵۷ درصد)، با نام داده مشکوک نشان‌گذاری شد؛ ازاین‌رو، به‌کارگیری حالت پیشرفته AutoQA4Env حتی با آزمایش آماری اندک در برخی از ایستگاه‌ها مناسب‌تر از حالت پایه است. شایان ذکر است کنترل کیفی داده‌ها به معنی تشخیص و حذف هرچه بیشتر خطای داده نیست، بلکه به معنی تشخیص صحیح خطا و تهیه گزارش کامل برای کاربر است. کاربر با توجه به نیاز و هدف مطالعه می‌تواند برای حذف داده‌های قبول‌نشدنی یا مشکوک اقدام کند. در برخی موارد، داده‌های مشکوک به پدیده‌های طبیعی نادر و کمیاب مانند موج گرمایی شدید در برخی از فصول یا نفوذ کم‌سابقه هوای پوشن‌سپهری به وردسپهر نسبت داده می‌شوند که می‌تواند هدف یک مطالعه باشد.

از مزایای به‌کارگیری AutoQA4Env، انعطاف‌پذیری، تکرارپذیری و توانایی نگهداری آن است. این ابزار را می‌توان گسترش داد؛ به این معنی که کاربر می‌تواند آزمایش‌های آماری دیگر و حتی پیشرفته‌تر را به ابزار اضافه کند. از این ابزار برای سری‌های زمانی دیگر آلاینده‌ها و میدان‌های هواشناختی نیز می‌توان استفاده کرد؛ لذا استفاده و توسعه هرچه بیشتر این ابزار جهت کنترل کیفی داده‌ها در زمینه‌های مختلف علمی توصیه می‌شود.

تشکر و قدردانی

نویسندگان از معاونت محترم علم و فناوری ریاست جمهوری برای حمایت مالی در انجام دادن این پژوهش، از شرکت کنترل کیفی هوای شهر تهران برای در اختیار قرار دادن داده و از مرکز تحقیقات یولیش برای در دسترس قرار دادن ابزار AutoQA4Env سپاسگزاری می‌کنند.

منابع

عابدینی، ع.، آزادی، م.، پرهیزگار، د.، ۱۳۸۲، کنترل کیفی داده‌های همدیدی سطح زمین و جو بالا: نشریه تحقیقات جغرافیایی، **۱۸**(۲)، ۷۴–۸۵

Campbell, J., 2013, Quantity is nothing without quality: BioScience, **63**(7), 574–585.

Gandin, L. S., 1988, Complex quality control of meteorological observations: Monthly Weather Review, **116**(5), 1137–1156.

Kaffashzadeh, N., and Schultz, M. G., 2020a, AutoQA4Env (basic flagging source code for EGU 2020 presentation), accessed 11 September 2021, https://b2share.fz-juelich.de/records/f79417f0a7eb4db7818e6e4e3c0163e7.

Kaffashzadeh, N., and Schultz, M. G., 2020b, AutoQA4Env (advanced flagging source code for EGU 2020 presentation), accessed 11 September 2021, https://b2share.fz-juelich.de/records/9afba748f2f943f5a73e6b6b919ce3c2.

Lorenc, A. C., 1981, A global three-dimentional multivaraite statistical interpolation scheme: Monthly Weather Review, **109**(4), 701–721.

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., ... , and Williams, M. L., 2015, Tropospheric ozone and its precursors from the urban to the global: Atmospheric Chemistry and Physics, **15**(15), 8889–8973.

Osborne, J. W., and Overbay, A., 2004, The power of outliers (and why researchers should always check for them): Practical Assessment, Research, and Evaluation, **9**(6), accessed 11 September 2021, https://scholarworks.umass.edu/pare/vol9/iss1/6.

Schultz, M. G. et al., 2017, Tropospheric ozone assessment report: database and metrics data

U.S. Integrated Ocean Observing System, 2014, Manual for real time quality control of in-situ temperature and salinity data: a guide to quality control and quality assurance for in-situ temperature and salinity observations: Silver Spring, MD, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, accessed 11 September 2021, https://repository.oceanbestpractices.org/handle/11329/269.

Zahumensky, I., 2016, Guidelines on quality control procedures for data from automatic weather stations, accessed 11 September 2021, https://www.researchgate.net/publication/228826920.

Zurbenko, I., Porter, P., Rao, S. T., Ku, J. Y., Gui, R., and Eskridge, R. E., 1996, Detecting discontinuities in time series of upper-air data: Development and demonstration of an adaptive filter technique: Journal of Climate, **9**, 3548–3560.

of global surface ozone observations: Elementa: Science of the Anthropocene, **5**:**58**.

Scully-Allison, C., Le, V., Fritzinger, E., Strachan, S., Harris, F. C., and Dascalu, S. M., 2018, Near real-time autonomous quality control for streaming environmental sensor data: Procedia Computer Science, **126**, 1656–1665.

Sofen, E. D., Bowdalo, D., Evans, M. J., Apadula, F., Bonasoni, P., Cuperio, M., and Ellul, R., 2016, Gridded global surface ozone metrics for atmosphere: Earth System Science Data, **8**, 41–59.

Steinacker, R., Mayer, D., and Steiner, A., 2011, Data quality control based on self-consistency: Monthly Weather Review, **139**(12), 3974–3991.

Tanhua, T., van Heuven, S., Key, R. M., Velo, A., Olsen, A., and Schirnick, C., 2010, Quality control procedures and methods of the CARINA database: Earth System Science Data, **2**, 35–49.

Quality control of measured surface Ozone at the Tehran city stations using a new sStatistical software

Najme Kaffashzadeh [1*] and Abbas Ali Aliakbari-Bidokhti [2]

[1] *Postdoctoral Fellowship, Space Physics Department, Institute of Geophysics, University of Tehran, Tehran, Iran*

[2] *Professor, Space Physics Department, Institute of Geophysics, University of Tehran, Tehran, Iran*

(Received: 19 September 2021, Accepted: 16 December 2021)

Summary

Being an inseparable part of environmental data, errors are generated due to several reasons, either natural or artificial. The first is produced from natural phenomena such as animal activities, storms, floods, etc. The later can be generated via human activities during data collecting, entering and processing that can be intentional or unintentional. Since errors can affect results of any analysis, distinguishing them via quality control is a prerequisite of any data usage. Because of unknown truth, this seemingly simple task becomes challenging. Although many efforts have been devoted to develop tests and tools for distinguishing errors in data, none of them can guarantee that all errors can be found. It is important as much as orthogonal testing to find more errors. Here we used a tool named AutoQA4Env, which has been developed for an automated quality control of environmental data. This tool consists of a series of statistical tests which have been used in various communities and organizations such as World Meteorological Organization and Environmental Protection Agency. The tests have been classified in several groups, based on their strictness. The tool has a setting menu by which users can add tests and modify the thresholds. Two versions of the tool, namely basic and advanced flagging system are open source and accessible via b2share. The tool was tested for the quality control of a set of data series of surface ozone measured at the pollution monitoring stations in the city of Tehran. These data are an important source to get information about the pollution levels and trends in Tehran; thus knowing their quality can improve and reduce the uncertainties in the results. The results indicate that gross errors exist in the most of the stations' data, even though these data are published and are publicly available. Applying the tool in the basic state finds most of the errors. About 0.02% of the data were erroneous for three years of data at 15 stations. Binary flagging system of the tool labels these failure data as an unacceptable data, although they were in fact acceptable. The advanced state of the tool was more moderate than the basic one and corrected these labels. In this state, 57.7% of the unacceptable data in the basic state were distinguished as a suspected value and only 5.6% of them were unacceptable. Therefore, we can conclude that the AutoQA4Env even at this stage could find and flag most of the data errors, at least gross errors. Besides, the advanced flagging system of the tool reduces errors in labeling.

Keywords: Errors, quality control, AutoQA4Env tool, surface ozone data

[*]Corresponding author: n.kaffashzadeh@ut.ac.ir