Journal of Environmental Studies
Vol. 48, No. 4, Winter 2023

Journal Homepage: www.Jes.ut.ac.ir
Print ISSN: 1025-8620 Online ISSN 2345-6922

Estimation of Missing Values in Time Series of Air Pollution Data in Tehran City

Moslem Dehnavi Eelagh, Rahim Ali Abbaspour*

Department of Spatial Information Systems, Faculty of Surveying Engineering and Geospatial Information, College of engineering, University of Tehran, Tehran, Iran

Document Type
Research Paper

Received
March 03, 2022

Accepted
December 11, 2022

DOI: 10.22059/JES.2022.339422.1008287

Abstract

Today, air pollution has become one of the most critical problems in densely populated cities, which causes many city residents to suffer from lung problems every year and can have irreparable effects on citizens' health. Air pollution recording devices in cities record pollution hourly. The technical issues of these devices sometimes cause some of the important data not to be recorded, and as a result, fixed values are created in the data. In this study, fixed values have been estimated. For this purpose, the study of air pollution events in Tehran including the concentration of PM2.5, PM10, SO2, NO2, O3 and CO pollutants was conducted. The LANN algorithm, used in the estimation and forecasting of single-variable time series, has been implemented and compared for all pollutants. Also, in another part of the study, other environmental pollutants have been considered in the estimation of fixed values, and by using the neural network method, the estimation of fixed values for all pollutants has been done. RMSE index was also used to check and compare algorithms. The value of RMSE in the LANN method was lower than other simpler models including mean, linear regression and LOCF, so its value was 30 to 50% lower, depending on the type of pollutant. Also, the neural network algorithm had lower RMSE than other methods in estimating PM2.5 values and its value was 7.78.

Keywords: Air pollution; Genetic programming; Missing data; Prediction

* Corresponding Author: Email: abbaspour@ut.ac.ir

Extended Abstract

Introduction

Air pollution has become one of the most important problems in large cities, where a significant number of citizens face various physical problems every year. Moreover, it can have irreversible effects on the health of citizens. Dealing with this problem requires complete and accurate information to be able to make suitable decisions based on that information. In cities facing this problem, devices that include pollution-sensitive sensors are placed with appropriate distribution throughout the city that continuously monitor air pollution. Air pollution recorders record information such as PM2.5, PM10, SO2, NO2, CO, O3, and other meteorological information. Factors such as power outages, systematic errors, and interruptions in maintenance repairs can cause missing data.

Air pollution data are mainly presented in the form of tables consisting of rows and columns. The lines that show the time dimension specify the recording time of each air pollution record, which are available hourly and daily. Also, each record contains a number of columns, which is related to a specific pollutant. Therefore, each position in the table has a numerical value that shows the level of a specific pollutant at a specific time.

In this study, the LANN algorithm, which is used in estimating and predicting univariate time series, is used and the missing values are implemented and compared for all pollutants. Also in another part of the study, other environmental pollutants have been considered in the estimation of missing data, it has been estimated by using the genetic programming method.

Materials and methods

Study area

Daily pollution data from Tehran are used in this study from April 2018 to April 2021. The city of Tehran has 23 active air pollution registration stations that are spread throughout the city. In this study, have worked on the events of Tarbiat Modares station. The data are for quarterly information and include 1081 records each record in addition to the number of days of the year, has 6 information columns that show the concentration of different particles with different sizes. These particles are PM2.5, PM10, SO2, NO2, O3, and CO.

LANN Algorithm

The LANN algorithm is based on statistical methods that find the most similar neighborhood for a series of searches. The searched neighborhood has the most similarity and proximity to the missing data and can be completed using statistical operators such as the average. Different neighborhoods can be considered for this method. For example, if neighborhood 3 is considered, the three neighbors before and after the vacancy are used to calculate the mean.

In this study, pollution data for three consecutive years have been examined. One solution is to use historical data. In this research, according to the history of the data, this case has been investigated. In this way, a threshold is considered for the difference between the neighborhood data, and if the difference is more than the threshold, the history data is also used for the neighborhood data.

Genetic programming (GP)

The GP algorithm is one of the population-based methods that fall into the category of meta-heuristic methods. This method can be considered an extension of the genetic algorithm (GA) method. The GA method has several regulatory parameters, including selection, crossover, and mutation. The GP method goes beyond this method and deals with operators and variables. Operators such as subtraction, addition, multiplication, division, and more. Each response in this method is called a program that is created to meet an objective. This method has better results than other methods used to replace multivariate. Each program is a set of operators and variables that represent a relation. Each program can be represented in a tree structure where nodes, operators, and leaves play the role of variables and fixed numbers.

In the GP method, as in the GA method, populations of responses are generated and evaluated for optimization. The stages of population production are selection, crossover, and mutation. In the selection stage, two programs are selected from the previous population, which uses methods such as tournaments. In the crossover phase, the two selected programs share their nodes. After the crossover process, the mutation process takes place, which affects a percentage of the population according to the rate set for it. Randomly replace sheets or subtrees of a program with random values.

Other algorithms

In this study, in addition to the two algorithms explained in the previous sections, Three algorithms including average algorithms, LOCF and linear regression were used to estimate empty places. In the algorithm, the average of empty positions is completed with the average number of the entire time series, which is considered a fast method in terms of implementation speed. This algorithm takes into account the general changes, so in the time series when the changes of the records are insignificant and the standard deviation of the data is small, the use of this algorithm can create a suitable approximation. Because the smaller the standard deviation, the data is closer to the average and fluctuates at a smaller distance from the average. In the LOCF algorithm, each empty position is filled using the data before it. This algorithm is suitable for completing series that have an ascending or descending pattern with a low slope, because in these patterns, positions close to each other have similar values.

Discussion

First, several data were randomly deleted on the data using an algorithm, then the missing data were estimated by applying the stated methods, and then the RMSE value was calculated based on the estimated values and actual values. This process was performed for different amounts of missing data. The problem was also solved 30 times each time. In this way, in the time series of the data, the missing data were created 30 times randomly and then these missing data were filled. In addition to the methods introduced in this study, simple mean, linear regression, and LOCF methods have been evaluated. This process has been applied to all PM2.5, PM10, SO2, O3, NO2, and CO pollutants.

The LANN method introduced in this study had a lower RMSE than all other methods in all pollutants, which means that their estimates were closer to reality. The simple mean method had the highest RMSE in all pollutants except O3. Also in the O3 pollutant, the highest amount of RMSE was related to the linear regression method. The LOCF method had almost better values in some pollutants such as O3 and PM2.5 than other methods, as the RMSE value of this method was slightly higher than the LANN methods. The important point about RMSE values is that in some pollutants the changes fluctuate in a small range, so the amount of RMSE in these pollutants is less than in pollutants that have a larger fluctuation range. This can be seen in NO2 and PM10 pollutants.

Following the implementation of the methods for estimating the missing values, the GP method was applied to the relevant data. In this method, which is a multivariate method, other environmental pollutants have been used to estimate PM2.5. These pollutants are PM10, SO2, NO2, CO, and O3. The output of the GP method is an equation based on other pollutants that can be used to estimate missing data. For all pollutants, a relationship based on other pollutants was obtained, which was the output of the GP method.

The GP method had a lower RMSE value in PM2.5 pollutants compared to the LANN methods which had better results than other univariate methods. In PM10, SO2, and CO pollutants, the amount of RMSE for the GP method and LANN methods was almost the same, and in O3 and NO2 pollutants, the amount of RMSE for the GP method was higher than LANN methods. The advantage of the GP method over univariate methods is that the method is not dependent on the data before and after the missing positions. Missing data increases the RMSE due to the dependence of LANN methods on pre-and post-empty values, so multivariate methods may be a better choice when there are large gaps in the data.

Based on the results of the implementation of the GP algorithm, once considering the radical operator and once without considering it, the results indicated a decrease in the value of RMSE, so that the effect of adding the radical operator on PM2.5, PM10, SO2, NO2 and CO pollutants , the reduction of 10.5%, 5.9%, 6.2%, 8.9% and 3.8% in RMSE value was respectively. The radical operator has no effect on the amount of O3 pollutant and the extracted relationship is the same in both cases and the radical operator has no role in forming the relationship.

In the estimation of PM2.5 pollutant, the GP algorithm had the best estimate and the RMSE value was 7.78. To estimate PM10 in this studytwo LANN and GP algorithms have worked properly, so that the RMSE value obtained by using two LANN and GP algorithms is 17.08 and 18.73 respectively. SO2 index has smaller fluctuations in recorded values compared to other pollutants, smaller RMSE values have been obtained. LANN and GP algorithms with RMSE values of 1.90 and 1.92 respectively have better results than LOCF, average and linear regression methods.

Also, pollutant O3 has the best results compared to other methods implemented in this study by using two algorithms, LANN and LOCF, with RMSE values of 4.35 and 5.40, respectively. The two mentioned algorithms also had the best results in the estimation of NO2, so that the LANN algorithm with an RMSE value of 8.44 and the LOCF algorithm with an RMSE value of 10.52 have estimated the NO2 index. In this study, to estimate CO, the LANN algorithm with RMSE value of 0.60 has been more efficient compared to other methods.
The main goal of this study was to investigate the way of placing missing values in time series, and the presented algorithms can be used as a tool for estimating missing values. Relevant organizations such as crisis management can complete the incomplete data recorded by air pollution recording stations using these algorithms, which can lead to a decision based on correct and complete information.

Conclusion

In this study, univariate algorithms and GP algorithm are used as multivariable algorithms. Of course, algorithms based on artificial intelligence can be used to estimate missing values. Also, this study has estimated each pollutant based on other pollutant. By adding meteorological parameters such as temperature, humidity and wind as well as information related to the land such as height, slope and aspect to the problem, the relationship of these parameters with the amount of air pollution can be investigated. Also, considering the effect of location on the results obtained from various methods, the methods presented in this study and previous studies can be investigated on the same area. The basic operators used in the GP algorithm, which was discussed in this study, included addition, division, multiplication and square root. It seems that the use of other operators in accordance with the relevant pollutant trend can provide more accurate estimates. It is suggested to use operators and functions that can show periodicity well for pollutants with alternating trends. Also, it seems that converting the data set into subsets that have a constant trend It can cause the efficiency of general algorithms such as regression.

محیط‌شناسی، دورهٔ ٤٨، شمارهٔ ٤، زمستان ١٤٠١، صفحهٔ ٤٣٩ – ٤٥٩

تخمین مقادیر جا افتاده در سری‌های زمانی داده‌های آلودگی هوای شهر تهران

مسلم دهنوی ئیلاق، رحیم علی‌عباسپور*

گروه سیستم‌های اطلاعات مکانی، دانشکده مهندسی نقشه‌برداری و سیستم‌های اطلاعات مکانی، دانشکدگان فنی، دانشگاه تهران، تهران، ایران.

تاریخ وصول مقاله: ١٤٠٠/١٢/١٧ تاریخ پذیرش مقاله: ١٤٠١/٠٩/٠٢

چکیده

امروزه آلودگی هوا به یکی از معضلات مهم در شهرهای پرجمعیت تبدیل‌شده است که هرساله تعداد قابل‌توجهی از ساکنان شهرها را با مشکلات ریوی روبرو می‌کند و می‌تواند تأثیرات جبران‌ناپذیری بر سلامت شهروندان داشته باشد. دستگاه‌های ثبت آلودگی هوا در شهرها، آلودگی را به‌صورت ساعتی ثبت می‌کنند. مشکلات فنی پیش‌آمده برای این دستگاه‌ها، در بعضی مواقع سبب می‌شود بخشی از داده‌های مهم ثبت نگردند و درنتیجه آن، مقادیر جاافتاده در داده‌ها ایجاد می‌گردد. در این مطالعه به تخمین مقادیر جاافتاده پرداخته‌شده است. این مطالعه روی‌داده‌های آلودگی هوای شهر تهران شامل غلظت آلاینده‌های PM2.5، PM10، SO2، NO2، O3 و CO انجام‌شده است. در این مطالعه الگوریتم LANN که در تخمین و پیش‌بینی سری‌های زمانی تک متغیره کاربرد دارد، استفاده و مقادیرجاافتاده برای تمامی آلاینده‌ها پیاده‌سازی و مقایسه شده است. هم‌چنین در بخشی دیگر از مطالعه، سایر آلاینده‌های محیطی در برآورد مقادیر جاافتاده در نظر گرفته شده‌اند که با به‌کارگیری روش شبکه عصبی، تخمین مقادیر جاافتاده برای همه آلاینده‌ها انجام شده است. همچنین برای بررسی و مقایسه الگوریتم ها از شاخص RMSE استفاده شده است. مقدار RMSE در روش LANN نسبت به سایر مدل‌های ساده‌تر شامل میانگین، رگرسیون خطی و LOCF مقدار کمتری داشت به نحوی که مقدار آن ٣٠ تا ٥٠ درصد، بسته به نوع آلاینده کمتر بوده است. همچنین الگوریتم شبکه عصبی نسبت به سایر روش‌ها در تخمین مقادیر PM2.5، RMSE کمتری داشت و مقدار آن ٧/٧٨ بوده است.

کلید واژه‌ها: آلودگی هوا، برنامه‌ریزی ژنتیک، داده جاافتاده ، پیش‌بینی، داده کاوی

سرآغاز

امروزه آلودگی هوا به یکی از معضلات مهم در شهرهای بزرگ تبدیل‌شده است که هرساله تعداد قابل‌توجهی از ساکنان شهرها را با مشکلات جسمی مختلف روبرو می‌کند و می‌تواند تأثیرات جبران‌ناپذیری بر سلامت شهروندان داشته باشد؛ مقابله با این معضل نیازمند داده‌ها و اطلاعات کامل و دقیق می‌باشد تا بتوان بر مبنای آن اطلاعات تصمیم‌گیری نمود. در شهرهایی که با این پدیده روبرو هستند دستگاه‌هایی که شامل سنسورهای حساس به آلودگی هوا هستند با توزیعی مناسب در سراسر شهر قرار داده می‌شود که به پایش مستمر آلودگی هوا می‌پردازند. این دستگاه‌های ثبت آلودگی هوا به‌صورت ساعتی داده‌های

* نویسنده مسئول
Email: abbaspour@ut.ac.ir

DOI: 10.22059/JES.2022.339422.1008287
DOR: 20.1001.1.10258620.1401.48.4.1.0

مربوط به آلودگی هوا را ثبت می‌کنند. دستگاه‌های ثبت آلودگی هوا اطلاعاتی نظیر غلظت آلاینده‌های PM2.5، PM10، SO2، NO2، CO، O3 و سایر اطلاعات هواشناسی را ثبت می‌نمایند که بروز عواملی نظیر قطع برق، خطاهای سیستماتیک و وقفه‌های ایجاد شده جهت تعمیرونگهداری، می‌تواند سبب ناپیوستگی در ثبت داده‌ها گردد(Liu et al., 2020). مقادیر جاافتاده سبب ایجاد خلل در فرآیند تصمیم‌گیری و انجام امور پژوهشی می‌گردد(Ghazali et al., 2020). به همین دلیل تکمیل این مقادیر جاافتاده با مقادیر مناسب، بسیار اهمیت دارد و منجر به تولید یک مجموعه داده کامل می‌گردد. به دلیل نامشخص بودن منابع آلودگی هوا و تأثیرگذاری عوامل فیزیکی و شیمیایی گوناگون در ایجاد آن، الگوهای متفاوتی برای آلودگی هوا می‌توان در نظر گرفت. بنابراین نحوه تولید و ارزیابی این الگوها اهمیت دارد (Seinfeld and Pandis, 2016). بخشهای خالی بر اساس تعداد مقادیر جاافتاده متوالی در یک مجموعه داده به سه دسته کوچک، متوسط و بزرگ تقسیم‌بندی می‌شوند که روش‌های مورد استفاده برای تکمیل این جایگاه‌های خالی باید بر اساس نوع آن‌ها تعیین گردد(Tito et al., 2019).

داده‌های آلودگی هوا عمدتا به صورت جداولی متشکل از سطر و ستون ارائه می‌شوند. سطرهای که بعد زمانی را نشان می‌دهند، زمان ثبت هر رکورد آلودگی هوا را مشخص می‌کند که به صورت ساعتی و روزانه دردسترس می‌باشند. همچنین هر رکورد ثبت‌شده شامل تعدادی ستون می‌باشد که هرکدام مربوط به یک آلاینده مشخص می‌باشد. بنابراین هر جایگاه در جدول دارای یک مقدار عددی می‌باشد که میزان شاخص یک آلاینده خاص در یک زمان مشخص را نشان می‌دهد.

برای حل این مشکل روش‌های متنوعی، از روش‌های ساده و سریع و درعین‌حال با دقت کمتر گرفته تا روش‌های پیچیده‌تر و دارای دقت بالاتر، توسط پژوهشگران و محققان علوم داده ارائه‌شده است. بعضی روش‌ها بر اساس اطلاعات گذشته، عملیات جایگذاری را انجام می‌دهند، ولی مشکل این روش‌ها این است که تغییرات ناشی از زمان را در نظر نمی‌گیرند (Junninen et al., 2004; Noor et al., 2015). برای مثال ممکن است احداث یک کارخانه در نزدیکی یک ایستگاه ثبت آلودگی هوا، سبب تغییرات کلی در مقادیر ثبت‌شده گردد که در این شرایط استفاده از داده‌های گذشته، راهکار مناسبی نیست. روش‌های ساده دیگری مانند میانگین و میانه داده‌ها می‌توانند به‌عنوان جایگزین در جایگاه‌های خالی قرار گیرند که این روش‌ها متناسب با نوع، ماهیت و ویژگی‌های آماری داده‌ها مورداستفاده قرار می‌گیرند. به‌عنوان‌مثال روش میانه در داده‌های اریب، کاربرد بیشتری دارد و نتایج بهتری نسبت به روش میانگین دارد(Junger and De Leon, 2015). همچنین برخی از روش‌ها بر اساس سایر اطلاعات محیطی نظیر جهت و سرعت باد و دما، مقادیر آلایندگی را تخمین زده‌اند. این روش‌ها براساس همبستگی موجود بین پارامترهای هواشناسی و شاخص‌های آلودگی هوا شکل گرفته‌اند.

به‌صورت کلی در مواجهه با مشکل مقادیر جاافتاده، دو رویکرد وجود دارد. یک رویکرد که تک متغیره نام دارد، بر اساس اطلاعات موجود در همان ستون، فرآیند جایگذاری صورت می‌گیرد. روش‌های درون‌یابی و آخرین مشاهده درحرکت روبه‌جلو (LOCF[1]) از روش‌هایی هستند که در رویکرد تک متغیره استفاده شده‌اند(Engels and Diehr, 2003; Plaia and Bondi, 2006). در رویکرد دیگر که حالت چندمتغیره می‌باشد، بر اساس سایر ستون‌ها و مقایسه ماتریس مشابهت بین ستون‌ها، رابطه و الگویی از بین داده‌ها استخراج می‌گردد و بر اساس الگوهای موجود، تخمین مدل پیش‌بینی صورت می‌گیرد و مقادیر جاافتاده ، برآورد می‌گردند.

روش LOCF آخرین مقدار ثبت‌شده را برای مقادیر جاافتاده در نظر می‌گیرد (Bokde et al., 2018; Zeileis and Grothendieck, 2005). این روش با توجه به اینکه از یک مقدار ثابت برای مقادیر جاافتاده استفاده می‌نماید، برای جایگاه‌های خالی با اندازه‌های کوچک مناسب‌تر است. روش عرشه داغ مقادیر جاافتاده را به‌صورت تصادفی از بین مقادیر

ثبت شده انتخاب می‌کند (Aljuaid and Sasi, 2016; Kowarik and Templ, 2016) . این روش با استفاده از کتابخانه‌های موجود در نرم‌افزار R قابل پیاده‌سازی است. روش میانگین متحرک (MA) جایگاه‌های مختلف را با میانگین مقادیر پر می‌نماید که برای محاسبه مقدار میانگین فرمول‌های مختلفی نظیر میانگین ساده (SMA)، میانگین وزن‌دارخطی (LWMA)، میانگین متحرک وزن‌دار توانی (EWMA) را می‌توان نام برد که در نرم‌افزار R قابل پیاده‌سازی هستند.

در خصوص موضوع مورد مطالعه، Caillault و همکاران (۲۰۲۰) روشی بر مبنای بسته‌بندی زمان پویا ارائه کرده‌اند. الگوریتم پیشنهادی بر اساس چهار شاخص Similarity، NMEA، RMSE و FSD با سایر الگوریتم‌ها مقایسه شده است و نتایج بهتری داشته است. Flores و همکاران (۲۰۱۹) از الگوریتمی سه مرحله‌ای برای تکمیل مقادیر جاافتاده از سری‌های زمانی استفاده کرده‌اند. با استفاده از روش میانگین بردار تاریخی به‌صورت فصلی مقادیر جاافتاده برآورد شده است. سپس از روش درون‌یابی نزدیک‌ترین همسایگی مقادیر قبل و بعد از مقدار جاافتاده با مقادیر برآورد شده تنظیم شده‌اند و درنهایت از فیلتر میانگین محلی نزدیک‌ترین همسایگی (LANNf [۲]) برای هموارسازی منحنی به‌دست آمده، از روش نزدیک‌ترین همسایگی اقدام شده است.

همچنین Flores و همکاران (۲۰۱۹) تخمین مقادیر جاافتاده را با به‌کارگیری الگوریتم LANN[۳] پیاده‌سازی کرده‌اند. برای داده‌های هواشناسی مثل بیشینه دما، استفاده از این دو الگوریتم پیشنهاد شده است. این روش‌ها نسبت به سایر روش‌های مرسوم مانند تلفیق دو الگوریتم فیلتر کالمن و میانگین متحرک یکپارچه خودرگرسیون ، عرشه‌داغ، LOCF و نزدیکترین همسایگی K نتایج بهتری داشته‌اند. Andiojaya و Demirhan (۲۰۱۹) بهبودی برروی روش‌های فیلتر کالمن و میانگین وزن‌دار انجام داده‌اند. درنهایت روش Kalman filtering همراه با روش ARIMA برای تخمین سری‌های زمانی پیشنهاد شده است. Shaadan و Rahim (۲۰۱۹) از الگوریتم‌های مختلفی ازجمله درون‌یابی خطی، درون‌یابی اسپلاین، میانگین متحرک توانی، نمونه تصادفی، فیلتر کالمن با به‌کارگیری مدل ARIMA و میانگین برای تخمین مقادیر جاافتاده داده‌های آلودگی هوا ازجمله PM10 استفاده کرده‌اند که درنهایت روش فیلتر کالمن با به‌کارگیری مدل ARIMA در مقایسه با سایر روش‌ها از نتایج بهتری برخوردار بوده است. Hamami و همکاران (۲۰۲۰) مدل حافظه کوتاه‌مدت طولانی (LSTM) را برای ایجاد مدل شبکه عصبی جهت پیش‌بینی آلاینده‌های موجود در هوا با خطای کمتر ارائه کرده‌اند. مدل‌های پیش‌بینی می‌توانند برای جایگذاری مقادیر جاافتاده استفاده گردند.

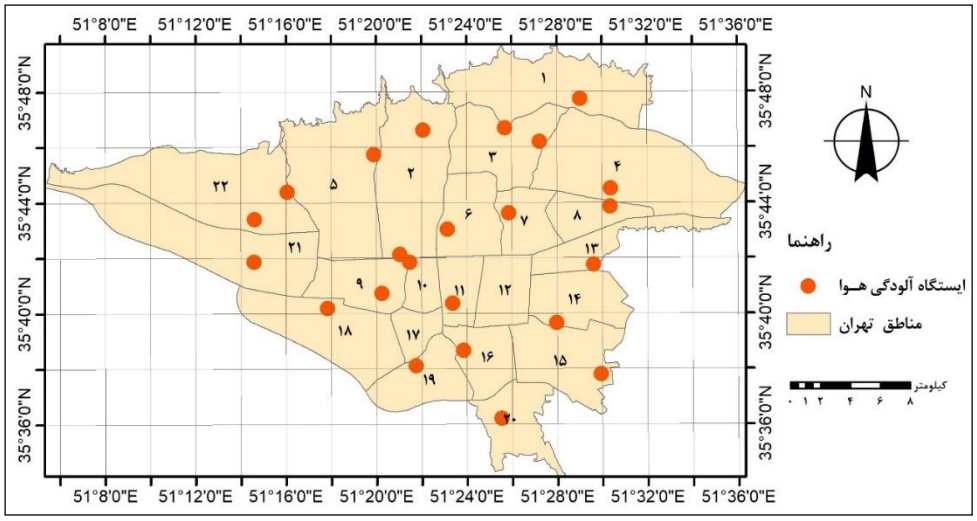
همچنین Hadeed و همکاران (۲۰۲۰) به مطالعه تخمین مقادیر PM2.5 با استفاده از روش‌های تک‌متغیره و چندمتغیره پرداخته‌اند. میانگین، میانه، LOCF، Kalman Filter، تخمین تصادفی و زنجیره مارکوف به عنوان روش‌های تک‌متغیره پیاده‌سازی شده‌اند. همچنین دو الگوریتم تطبیق میانگین پیش‌بینی کننده و میانگین ردیف به عنوان روش‌های چندمتغیره پیاده‌سازی شده اند. دراین مطالعه از داده‌های میانگین ۲۴ ساعته PM2.5 استفاده شده است. همچنین یک مجموعه داده کوچک ۲۰ تایی برای بررسی مدل‌ها استفاده شده است. Mishchunk و همکاران (۲۰۱۹) برای تکمیل جایگاه‌های جاافتاده، مدل تبدیلات هندسی متوالی (SGTM) را پیشنهاد داده‌اند. دقت مدل ارائه شده در مقایسه با روش میانگین حسابی سه برابر دقت بالاتری داشته است. دراین مطالعه آلاینده‌های CO، NO2 و NO مورد بررسی قرار گرفته است. Yuan و همکاران (۲۰۱۸) مدل LSTM را برای تکمیل جایگاه‌های خالی برای آلاینده PM2.5 مورد استفاده قرار داده‌اند. از اطلاعات سایر آلاینده‌ها برای تخمین PM2.5 استفاده شده است. دراین مطالعه رکوردهای ساعتی آلاینده‌ها در نظر گرفته شده است و ۸۰ درصد داده‌ها آموزشی و ۲۰ درصد داده‌ها برای ارزیابی در نظر گرفته شده است. Shahbazi و همکاران (۲۰۱۸) از سیستم استنتاج فازی مبتنی بر شبکه تطبیقی(ANFIS) برای

تکمیل مجموعه داده‌های آلاینده‌های O3، NO، PM2.5 و PM10 استفاده کرده‌اند. داده‌های مورد استفاده در این مطالعه مربوطه به شهر تهران و به صورت ساعتی در نظر گرفته شده‌اند.

همانطور که بیان شد برخی روش‌ها براساس همبستگی موجود بین پارامترهای هواشناسی و شاخص‌های آلودگی هوا شکل گرفته‌اند. مطالعاتی که در این خصوص انجام گرفته است با بررسی تغییرات پارامترهای هواشناسی و شاخص‌های آلایندگی، میزان مشابهت بین روند تغییرات را به عنوان معیاری برای ارتباط بین پارامترها در نظر می‌گیرند. Ashrafi و Ahmadi Orkomi (۲۰۱۴) طی پژوهشی همبستگی را بین غلظت آلاینده‌های CO و NO2 با پارامترهای هواشناسی مانند دما، سرعت و جهت باد و عدد بی بُعد ریچاردسون بررسی کردند که به این منظور از روش رگرسیون چندمتغیره ، اطلاعات هواشناسی و آلودگی هوای شهر تهران استفاده شده است. که در نتیجه آن ارتباط معنی‌داری بین پارامترهای هواشناسی و دوآلاینده CO و NO2 کشف گردید. همچنین Yicun و همکاران (۲۰۲۰) در پژوهشی به بررسی ارتباط و همبستگی بین پارامترهای جوی و اطلاعات سینوپتیکی با غلظت آلاینده PM2.5 پرداخته‌اند که از اطلاعات شهر تبریز استفاده شده است که در آن از مدل لاگرانژی پخش ذرات HYSPLIT برای بررسی توده‌های هوایی شهر تبریز در شرایط آلودگی هوا استفاده شده است. در مطالعه‌ای دیگر Ashrafi و Hoshyaripour (۲۰۱۰) به بررسی در خصوص میزان پایداری شرایط جوی و ارتباط آن با غلظت آلاینده CO پرداخته‌اند که به صورت ساعتی و فصلی این ارتباط بررسی شده است. به این منظور از روش پاسکیل- ترنر استفاده شده است. استفاده از این قبیل اطلاعات و همبستگی‌ها می‌تواند در جهت بهبود الگوریتم‌های تخمین و افزایش دقت برآورد استفاده شوند.

مواد و روش بررسی

شهر تهران به سبب جمعیت بالا و همچنین صنایع مهمی که در اطراف آن قرار دارد، همه‌روزه با مشکلات آلودگی هوا مواجه است که عمده منشأ این آلودگی‌ها ناشی از وسایل نقلیه و کارخانه‌های صنعتی قرارگرفته در غرب آن می‌باشد که به دلیل وزش بادهای غربی – شرقی که در تهران می‌وزد، این آلودگی‌ها به سمت مناطق شهری تهران رانده می‌شوند. شهر تهران در فصول سرد سال با افزایش میزان آلاینده‌های PM2.5 روبرو است. این ذرات به دلیل ابعاد بسیار کوچکی که دارند (کمتر از ۲/۵ میکرون) برای سلامت انسان بسیار خطرناک هستند. داده‌های استفاده شده در این مطالعه، داده‌های روزانه آلودگی شهر تهران در بازه زمانی فروردین ۹۷ تا فروردین ۱۴۰۰ می‌باشد و داده‌های تعداد زیادی از ایستگاه‌های تهران در دسترس می‌باشد. در شکل ۱ می‌توان پراکندگی ایستگاه‌های ثبت آلودگی هوا شهر تهران را مشاهده نمود.



شکل ۱ - توزیع مکانی ایستگاه‌های ثبت آلودگی هوا

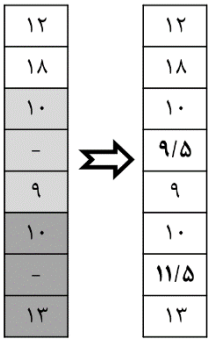
همان‌طور که در شکل ۱ مشخص می‌باشد، شهر تهران دارای ۲۳ ایستگاه فعال ثبت آلاینده‌های هوا می‌باشد که در سراسر شهر گسترده شده‌اند. در این مطالعه بر روی‌داده‌های ایستگاه تربیت مدرس تمرکز شده است. همچنین در این مطالعه از داده‌های آلودگی هوای شهرتهران برای سه سال متوالی استفاده شده است که شامل ۱۰۸۱ رکورد هستند که هر رکورد علاوه بر شماره روز سال، دارای ۶ ستون اطلاعاتی می‌باشد که این اطلاعات میزان غلظت آلاینده‌های مختلف با اندازه‌های مختلف را نشان می‌دهد. این آلاینده‌ها عبارت‌اند از:PM2.5، PM10، SO2، NO2، O3، CO. مطابق پیشینه مطرح‌شده، برای تکمیل مقادیر جاافتاده روش‌های مختلفی ارائه‌شده است الگوریتم LANN به‌عنوان الگوریتمی مؤثر نسبت به سایر روش‌ها معرفی‌شده است(Tito et al., 2019). در این مطالعه با رویکرد الگوریتم فوق و اعمال تغییراتی در نحوه پیاده‌سازی جهت بهبود نتایج آن، اقدام شده است. همچنین الگوریتم برنامه‌ریزی ژنتیک(GP[۴]) به‌عنوان یک روش چندمتغیره معرفی و پیاده‌سازی شده است. الگوریتم های میانگین، LOCF و رگرسیون خطی نیز به عنوان روش هایی که در مطالعات پیشینه ارائه شده بود، نیز در این مطالعه مورد استفاده قرار گرفته‌اند. در ادامه به این روش‌ها که روش‌هایی ساده محسوب می‌گردند، اشاره شده است.

الگوریتم LANN

مطابق پیشینه مطرح شده، برای تکمیل مقادیر جاافتاده روش‌های مختلفی ارائه‌شده است که الگوریتم LANN به‌عنوان الگوریتمی مؤثر نسبت به سایر روش‌ها معرفی‌شده است (Tito et al., 2019). الگوریتم LANN مبتنی بر روش‌های آماری می‌باشد که شبیه‌ترین همسایگی را برای یک سری زمانی جستجو می‌نماید. منطق اصلی این الگوریتم بر این مبنا می‌باشد که رکوردهایی که ثبت می‌شوند بیشترین شباهت را به رکوردهای قبلی و بعدی خود دارند. بنابراین طبق این الگوریتم مقدار میانگین همسایگی در نظر گرفته می‌شود. این همسایگی می‌تواند مقادیر متفاوتی داشته باشد(n). همسایگی جستجو شده دارای بیشترین شباهت و نزدیکی به مقدار جاافتاده می‌باشد و می‌توان با استفاده از عملگرهای آماری مانند میانگین مقادیر جاافتاده را تکمیل نمود. برای این روش می‌توان همسایگی‌های متفاوتی در نظر گرفت. مقادیر تخمین طبق فرمول ۱ به دست می‌آید:

$$y = \frac{1}{2n} \sum_{t=i-n}^{i+n} x_t \quad t \neq i \qquad (1)$$

در رابطه فوق y نشان‌دهنده مقدار برآوردشده برای جایگذاری در مقدار جاافتاده می‌باشد. همچنین n همسایگی را نشان می‌دهد. مثلاً اگر $n=1$ آنگاه فقط داده‌های قبل و بعد جایگاه، در تخمین نقش دارند. در فرمول فوق t و i به ترتیب نشان‌دهنده شماره جایگاه‌های همسایگی و شماره مقدار جاافتاده می‌باشند. در شکل ۲ نمونه‌ای از پیاده‌سازی روش با فرض اینکه همسایگی با اندازه ۱، بیشترین شباهت را به داده‌های ثبت‌شده در سری زمانی داراست، حل گردیده است.



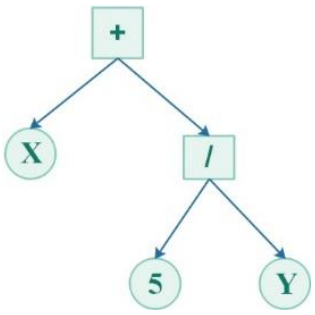
شکل ۲ – تکمیل مقادیر جاافتاده با میانگین داده‌های قبل و بعد از جایگاه خالی

محدوده‌های خاکستری نمایانگر همسایگی مقدار جاافتاده هستند. اعداد ۹/۵ و ۱۱/۵ حاصل میانگین اعداد قبل و بعد از جایگاه خالی می‌باشند. روش LANN برای داده‌های تک متغیره که موضوع این مطالعه می‌باشد کاربرد زیادی دارد.

روش برنامه‌ریزی ژنتیک

الگوریتم GP از جمله روش‌های مبتنی بر جمعیت می‌باشد که در دسته روش‌های فراابتکاری قرار می‌گیرد. این روش را می‌توان توسعه‌یافته روش الگوریتم ژنتیک دانست.

روش GA تعدادی پارامتر تنظیم‌کننده دارد شامل انتخاب، تلفیق و جهش. روش GP فراتر از این روش می‌باشد و با عملگرها و متغیرها سروکار دارد. عملگرها مثل تفریق، جمع، ضرب، تقسیم و موارد دیگر. هر پاسخ تولیدشده در روش برنامه‌ریزی ژنتیک، که در این مطالعه روابط استخراج‌شده از سری‌های زمانی می‌باشند، برنامه (Program) خوانده می‌شود که برای برآورده ساختن یک هدف (Objective)، ساخته می‌شود. در این مطالعه هدف مینیمم نمودن مجموع مربعات اختلاف بین مقادیر اصلی (داده‌های آموزشی) و مقادیر برآوردشده می‌باشد. این روش در مقایسه با سایر روش‌هایی که برای جایگزینی چند متغیره استفاده می‌شود، نتایج بهتری دارد(Tran, 2018). هر برنامه مجموعه‌ای از عملگرها و متغیرهاست که معرف یک رابطه هستند. شکل ۳ یک برنامه ساده را نشان می‌دهد که از عملگرهای ضرب و تقسیم و همچنین دو متغیر و یک عدد ثابت تشکیل شده است. هر برنامه را می‌توان در یک ساختار درختی نمایش داد که گره‌ها (Node)، نقش عملگرها و برگ‌ها نقش متغیرها و اعداد ثابت را ایفا می‌کنند.



شکل ۳ – ساختار یک برنامه

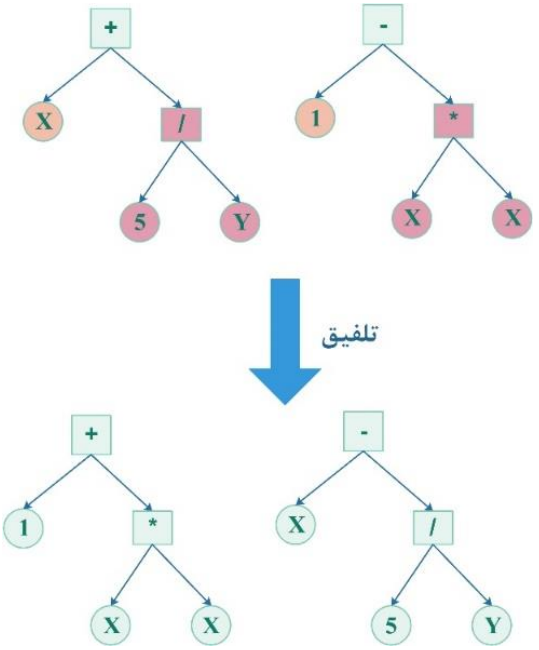
شکل ۳ یک ساختار درختی را برای یک برنامه ساده نشان داده است. این برنامه معرف رابطه « $X+\frac{5}{Y}$ » می‌باشد. در روش GP همانند روش GA، جمعیتی از جواب‌ها (برنامه‌ها) تولید شده و ازلحاظ بهینگی، مورد ارزیابی قرار می‌گیرند. برای تولید جمعیت اولیه، تعدادی برنامه براساس عملگرهای پایه ای که مشخص شده است، به صورت تصادفی تولید می‌گردد. برای ارزیابی هر برنامه از شاخص ریشه میانگین مربع انحرافات(RMSE) استفاده شده است که در رابطه ۲ مشخص شده است.

$$RMSE=\sqrt{\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n}} \qquad (۲)$$

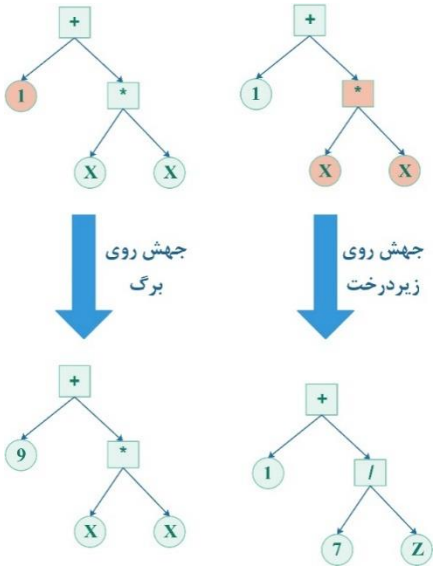
که در آن y_i مشاهده رکورد iام سری‌زمانی، $\hat{y}_i$ مقدار برآورد شده iام سری‌زمانی و n تعداد مقادیر جاافتاده می‌باشد. مراحل تولید جمعیت عبارت‌اند از: انتخاب، تلفیق و جهش. در مرحله انتخاب از بین جمعیت قبلی دو برنامه انتخاب می‌گردد که از روش‌هایی مثل رقابت استفاده می‌شود (Al-Helali et al., 2021; Tran et al., 2017). در روش رقابت، تعدادی برنامه به صورت تصادفی با شانس یکسان از بین جمعیت انتخاب می‌گردند و از بین برنامه‌های انتخاب شده، آن برنامه‌ای به عنوان والد انتخاب می‌گردد که دارای بهینگی بهتری باشد. این فرآیند برای تولید تعداد والد مورد نیاز به منظور تولید جمعیت جدید تکرار می‌گردد. در مرحله تلفیق دو برنامه انتخاب‌شده، گره‌های خود را به اشتراک می‌گذارند. مطابق با شکل ۴، دو برنامه نشان داده شده است. که روابط به صورت « X^2-1 »، درسمت راست تصویر و « $X+\frac{5}{Y}$ »، درسمت چپ تصویر نشان داده شده است. بعد از اعمال فرآیند تلفیق، گره‌های اصلی دو برنامه جابجا شده‌اند که در نتیجه آن روابط به صورت « X^2+1 » و « $X-\frac{5}{Y}$ » درآمده‌اند.

بعد از فرآیند تلفیق، فرآیند جهش اتفاق می‌افتد که با توجه به نرخی که برای آن تعیین شده است، درصدی از جمعیت تحت تأثیر قرار می‌گیرد. به این صورت که به‌صورت اتفاقی، برگ‌ها یا زیردرخت های یک برنامه با مقادیر اتفاقی، جایگزین می‌شود(شکل ۵). مطابق با شکل ۵ جهش می‌تواند به دو صورت انجام گیرد. در حالت جهش روی زیردرخت، یک گره و تمام مجموعه زیرگره‌های متصل به آن تغییر پیدا می‌کنند به عنوان مثال در شکل ۵ رابطه « X^2+1 » بعد از اعمال جهش روی زیردرخت، به رابطه

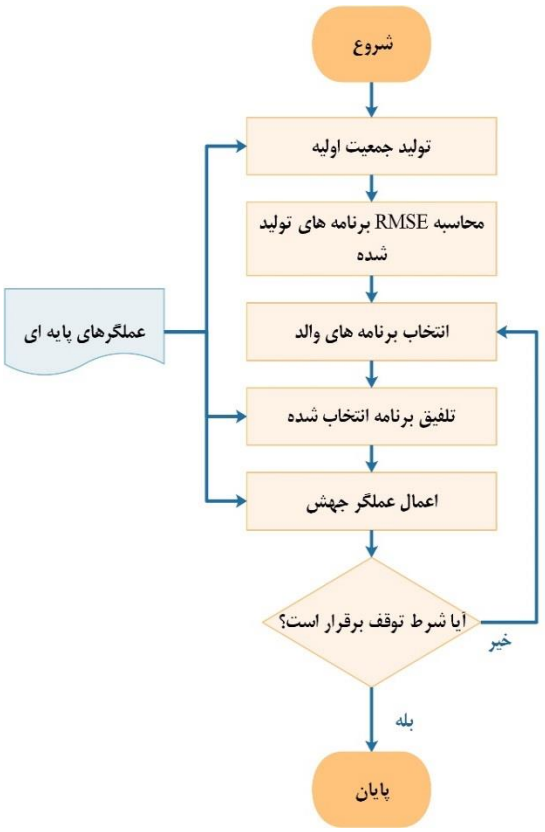
«$\frac{۷}{Z}+۱$» تبدیل شده است. در حالی‌که در حالت جهش روی برگ، فقط یک گره انتخاب می‌شود و مقداری تصادفی جایگزین آن می‌گردد. به عنوان مثال در شکل ۵ رابطه «$X^۲+۱$» بعد از اعمال جهش روی برگ، به رابطه «$X^۲+۹$» تبدیل شده است. همچنین در شکل ۶ فلوچارت الگوریتم پیشنهادی GP آمده است:



شکل ۴ – فرآیند تلفیق دو برنامه



شکل ۵ – فرآیند اعمال جهش بر روی برنامه‌ها



شکل ۶ – فلوچارت روش پیشنهادی GP

الگوریتم‌های ساده تر

در این مطالعه، علاوه‌بر دو الگوریتم توضیح داده شده در بخش‌های قبلی از سه الگوریتم میانگین، LOCF و رگرسیون خطی نیز برای تخمین مکان‌های خالی استفاده شده است. در الگوریتم میانگین جایگاه‌های خالی با عدد میانگین کل سری زمانی تکمیل می‌شود که روشی سریع از لحاظ سرعت پیاده‌سازی محسوب می‌گردد. این الگوریتم تغییرات کلی را در نظر می‌گیرد بنابراین در سری‌های زمانی که تغییرات رکوردها ناچیز و انحراف‌معیار داده‌ها کوچک می‌باشد، استفاده از این الگوریتم می‌تواند تقریب مناسبی ایجاد کند. زیرا هرچه انحراف‌معیار کوچکتر باشد، داده‌ها به میانگین نزدیک‌تر بوده و در فاصله کمتری از میانگین نوسان می‌کنند. در الگوریتم LOCF هر جایگاه خالی با استفاده از داده قبل از خود تکمیل می‌گردد. این الگوریتم مناسب تکمیل سری‌هایی است که دارای یک الگوی صعودی و یا نزولی با شیب کم هستند، زیرا در این الگوهـا، جایگاه‌های نزدیک به

هم دارای مقادیر مشابهی هستند.

الگوریتم رگرسیون خطی که هدف آن تخمین پارامترهای یک خط می‌باشد، داده های یک سری‌زمانی را با استفاده از یک خط شبیه‌سازی می‌نماید. به این صورت که مجموع مربعات اختلاف بین داده‌های اصلی(آموزشی) و داده‌های برآوردشده مینیمم گردد. رابطه ۳ معادله خط را نشان می‌دهد که در رگرسیون خطی به‌دنبال زوج مرتب (a,b) هستیم به نحوی که بیشترین شباهت را به روند داده‌ها داشته باشد. رابطه ۴ تابع بهینه‌سازی جهت کمینه نمودن مجموع مربعات اختلاف بین داده‌های اصلی(آموزشی) و داده‌های برآوردشده می‌باشد.

$$\hat{y} = ai + b \qquad (۳)$$

$$\min f = \sum_{i \in I} (y_i - \hat{y}_i)^2 \qquad (۴)$$

که در روابط بالا، a شیب خط رگرسیون، b عرض از مبدا خط، $\hat{y}$ مقدار برآوردشده به ازای شماره جایگاه در سری‌زمانی می‌باشد که در رابطه ۳ با نماد i نشان داده شده است. همچنین y_i مقدار واقعی جایگاه iام را نشان می‌دهد که مقدار برآوردشده آن با نماد $\hat{y}_i$ نشان داد هشده است.

ضریب همبستگی

شاخص ضریب همبستگی به منظور ارزیابی میزان مشابهت بین دو سری زمانی می‌تواند مورد استفاده قرار گیرد که عددی بین ۱- و ۱ می‌باشد. عدد ۱ معرف بیشترین شباهت، عدد صفر معرف عدم وجود شباهت و عدد ۱- معرف شباهت در جهت راستای منفی می‌باشد. به بیان دیگر زمانی مقدار این شاخص منفی می‌گردد که با افزایش مقدار پارامتر در یک سری زمانی، مقدار پارامتر دیگر در یک سری زمانی دیگر، کاهش داشته باشد. در این صورت مقدار ضریب همبستگی عددی منفی به خود می‌گیرد. رابطه ضریب همبستگی برای دو سری زمانی X و Y با استفاده از روابط ۵ و ۶ تعریف می‌گردد.

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \qquad (۵)$$

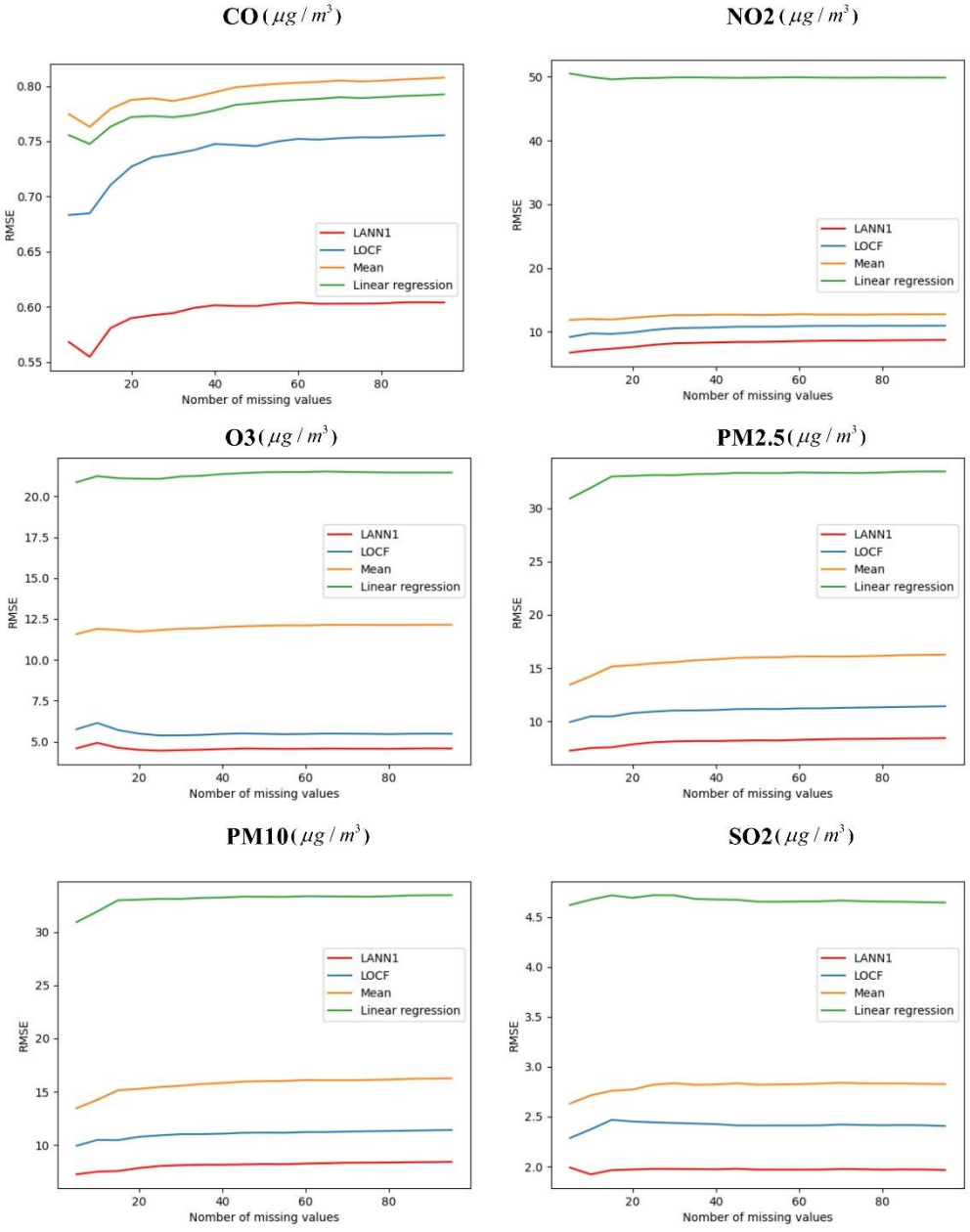
$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n} \qquad (۶)$$

که در رابطه ۵، $Corr(X,Y)$ نشان‌دهنده مقدار همبستگی می‌باشد. همچنین مقدار $Cov(X,Y)$ مقدار کوواریانس بین دو سری زمانی X و Y می‌باشد که از رابطه ۶ بدست می‌آید. همچنین σ_x و σ_y به ترتیب انحراف‌معیار داده‌های سری زمانی X و Y می‌باشند. همچنین مطابق با رابطه ۶، داده iام در سری زمانی های مربوطه به صورت X_i و Y_i نشان داده شده است. همچنین $\bar{X}$ و $\bar{Y}$ میانگین داده‌ها بوده و n تعداد کل داده‌ها می‌باشد.

نتایج

در این مطالعه برای پیاده‌سازی الگوریتم‌های LANN، LOCF، برنامه‌ریزی ژنتیک، میانگین و رگرسیون خطی، از داده‌های روزانه آلودگی هوای شهر تهران در بازه زمانی سال‌های ۱۳۹۷ تا ۱۴۰۰ به‌صورت متوالی استفاده شده است. به این منظور الگوریتمی جهت حذف مقادیری از سری زمانی در نظر گرفته شد. مطابق با الگوریتم مربوطه، تعدادی جایگاه به صورت تصادفی انتخاب گردیدند و مقادیر آن جایگاه‌ها حذف گردید. همچنین برای جلوگیری از به وجود آمدن جایگاه‌های حذف شده متوالی با طول زیاد، قیدی به الگوریتم مربوطه اضافه گردیدکه طبق آن دوجایگاه خالی و بیشتر نمی‌توانند به صورت متوالی در سری زمانی ایجاد شوند. بنابراین در این مطالعه فقط جایگاه‌های خالی منفرد (یک جایگاه خالی با همسایگی کامل) در نظر گرفته شده است. در ابتدا بر روی داده‌ها با استفاده از الگوریتم ذکر شده، به‌صورت تصادفی تعدادی از داده‌ها حذف گردید سپس با اعمال روش‌های بیان‌شده مقادیر حذف شده برآورد شدند و بعد از آن بر اساس مقادیر تخمینی و مقادیر واقعی مقدار RMSE محاسبه گردید. این فرآیند به ازای تعداد متفاوت مقادیر حذف شده انجام گردید. هم‌چنین در هر بار، مسئله ۳۰ مرتبه حل گردید. به این صورت که در سری زمانی داده‌ها ۳۰ مرتبه به‌صورت تصادفی مقادیری حذف گردیدند و سپس این مقادیر حذف شده با استفاده از روش‌های ذکر شده

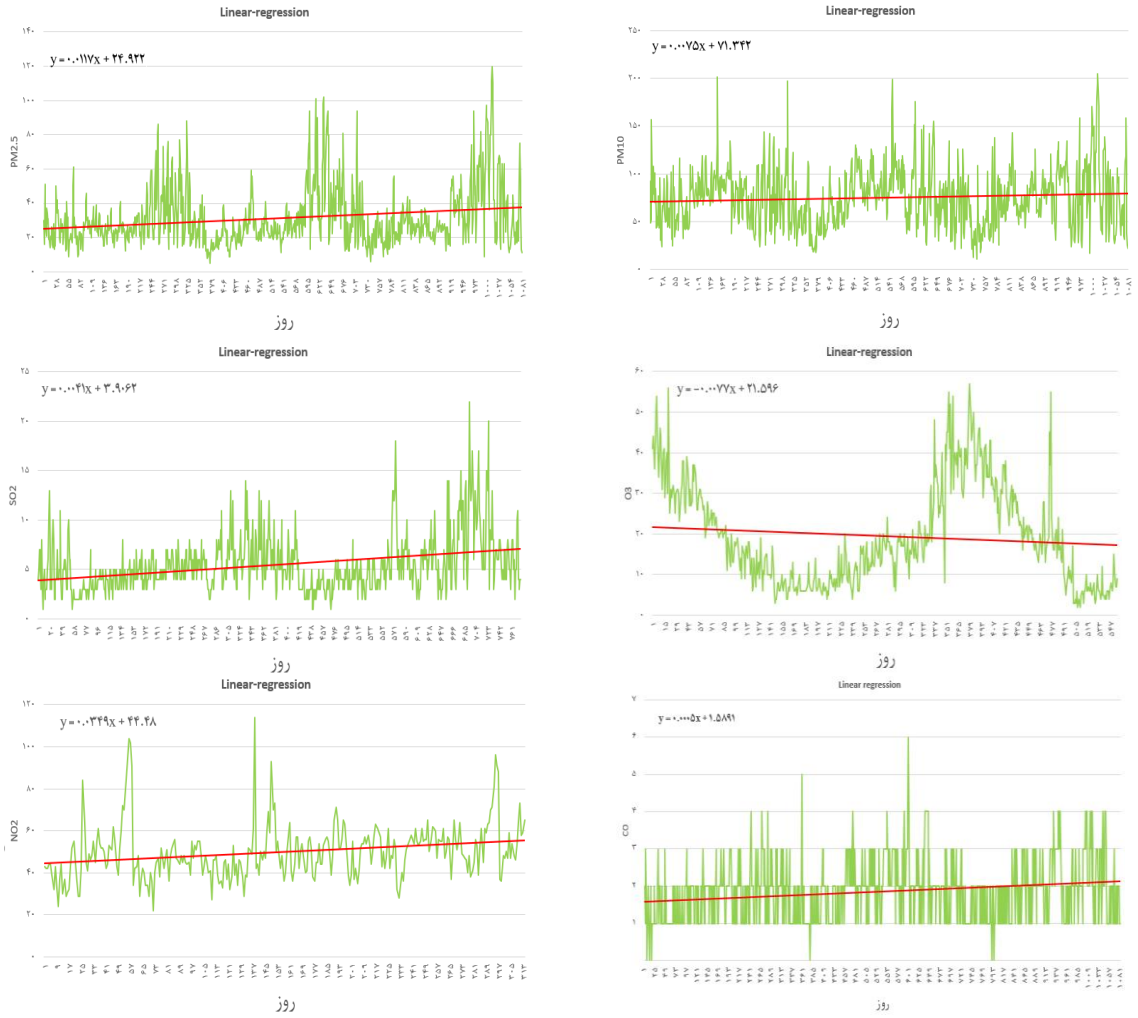
برآورد شدند. در شکل ۶ علاوه بر روش‌های معرفی‌شده در این مطالعه، روش‌های میانگین ساده، رگرسیون خطی و LOCF مورد ارزیابی قرار گرفته‌اند. این فرآیند روی تمامی آلاینده‌های PM2.5، PM10، SO2، O3، NO2 و CO اعمال گردیده است که در شکل ۶ که در بالا برای معرفی فرآیند ژنتیک معرفی شده است، قابل مشاهده می‌باشد.



شکل ۷ – مقایسه مقدار RMSE روش‌های مختلف تخمین مقادیر جاافتاده برای آلاینده‌های مختلف

همان‌طور که در شکل ۷ مشخص است روش LANN که در این مطالعه معرفی شده‌اند در همه آلاینده ها دارای RMSE کمتری نسبت به سایر روش‌ها بوده‌اند که این بدین معناست که برآوردهای آن‌ها به واقعیت نزدیک‌تر بوده است. روش میانگین ساده که در نمودار با نماد Mean و با رنگ نارنجی نشان داده شده است برای همه آلاینده ها جز آلاینده CO مقدار RMSE کمتری نسبت به رگرسیون خطی داشته است. هم‌چنین در آلاینده O3 بیشترین مقدار RMSE مربوط

به روش رگرسیون خطی بوده است. روش LOCF در برخی آلاینده‌ها مثل O3 و PM2.5 مقادیر تقریباً بهتری نسبت به سایر روش‌ها داشته‌اند به‌نحوی‌که مقدار RMSE این روش اندکی از روش LANN بیشتر بوده است. نکته حائز اهمیت در مقدار RMSE های شکل ۷ این است که در برخی آلاینده‌ها تغییرات در محدوده کوچکی نوسان می‌کند به همین دلیل مقدار RMSE در این آلاینده‌ها کمتر از آلاینده‌هایی است که دامنه نوسان بزرگ‌تری دارند. در آلاینده‌های NO2 و PM10 این مورد قابل مشاهده است. در روش رگرسیون خطی برای هر سری زمانی خطی برازش داده شده است که نتایج آن در شکل ۸ قابل مشاهده است.



شکل ۸– سری زمانی و معادله خطوط رگرسیون برای تمامی آلاینده‌های موردمطالعه برحسب ($\mu g / m^3$)

برای پیاده‌سازی روش رگرسیون خطی بر اساس اطلاعات سری زمانی بهترین معادله خطی به کل مجموعه برازش داده شده است که معادله خطوط در شکل ۸ قابل مشاهده است.

در ادامه پیاده‌سازی روش‌های تخمین مقادیر جاافتاده، روش GP بر روی داده‌های مربوطه اعمال گردید. قبل از پیاده‌سازی الگوریتم GP یک بررسی بر روی همبستگی روندهای موجود بین آلاینده‌های مختلف صورت گرفته است که در جدول ۱ به صورت ماتریسی نشان داده شده است.

جدول ۱ – همبستگی بین روندهای مربوط به آلاینده‌های مختلف

	PM2.5	PM10	SO2	NO2	CO	O3
PM2.5	۱	۰/۷۱	۰/۶۶	۰/۶۳	۰/۵۳	−۰/۴
PM10	۰/۷۱	۱	۰/۵۳	۰/۴۵	۰/۵۶	−۰/۰۵
SO2	۰/۶۶	۰/۵۳	۱	۰/۴۹	۰/۴۶	−۰/۳۰
NO2	۰/۶۳	۰/۴۵	۰/۴۹	۱	۰/۵۸	−۰/۳۶
CO	۰/۵۳	۰/۵۶	۰/۴۶	۰/۵۸	۱	−۰/۱۸
O3	−۰/۴	−۰/۰۵	−۰/۳۰	−۰/۳۶	−۰/۱۸	۱

همانطور که در جدول ۱ مشخص می‌باشد،همبستگی به صورت زوجی بین آلاینده‌ها نشان داده شده است. به عنوان مثال آلاینده PM2.5 بیشترین شباهت را با PM10 دارد که مقدار ضریب همبستگی برای آن ۰/۷۱ می‌باشد. همچنین مقدار ضریب همبستگی برای آلاینده O3 و سایر آلاینده‌ها مقداری منفی دارد که نشان می‌دهد روند سری زمانی O3 تقریبا در خلاف سایر آلاینده‌ها بوده است. البته این مقدار بین آلاینده‌های O3 و PM10 نزدیک به صفر بوده است که نشان می‌دهد که شباهت چشم‌گیری بین روند این دو آلاینده وجود ندارد. در روش GP که از روش‌های چندمتغیره می‌باشد، برای تخمین آلاینده PM2.5 از سایر آلاینده‌های محیطی استفاده شده است. این آلاینده‌ها عبارت‌اند از PM10، SO2، NO2، CO و O3. خروجی روش GP یک معادله بر اساس سایر آلاینده‌ها می‌باشد که بر اساس آن می‌توان مقادیر جاافتاده را تخمین زد. همانطور که گفته شد روش GP براساس عملگرهای پایه‌ای که به الگوریتم معرفی می‌گردد، اقدام به تولید جمعیت اولیه می‌نماید همچنین دو مرحله تلفیق و جهش براساس این عملگرهای پایه صورت می‌گیرد . تنظیمات الگوریتم GP که در این مطالعه استفاده شده است، مطابق جدول ۲ می‌باشد که از مطالعات Al-Helali و همکاران (۲۰۲۱) و Tran و همکاران (۲۰۱۷) استفاده شده است. عملگرهای پایه‌ای مورد استفاده در این مطالعه عبارتند از جمع، تقسیم، ضرب و جذر که با تلفیق ضرب و جمع می‌توان به عملگر تفریق رسید. همچنین با چندین بار استفاده از عملگر ضرب می‌توان عملگر توان را ایجاد نمود. همچنین الگوریتم GP یکبار با عملگرهای پایه‌ای جمع، ضرب و تقسیم و یکبار با درنظرگرفتن عملگرهای پایه‌ای جمع، ضرب، تقسیم و جذر انجام گرفته و اثر اضافه نمودن عملگر جذر بررسی شده است.

جدول ۲ – پارامترهای الگوریتم GP

پارامتر	مقدار(ها)
جمعیت اولیه	۵۱۲
روش انتخاب	رقابت
نرخ تلفیق	۰/۸
نرخ جهش	۰/۲
عملگرهای پایه‌ای	جمع– تقسیم– ضرب – جذر
تعداد تکرار (شرط توقف)	۱۰۰

این الگوریتم برای همه آلاینده‌ها رابطه‌ای بر اساس سایر آلاینده‌ها به دست آورده است که این روابط در جدول ۳ آورده شده است.

جدول ۳ – رابطه خروجی روش GP برای آلاینده‌های مختلف

فرمول های خروجی روش GP	
$PM2.5 = \sqrt{NO2} + SO2 + \frac{PM10}{\sqrt{O3}}$	(۷)
$PM10 = PM2.5 + \sqrt{\frac{CO \times SO2}{0.03}}$	(۸)
$SO2 = \sqrt{0.743 \times PM2.5}$	(۹)
$NO2 = 2\sqrt{O3} + CO + SO2 + PM2.5$	(۱۰)
$CO = \sqrt{SO2}$	(۱۱)
$O3 = \frac{PM10}{0.13 \times PM2.5}$	(۱۲)

در فرآیند مدل‌سازی از ۷۰ درصد داده‌ها به‌عنوان داده‌های آموزشی و ۳۰ درصد داده‌ها به‌عنوان داده‌های ارزیابی استفاده شده است پس از استفاده از الگوریتم برنامه‌ریزی ژنتیک و کشف روابط موجود بین داده‌ها، داده‌های ارزیابی جهت محاسبه مقدار RMSE مورد استفاده قرار گرفتند و رابطه‌ای که بیشترین مقدار RMSE را دارا بود، برای هر آلاینده انتخاب گردید . درنهایت برای بررسی میزان تطبیق‌پذیری روابط با روندهای موجود، روابط برای تخمین همه داده‌ها (روزهای سال) مورد استفاده قرار گرفتند که در شکل ۹ مقدار داده‌های اصلی و برآوردشده به ترتیب با رنگ‌های آبی و قرمز نشان داده شده‌اند که برای همه آلاینده‌ها ترسیم شده است. روابط استخراج شده از روند سری‌های زمانی آلاینده‌ها که در جدول ۳ قرار داده شده‌اند، حاصل رقابتی می‌باشند که در فرآیند GP صورت گرفته است و الگوریتم مربوطه با کشف روابط نهفته‌ای که بین آلاینده‌ها برقرار می‌باشد، روابطی را پیشنهاد می‌دهد که می تواند تطبیق‌پذیری مناسبی با روندهای موجود داشته باشد و الزاما خواص فیزیکی و شیمیایی آلاینده‌ها و تاثیراتی که به سبب این خواص بر روی همدیگر می‌گذارند، در نظر گرفته نشده است. همچنین روابط ارائه شده در جدول ۳ الزاما برگشت‌پذیر نیستند؛ زیرا روش GP با توجه به ماهیت تقریبی که دارد به جمعیت اولیه تولید شده وابسته می‌باشد و روابطی تقریبی تولید می‌کند؛ یعنی هر بار که الگوریتم اجرا می‌گردد ممکن است رابطه‌ای جدید ارائه دهد که براساس عملگرهای پایه‌ای که تعریف می‌گردد به شبیه ترین روابط به روندهای موجود دست پیدا می‌کند. بنابراین با تغییر عملگرهای پایه‌ای، فضای جستجوی الگوریتم به کلی تغییر می‌کند و روابطی ترکیبی از عملگرهای معرفی‌شده ارائه می‌کند. مثلاً، اگر برای عملگرهای پایه‌ای از توابع مثلثاتی استفاده گردد، آنگاه خروجی روش GP به صورت ترکیبی از روابط مثلثاتی بدست می‌آید که ماهیتی متناوب دارند.

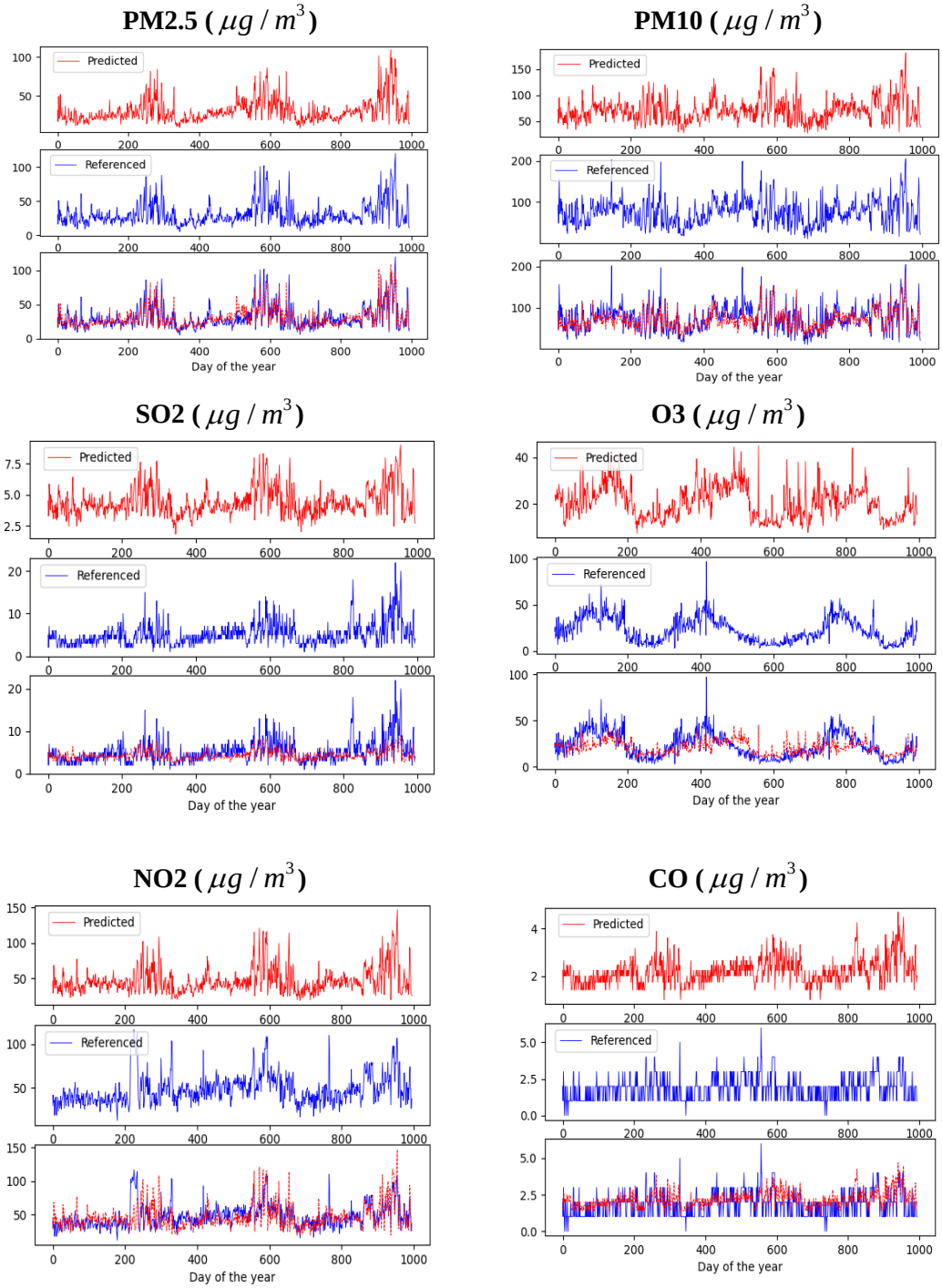
در جدول ۳ نتایج پیاده‌سازی تمامی روش‌ها آمده است به این صورت که به ازای تعداد مختلف جایگاه‌های خالی (بین ۵ تا ۱۰۰) فرآیند تکمیل آن‌ها با استفاده از الگوریتم های مختلف پیاده‌سازی شد و میانگین مقادیر RMSE به ازای هر روش مورد استفاده، در جدول ۴ قرار داده شده است.

جدول ۴ – مقایسه مقادیر RMSE روش‌های مورد بررسی برای آلاینده‌های مختلف برحسب ($\mu g / m^3$)

آلاینده	GP (بدون عملگر جذر)	GP (با عملگر جذر)	Linear Regression	Mean	LOCF	LANN
PM2.5	۸/۶۰	۷/۷۸	۱۶/۰۰	۱۶/۳۹	۱۰/۹۲	۸/۳۳
PM10	۱۹/۸۴	۱۸/۷۳	۲۹/۲۷	۲۹/۴۱	۲۳/۱۶	۱۷/۰۸
SO2	۲/۰۴	۱/۹۲	۲/۶۰	۲/۷۷	۲/۳۳	۱/۹۰
O3	۱۱/۵۲	۱۱/۵۲	۱۲/۸۳	۱۲/۰۴	۵/۴۰	۴/۳۵
NO2	۱۵/۴۳	۱۴/۱۶	۱۲/۱۷	۱۲/۷۳	۱۰/۵۲	۸/۴۴
CO	۰/۸۱	۰/۷۸	۰/۷۹	۰/۸۰	۰/۷۵	۰/۶۰

بحث و جمع بندی

در این مطالعه روش LANN برای تخمین مقادیر جاافتاده پیاده‌سازی شده است. پس از پیاده‌سازی روش، جهت اعتبارسنجی روش نتایج برای تعداد مقادیر جاافتاده مختلف با تعدادی بین ۵ تا ۱۰۰ به دست آمد و با مقایسه با سایر روش‌های استفاده شده در سایر مطالعات شامل روش‌های میانگین ساده، رگرسیون خطی و LOCF مقایسه گردید. از بین چهار روش رگرسیون خطی، LOCF، میانگین و LANN روش رگرسیون خطی در تمامی آلاینده‌ها با فاصله ای معنادار از سایر روش‌ها قرار گرفته است که نشان می‌دهد رگرسیون خطی برای آلاینده‌هایی که دارای روند مشخصی (صعودی یا نزولی) نیستند، عملکرد مناسبی نداشته است



شکل ۹– سری‌های زمانی اصلی و تخمینی برای آلاینده‌های مختلف براساس پیاده‌سازی الگوریتم GP

اگرچه در آلاینده CO که روند نسبتا ثابتی دارد دو الگوریتم میانگین و رگرسیون خطی وضعیت مشابهی داشته‌اند. درواقع برای تخمین آلاینده‌هایی که یک روند متناوب داشته و در برخی فصل‌ها افزایش و دربرخی فصل‌ها کاهش می‌یابند، روش رگرسیون RMSE بالایی را نشان می‌دهد.همچنین دو الگوریتم LANN و LOCF مطابق شکل ۷، دارای وضعیت بهتری بوده‌اند و در آلاینده‌هایی که در نواحی محلی کوچک تغییرات جزئی داشته‌اند این دو الگوریتم دارای نتایج بهتری بوده‌اند مثل دو آلاینده NO2 و O3. این دو آلاینده، دارای یک روند مشخص می‌باشند. آلاینده NO2 با توجه به ماهیت

متناوبی که دارد نیازمند الگوریتم‌هایی می‌باشند که بتواند ویژگی تناوب را به خوبی تخمین بزند. مثل رگرسیون‌هایی که بر مبنای توابع متناوب شکل می‌گیرند. بابررسی صورت گرفته برروی سری‌زمانی آلاینده‌ها، تمامی آلاینده‌ها به غیراز آلاینده های CO و NO2 دارای روندی نسبتا متناوب می‌باشند که این روند در سری‌زمانی مربوط به آلاینده O3 ملموس‌تر می‌باشد. استفاده از الگوریتم‌هایی که روند کلی سری زمانی را برای تخمین در نظر می‌گیرند مثل میانگین و رگرسیون خطی، نمی توانند دارای کارآمدی مناسبی باشند.

همچنین روش LANN نسبت به سایر روش‌ها نتایج بهتری داشته است و مقدار RMSE در این روش در حدود ۳۰ تا ۵۰ درصد کمتر بوده است. هم‌چنین در ادامه روش‌های قبلی با استفاده از روش GP که از روش‌های چندمتغیره می‌باشد روابطی برای تخمین آلاینده‌های مختلف استخراج گردید و مقدار RMSE محاسبه شده برای این روش با روش‌های تک‌متغیره، مقایسه گردید. این الگوریتم در برآورد شاخص PM2.5 نسبت به سایر روش‌های تک‌متغیره، موفق‌تر بوده است. یکی از مهم‌ترین مزیت‌های روش GP بر روش‌های تک متغیره، عدم وابستگی نتایج به داده‌های قبل و بعد مقدار جاافتاده موردنظر است که برای مواقعی که تعداد مقادیر جاافتاده زیاد و سری زمانی دارای مقادیر جاافتاده متوالی بزرگ است، روش GP مناسب‌تر از روش‌های تک متغیره می‌باشد. همچنین روش GP با استفاده از سری‌های زمانی موجود به روابطی رسیده است که ممکن است با روابط تجربی که براساس خصوصیات فیزیکی و شیمیایی آلاینده‌ها تعیین شده‌اند همخوانی نداشته باشند، درحالی که رویکرد روش GP براساس روندهای سری زمانی شکل می‌گیرد و یک رابطه کلی برای تقریب سری زمانی پیشنهاد می‌دهد و هدف کمینه نمودن RMSE کل مجموعه می‌باشد. با توجه به نتایج بدست آمده در این مطالعه، روش GP می‌تواند در مواردی برآوردهای مناسبی داشته باشد. البته با توجه به ماهیت تقریبی روش GP تعیین مناسب عملگرهای پایه‌ای و تنظیم پارامترهای الگوریتم نقش موثری در روابط پیشنهادی دارد.

براساس نتایج پیاده‌سازی الگوریتم GP یکبار با در نظر گرفتن عملگر جذر و یک بار بدون در نظرگرفتن آن، نتایج حاکی از کاهش مقدار RMSE بوده است به صورتی که تأثیر اضافه نمودن عملگر جذر برروی آلاینده‌های PM2.5، PM10، SO2، NO2 و CO، به ترتیب کاهش ۱۰/۵، ۵/۹، ۶/۲، ۸/۹ و ۳/۸ درصدی در مقدار RMSE بوده است. عملگر جذر تأثیری برروی مقدار آلاینده O3 نداشته است و رابطه استخراج شده در هر دو حالت یکسان بوده و عملگر جذر نقشی در تشکیل رابطه نداشته است.

در تخمین آلاینده PM2.5 الگوریتم GP بهترین برآورد را داشته است و مقدار RMSE، ۷,۷۸ بدست آمده است. در مطالعه Hadeedو همکاران (۲۰۲۰)، مقدار RMSE برای آلاینده PM2.5 در بهترین حالت ۱۴/۳۴ به ازای ۲۰ درصد ناکاملی مجموعه داده، بدست آمده است. همچنین در مطالعه Yuan و همکاران (۲۰۱۸)، مقدار RMSE برای آلاینده PM2.5، ۹/۰ بدست آمده است. همچنین در مطالعه Shahbazi و همکاران (۲۰۱۸) مقدار RMSE برای تخمین PM2.5، ۹/۴۷ ±۱/۲۵ بدست آمده است.

برای برآورد PM10 در این مطالعه در مقایسه با مطالعه Shahbazi و همکاران (۲۰۱۸) ، دو الگوریتم LANN و GP مناسب عمل کرده‌اند به نحوی که مقدار RMSE بدست‌آمده با به‌کارگیری دو الگوریتم LANN و GP به ترتیب ۱۷/۰۸ و ۱۸/۷۳ بوده است. در مطالعه Shahbazi و همکاران (۲۰۱۸) ، این شاخص با مقدار RMSE، ۲۰/۷ ±۳/۱ بدست آمده است. شاخص SO2 به دلیل نوسانات کوچکتری که در مقادیر ثبت شده نسبت به سایر آلاینده‌ها دارد، مقادیر RMSE کوچکتری بدست آمده است. دو الگوریتم LANN و GP با مقادیر RMSE به ترتیب ۱/۹۰ و ۱/۹۲ دارای نتایج بهتری نسبت به روش‌های LOCF، میانگین و رگرسیون خطی بوده‌اند.

همچنین آلاینده O3 با به‌کارگیری دو الگوریتم LANN و LOCF با مقادیر RMSE به ترتیب ۴/۳۵ و ۵/۴۰ بهترین نتایج را در مقایسه با سایر روش‌های پیاده‌سازی شده در این مطالعه داشته است. در مطالعه Shahbazi و همکاران (۲۰۱۸)، این آلاینده با مقدار RMSE، ۱/۵۷ ± ۶/۶۷ برآورد شده است. دو الگوریتم ذکر شده در برآورد NO2 نیز بهترین نتایج را داشته اند به نحوی که الگوریتم LANN با مقدار RMSE، ۸/۴۴ و الگوریتم LOCF با مقدار RMSE ۱۰/۵۲ شاخص NO2 را برآورد کرده‌اند. مطالعه Mishchuk و همکاران (۲۰۱۹) با به‌کارگیری الگوریتم SGTM توانسته آلاینده NO2 را با مقدار RMSE، ۷/۴۰ برآورد کند. دراین مطالعه، برای برآورد CO، الگوریتم LANN با مقدار RMSE، ۰/۶۰ در مقایسه با سایر روش‌ها کارایی بهتری داشته است. همچنین این آلاینده در مطالعه Yuan و همکاران (۲۰۱۸)، با مقدار RMSE، ۰/۲۴۶ برآورد شده است.

هدف اصلی این مطالعه بررسی نحوه جایگذاری مقادیر جاافتاده در سری‌های زمانی بوده است که الگوریتم‌های ارائه شده می‌توانند به عنوان یک ابزار برای تخمین مقادیر جاافتاده مورد استفاده قرار گیرند. سازمان های ذیربط مثل مدیریت بحران می‌توانند داده‌های ناقص ثبت‌شده توسط ایستگاه‌های ثبت آلودگی هوا را با استفاده از این الگوریتم‌ها تکمیل نمایند که می‌تواند منجر به یک تصمیم‌گیری با پشتوانه اطلاعات صحیح و کامل شود.

نتیجه گیری

در این مطالعه از الگوریتـم‌های تک متغیره و الگوریتـم GP به‌عنوان الگوریتم چندمتغیره استفاده شده است. البته الگوریتم‌های مبتنی بر هوش مصنوعی می‌توانند در تخمین مقادیر جاافتاده به‌کارگرفته شوند. همچنین این مطالعه هر آلاینده را بر اساس سایر آلاینده‌ها برآورد کرده است. می‌توان با اضافه نمودن پارامترهای هواشناسی مثل دما، رطوبت و باد و نیز اطلاعات مربوط به زمین مثل ارتفاع، شیب و جهت شیب به مسئله، ارتباط این پارامترها با میزان آلودگی هوا بررسی گردد. همچنین با توجه به تاثیر مکان بر نتایج بدست آمده از روش‌های گوناگون، روش‌های ارائه شده در این مطالعه و مطالعات پیشین می‌توانند برروی یک منطقه یکسان بررسی گردند. عملگرهای پایه‌ای استفاده شده در الگوریتم GP که در این مطالعه به آن پرداخته شده است، شامل جمع، تقسیم، ضرب و جذر بوده است به نظر می‌رسد استفاده از سایر عملگرها متناسب با روند آلاینده مربوطه می‌تواند تخمین‌های دقیق تری را ارائه دهد. پیشنهاد می‌گردد برای آلاینده هایی با روندهای متناوب از عملگرها و توابعی استفاده گردد که بتواند ویژگی تناوب را به خوبی نمایش دهد. همچنین به نظر می‌رسد تبدیل کردن مجموعه داده‌ها به زیرمجموعه‌هایی که دارای روند ثابتی هستند، می‌تواند سبب کارآمدی الگوریتم‌هایی کلی‌نگر مثل رگرسیون گردد.

یادداشت‌ها

1. Last Observation Carried Forward (LOCF)
2. Local Average of Nearest Neighbor filter (LANNf)
3. Local Average of Nearest Neighbor (LANN)
4. Genetic Programming (GP)

فهرست منابع

Al-Helali, B., Chen, Q., Xue, B., & Zhang, M. (2021). A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data. *Soft Computing*, 1-20 .

Aljuaid, T., & Sasi, S. (2016). *Proper imputation techniques for missing values in data sets*. 2016 International Conference on Data Science and Engineering (ICDSE) ,

Andiojaya, A., & Demirhan, H. (2019). A bagging algorithm for the imputation of missing values in time series. *Expert Systems with Applications*, *129*, 10-26.

Ashrafi, K., & Ahmadi Orkomi, A. (2014). Atmospheric stability analysis andits correlationwith the concentration of air pollutants: a case study ofa critical air pollution episode in Tehran. *Iranian Geophys*, 8(3), 49−61

Ashrafi, K., & Hoshyaripour, G. A. (2010). A model to determine atmospheric stability and its correlation with CO concentration. *International Journal of Civil and Environmental Engineering*, *2*(2), 82-88 .

Bokde, N., Beck, M. W., ءlvarez, F. M., & Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern Recognition Letters*, *116*, 88-96 .

Caillault, É. P., Lefebvre, A., & Bigand, A. (2020). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*, *139*, 139-147 .

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, *56*(10), 968-976.

Flores, A., Tito, H., & Centty, D. (2019). Model for time series imputation based on average of historical vectors, fitting and smoothing. *IJACSA) International Journal of Advanced Computer Science and Applications*, *10*(10), 346-352 .

Flores, A., Tito, H., & Silva, C. (2019). Local average of nearest neighbors: Univariate time series imputation. *International Journal of Advanced Computer Science and Applications*, *10*(8), 45-50 .

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), 755-763 .

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*, *730*, 139-140 .

Hamami, F., & Dahlan, I. A. (2020). *Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network*. 2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS) ,

Junger, W., & De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96-104 .

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), 2895-2907 .

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, *74*(7), 1-16 .

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*, 105713 .

Mishchuk, O., Tkachenko, R., & Izonin, I. (2019). *Missing data imputation through SGTM neural-like structure for environmental monitoring tasks*. International Conference on Computer Science, Engineering and Education Applications ,

Noor, N. M., Yahaya, A. S., Ramli, N. A., & Al Bakri Abdullah, M. M. (2015). Filling the missing data of air pollutant concentration using single imputation methods. In *Applied Mechanics and Materials* (Vol. 754, pp. 923-932). Trans Tech Publications Ltd.

Plaia, A., & Bondi, A. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, *40*(38), 7316-7330 .

Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Shaadan, N., & Rahim, N. (2019). *Imputation Analysis for Time Series Air Quality (PM10) Data Set: A Comparison of Several Methods*. Journal of Physics: Conference Series ,

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*, 24-33 .

Tito, H., Flores, A., & Silva, C. (2019). Local average of nearest neighbors: univariate time series imputation. *International Journal of Advanced Computer Science and Applications*, *10*(8), 45-50 .

Tran, B. N. (2018). *Evolutionary computation for feature manipulation in classification on high-dimensional data*.Victoria University of Wellington.

Tran, C. T., Zhang, M., Andreae, P., & Xue, B. (2017, July). Multiple imputation and genetic programming for classification with incomplete data. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 521-528).

Yicun, G., Mohammad Khorshiddoust, A., Mohammadi, G. H., Hoseini Sadr, A., & Aghlmand, F. (2020). The relationship between PM2. 5 concentrations and atmospheric conditions in severe and persistent urban pollution in Tabriz, northwest of Iran. *Arabian Journal of Geosciences*, *13*(5), 1-12 .

Yuan, H., Xu, G., Yao, Z., Jia, J., & Zhang, Y. (2018, October). Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (pp. 1293-1300).

Zeileis, A., & Grothendieck, G. (2005). Zoo: S3 infrastructure for regular and irregular time series. *ArXiv preprint math/0505527* .