2021, Vol.8, No.3, Pages 31 to 42
Journal of Water and Sustainable Development

سال هشتم، شماره ۳، ۱۴۰۰، صفحات ۳۱ تا ۴۲
نشریه آب و توسعه پایدار

Article Type: Applied

نوع مقاله: کاربردی

Introduction of MICE Method for Imputation Missing Meteorological Data and Comparison by Regression; Case Study: 130 Years of Monthly Temperature in Mashhad, Jask and Bushehr

M. Farzandi[1*], H. Rezaei-Pazhand[2]

1- Ph.D. in Agricultural Meteorology, Ferdowsi University of Mashhad, Faculty of Agriculture, Department of Water Engineering, Iran. 2- MSc in Hydrology, Islamic Azad University, Mashhad Branch, Iran.

*(Corresponding Author Email: mhb_farzandi@yahoo.com)

Received: 16-04-2021 Revised: 03-06-2021
Accepted: 28-06-2021 Available Online: 06-12-2021

معرفی روش MICE در ترمیم داده‌های گمشده هواشناسی و مقایسه با رگرسیون؛ مطالعه موردی: ۱۳۰ سال دمای ماهانه مشهد، جاسک و بوشهر

محبوبه فرزندی[1*]، حجت رضایی‌پژند[2]

۱- دکترای هواشناسی کشاورزی، دانشگاه فردوسی مشهد، دانشکده کشاورزی، ایران.
۲- کارشناسی‌ارشد هیدرولوژی، دانشگاه آزاد اسلامی مشهد، ایران.

*(نویسنده‌ی مسئول، E-Mail: mhb_farzandi@yahoo.com)

تاریخ دریافت: ۱۴۰۰/۰۱/۲۷ تاریخ بازنگری: ۱۴۰۰/۰۳/۱۳
تاریخ پذیرش: ۱۴۰۰/۰۴/۰۷ تاریخ انتشار: ۱۴۰۰/۰۹/۱۵

Abstract

Requiring accurate, complete and reliable data is the first step in climate studies. Incomplete data challenges climate analysis. Missing (incomplete) data is often found in meteorology. Therefore, completing the data (imputation) is the primary need for analysis. There are several ways to imputation missing data that vary depending on the data type and climatic characteristics of each region. Precipitation and temperature are the most important variables of meteorology and climatology. The length of the statistical period plays a pivotal role in the accurate analysis of these variables. The monthly temperature of three cities in Iran, including Mashhad, Bushehr and Jask, has been available in a book called World Weather Records since about 1890. This information contains missing data, especially during World War II (1941-1949). This missing data is more visible. The purpose of this study is to increase the accuracy of estimating these missing data by introducing the applied MICE method and providing a complete series of monthly temperatures over 130 years. Stations from neighboring countries were selected as independent (predictor) stations in the patterns. First, the missing monthly temperature data of these three stations were estimated by fitting regression patterns (RMSE of 0.71 to 0.94 OC). The classical regression method requires the study of basic hypotheses and pattern pathology. These patterns were also estimated by the MICE method (RMSE of 0.39 to 0.82 OC). The results of the study and implementation of this package in Rstudio show the superiority of this method. This method is designed for missing data, does not have regression problems, and has many capabilities. Therefore, it is recommended to estimate missing meteorological data.

Keywords: Missing Data, Regression, MICE Algorithm, Temperature, Basic Regression Hypotheses.

چکیده

نیاز به داده‌های کامل و قابل اطمینان اولین گام در مطالعات اقلیمی است. داده‌های ناکامل، تحلیل‌های اقلیمی را دچار چالش می‌کند. اغلب در آب و هواشناسی داده‌های گمشده (ناکامل) وجود دارد. بنابراین کامل‌کردن داده‌ها (ترمیم) نیاز اولیه تحلیل‌هاست. روش‌های متعددی برای بازسازی داده‌ها وجود دارند که بسته به نوع داده و خصوصیات آب و هوایی هر منطقه متفاوت می‌باشند. بارش و دما از مهم‌ترین متغیرهای هوا و اقلیم‌شناسی هستند. طول دوره آماری اهمیت بسزایی در دقت تحلیل این دو متغیر دارد. دمای ماهانه سه شهر ایران شامل مشهد، بوشهر و جاسک از سال حدود ۱۸۹۰ در کتبی به نام World Weather Records موجود است. این اطلاعات دارای داده‌های گمشده می‌باشد، مخصوصاً همزمان با جنگ جهانی دوم (۱۹۴۱-۱۹۴۹) این داده‌های گمشده مشهودتر هستند. هدف این پژوهش، افزایش دقت برآورد این داده‌های مفقود با معرفی روش کاربردی MICE و ارائه سری کامل دمای ماهانه در طول ۱۳۰ سال است. بدین منظور، ایستگاه‌هایی از کشورهای مجاور به‌عنوان ایستگاه‌های مبنا انتخاب شدند. ابتدا داده‌های مفقود دمای ماهانه این سه ایستگاه با برازش الگوهای رگرسیونی ترمیم شدند (ریشه میانگین مربعات خطا ۰/۷۱ تا ۰/۹۴ درجه سانتیگراد). روش کلاسیک رگرسیون نیازمند بررسی فرض‌های زیربنایی و آسیب‌شناسی است. این الگوها با روش MICE نیز برآورد شدند (ریشه میانگین مربعات خطا ۰/۳۹ تا ۰/۸۲ درجه سانتیگراد). نتایج مطالعه و اجرای این بسته در محیط Rstudio نشان از برتری این روش دارد. این روش برای داده‌های مفقود طراحی شده، مشکلات رگرسیون را نداشته و قابلیت‌های زیادی دارد. لذا برای ترمیم داده‌های گمشده آب و هواشناسی پیشنهاد می‌شود.

واژه‌های کلیدی: داده گمشده، رگرسیون، الگوریتم MICE، دما، فرض‌های زیربنایی رگرسیون.

HomePage: https://jwsd.um.ac.ir

How to cite this article: Farzandi M. and Rezaei-Pazhand H. 2021. Introduction of MICE Method for Imputation Missing Meteorological Data and Comparison by Regression; Case Study: 130 Years of Monthly Temperature in Mashhad, Jask and Bushehr. Journal of Water and Sustainable Development, 8(3): 31-42. doi: 10.22067/jwsd.v8i3.2104.1038

مقدمه

ترمیم داده‌های گمشده[1] (مفقود) در علوم مختلف اهمیت ویژه‌ای دارد. تحلیل‌های کلاسیک آماری با نمونه‌های کامل (بدون مفقودی) ممکن است و در نمونه‌های شامل مفقودی نتایج تحلیل‌ها اریب هستند، یعنی در صحت آنها تردید وجود دارد (ارقامی و همکاران، ۱۳۸۰). نیاز به داده‌های صحیح، کامل و قابل اطمینان اولین گام در مطالعات اقلیمی است. داده‌های ناکامل، تحلیل‌های اقلیمی را دچار چالش می‌کند. اغلب در آب و هواشناسی داده‌های گمشده (ناکامل) موجود است. بنابراین کامل کردن داده‌ها (ترمیم) نیاز اولیه تحلیل‌هاست. روش‌های متعددی برای بازسازی داده‌ها وجود دارند که بسته به نوع داده و خصوصیات آب و هوایی هر منطقه متفاوت می‌باشند. دما و بارش از مهم‌ترین متغیرهای آب و هوایی هستند. داده‌های طولانی مدت و کامل متغیرهای دما و بارش اولین نیاز تحلیل‌های قابل اعتماد در آب و هواشناسی و علوم مرتبط است. طول دوره آماری اهمیت بسزایی در دقت تحلیل این دو متغیر دارد. حجم نمونه کمتر از ۱۰۰ سال (بویژه در مناطق خشک و نیمه خشک) نمی‌تواند نوسانات دراز مدت را به خوبی منعکس کند (Edmond و همکاران، ۱۹۷۳).

آمار مشاهده‌ای متغیرهای هواشناسی در چند ایستگاه همدید ایران از سال ۱۳۳۰ (۱۹۵۱ میلادی) در سایت سازمان هواشناسی ایران قابل دسترس است (www.weather.ir). بنابراین حداکثر طول دوره آماری رسمی کشور ۷۰ سال است. آمار قدیمی و طولانی مدت دمای ماهانه سه شهر و بارش ماهانه پنج شهر ایران توسط سفارت امریکا و انگلیس در دوره قاجار و قبل از سال۱۳۳۰ اندازه‌گیری و در کتبی به نام World Weather Records ثبت شده است (Smithsonian Institution، ۱۹۲۷). این اطلاعات نشان می‌دهد که طول دوره آماری دما و بارش ماهانه این ایستگاه‌ها حدود ۱۳۰ سال است. ایستگاه‌های فوق، تنها ایستگاه‌های با آمار طولانی مدت در ایران هستند. متاسفانه این آمار که می‌تواند مبنای ارزشمندی در تحقیقات و نشان‌دهنده تغییرات بلندمدت دوره‌ای و روند و ... باشند، دارای داده گمشده بوده و تا به حال در دسترس پژوهشگران قرار نگرفته است. بررسی گمشده‌های ماهانه این ایستگاه‌ها نشان می‌دهد که اکثر آنها به دنبال جنگ جهانی اول (۱۹۱۸ و ۱۹۱۹) و جنگ جهانی دوم (۱۹۴۱-۱۹۴۹) رخ داده‌اند. برخی داده‌های گمشده ماهانه نیز به طور پراکنده در طول دوره آماری وجود دارند (Smithsonian Institution، ۱۹۴۷). نام این پنج شهر و سال شروع آماربرداری آنها عبارت‌اند از: مشهد (دما ۱۸۸۵و بارش ۱۸۹۳)، تهران (بارش ۱۸۸۴)، اصفهان (بارش ۱۸۹۳)، جاسک (دما ۱۸۹۳ و بارش ۱۸۹۳) و بوشهر (دما ۱۸۷۸ و بارش ۱۸۷۸). بنابراین این پنج ایستگاه دارای بارش و دمای ماهانه دراز مدت (۱۲۵ تا ۱۴۰ سال) است. آمار طولانی مدت بارش این ایستگاه‌ها پس از ترمیم به بیش از ۱۰۰ سال خواهد رسید. بنابراین می‌تواند مبنای تحلیل‌های دقیق‌تر و مطمئن‌تری برای دما و بارش ماهانه و سالانه این پنج شهر باشد. چون این پنج شهر در اقصا نقاط ایران قرار دارند، هریک می‌توانند نماینده ناحیه‌ای از کشور باشد. تاکنون ترمیم این داده‌ها در مقیاس ماهانه (به جز دمای مشهد توسط فرزندی و همکاران، ۱۳۹۳) انجام نشده است. ترمیم این آمار در قلمروی برآورد داده‌های گمشده است. داده‌های گمشده سبب کاهش دقت برآورد می‌شود. یعنی برآوردها اریب هستند و صحت تحلیل‌های ما در مورد جامعه دچار تردید است (Little وRubin ، ۲۰۰۲؛ ارقامی و همکاران، ۱۳۸۰).

حجم نمونه (طول دوره آماری) به ویژه در مناطق خشك و نیمه خشك برای تحلیل فراوانی، تحلیل سری‌های زمانی، تحلیل خشکسالی‌ها و ... باید حداقل۱۰۰ سال باشد، زیرا داده‌هایی با طول کمتر نوسانات دراز مدت را منعکس نمی‌کنند. دوره بازگشت نیز به طول دوره آماری وابسته است. برآورد بزرگترین دوره بازگشت با دقت قابل قبول معادل یك پنجم طول داده‌ها است (Jacobو همکاران، ۱۹۹۹). بنابراین در اختیار داشتن آمار طولانی مدت و کامل اولین نیاز تحلیل‌های قابل اعتماد در آب و هواشناسی است. این خود نیاز به روش‌های دقیق‌تری برای برآورد داده‌های مفقود دارد. زیرا مفقودی در داده‌های آب و هواشناسی معمول است.

تاکنون روش‌های مختلفی برای ترمیم داده‌های مفقود پیشنهاد شده است که هر یک بر اصول ریاضی خاصی بنا شده‌اند. رگرسیون به عنوان یک روش کلاسیک آماری با روش برآورد کمترین مربعات، کاربرد زیادی در آب و هواشناسی دارد (رضایی پژند و بزرگ نیا، ۱۳۸۱). Iqbal و همکاران (۲۰۱۸) تغییرات بارش در دریای زرد چین را در دوره آماری ۲۰۱۴-۱۹۶۰ بررسی کرده‌اند. آن‌ها روش رگرسیون خطی را برای برآورد داده‌های مفقود بارش به کار برده‌اند. طولانی‌ترین آمار دما و بارش ماهانه ایران در شهر مشهد و از سال۱۸۹۳ (بیش از ۱۲۵ سال) قابل دسترس است (فرزندی و همکاران، ۱۳۹۳). این سری‌های زمانی دارای داده‌های گمشده هستند. ترمیم ماهانه داده‌های مشهد با الگوهای رگرسیونی و بهینه‌سازی انجام شده است (فرزندی و همکاران، ۱۳۹۳). همچنین دو پژوهش در مقیاس سالانه نیز بر روی بارش داده‌های فوق صورت گرفته است. ابتدا مفقودی‌ها ترمیم و سپس داده‌های کامل سالانه تحلیل شده‌اند. خلیلی و بذرافشان (۱۳۸۷) تداوم خشکسالی‌ها را با تحلیل فراوانی بارش سالانه طولانی مدت مشهد بررسی کردند. آن‌ها روش خودهمبستگی را برای ترمیم بارش سالانه انتخاب کردند. Ghahraman و Ahmadi (۲۰۰۷) پانزده سال مفقودی بارش سالانه مشهد را

با روش کریجینگ و برازش رگرسیون چندگانه بر میانگین‌های متحرک باران سالانه ترمیم کردند. روش خودهمبستگی فقط از اطلاعات درون خود داده‌ها استفاده می‌کند. MICE

داده‌های گمشده در سری داده‌ها باید شرایطی داشته باشند که معمولاً روش‌های برآورد داده مفقود بتوانند آنها را با دقت مناسبی ارزیابی کنند. مثلا باید تصادفی یا کاملاً تصادفی باشند. از روش‌هایی که برای برآورد داده مفقود طراحی شده‌اند می‌توان به الگویتم‌های EM[2]، MI[3]، MICE[4] و .... اشاره کرد. Porto و همکاران (۲۰۱۷) بارش روزانه دو منطقه همگن اقلیمی را در تاریخ‌های مشخص شده با سه روش MICE، کریجینگ و کوکریجینگ برآورد نمودند. آنها نتیجه گرفتند روش MICE در مقایسه با روش‌های زمین آماری در منطقه اول ۰/۱۶ و در منطقه دوم ۰/۲۶ بهتر بود. Yozgatligil و همکاران (۲۰۱۳) شش روش ترمیم داده‌های مفقود را برای بارش و دمای ماهانه ترکیه ارزیابی و مقایسه نمودند. روش‌ها در این تحقیق به دو دسته ساده و پیچیده تقسیم شدند. روش‌های ساده عبارتند از میانگین حسابی، نسبت نرمال (NR)[5] و نسبت نرمال وزنی با روش همبستگی. روش‌های پیچیده شامل پرسپترون چندلایه شبکه عصبی مصنوعی (ANN)[6]، استراتژی جاگذاری چندگانه با زنجیره مارکوف-مونت کارلو بر اساس حداکثرکردن امید ریاضی (EM-MCMC)[7] و روش اصلاح شده EM-MCMC است. آنها مجموع مربعات خطا را برای ارزیابی و انتخاب روش برتر به کار بردند. افزون بر این روش تحلیل سری‌های زمانی پویای غیرخطی را با روش همبستگی بی‌بعد نیز برای وابستگی فضایی مکانی داده‌های جاگذاری شده به کار گرفتند. تحلیل آنها نشان داد که دو روش ANN و EM-MCMC از بقیه بهتر عمل می‌کنند.

هدف این پژوهش افزایش دقت برآورد داده‌های مفقود در آمار دمای طولانی مدت ماهانه سه شهر ایران و ارائه سری‌های بلندمدت است. در این راستا روش MICE برای برآورد داده‌های مفقود هواشناسی برای اولین بار در ایران معرفی شده و با روش کلاسیک رگرسیون مقایسه می‌شود. این تحقیق از دو جنبه حائز اهمیت است. ۱- معرفی یک روش کارا و ساده برای برآورد داده‌های مفقود هواشناسی، ۲- تولید آمار بلندمدت دمای ماهانه سه شهر ایران که می‌تواند مبنای ارزشمندی برای مطالعات منابع آب، خشکسالی‌ها، تغییر اقلیم، گرمایش جهانی و... باشد.

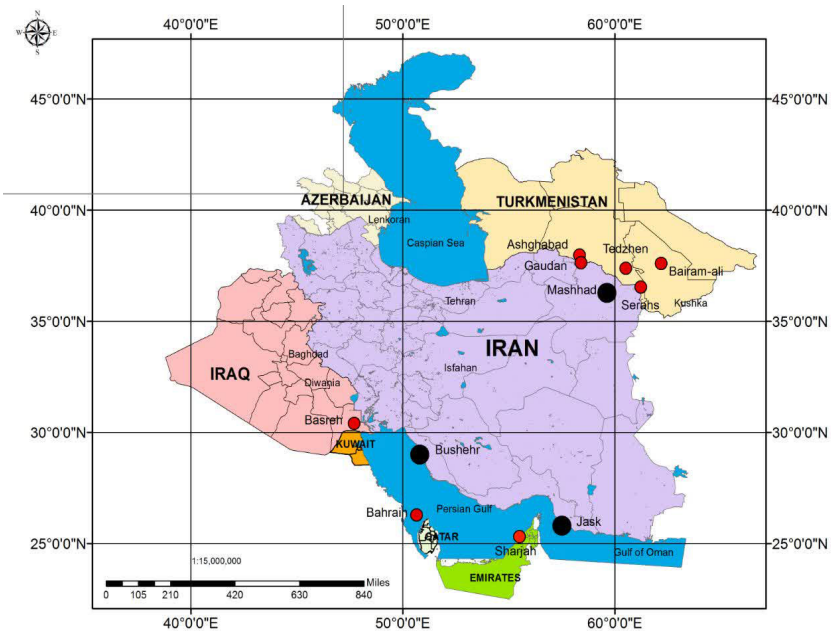
مواد و روش

دمای ماهانه سه ایستگاه در انتشارات WWR[8] منتسب به ایران با قدمتی ۱۲۵ تا ۱۴۰ سال موجود است (Smithsonian Institution، ۱۹۲۷؛ ۱۹۳۴؛ ۱۹۴۷). این ایستگاه‌ها شامل مشهد، جاسک و بوشهر است. متاسفانه این آمار که می‌تواند مبنای ارزشمندی در تحقیقات و نشان‌دهنده تغییرات دوره‌ای، روند و ... باشد، دارای داده‌های گمشده است. داده‌های گمشده عمدتاً مربوط به جنگ‌های جهانی اول و دوم می‌باشد.

این تحقیق قصد دارد ضمن جمع‌آوری این اطلاعات، آنها را با روش‌های دقیق و نوین ترمیم نموده و آمار کامل و طولانی مدت را برای اولین بار در اختیار پژوهشگران قرار دهد. این آمار از کتب WWR و سازمان هواشناسی کشور جمع‌آوری و سپس ترمیم شده است. دو روش رگرسیون و MICE در این تحقیق بررسی و مقایسه شده است. ایستگاه‌های مورد استفاده در این پژوهش در جدول (۱) و شکل (۱) آمده است. سه ایستگاه پاسخ با دایره‌های قرمز و ایستگاه‌های پیشگو با دایره‌های سیاه در شکل (۱) نشان داده شده است.

جدول ۱- مشخصات ایستگاه‌های منتخب مورد بررسی در الگوهای دما و بارش

نام ایستگاه	نام کشور	کد WMO	طول جغرافیایی	عرض جغرافیایی	ارتفاع (متر)	سال شروع آماربرداری	درصد گمشدگی بارش	درصد گمشدگی دما
مشهد	ایران	۴۰۷۴۵	۵۹/۶۳	۳۶/۲۷	۹۸۰	۱۸۹۳	۹/۱	۱۰/۴
جاسک	ایران	۴۰۸۹۳/۲	۵۷/۵۰	۲۵/۸۰	۴	۱۸۹۳	۱۹/۶	۷/۷
بوشهر	ایران	۴۰۸۵۸	۵۰/۸۰	۲۹/۰۰	۱۴	۱۸۷۸	۱۹/۵	۶/۴
سرخس	ترکمنستان	۳۸۹۷۴	۶۱/۲۲	۳۶/۵۳	۲۷۹	۱۹۰۲	-	-
کوشکا	ترکمنستان	۳۸۹۸۷	۶۲/۳۵	۳۵/۲۸	۵۷	۱۸۹۷	-	-
گودان	ترکمنستان	۳۸۸۸۱	۵۸/۴۰	۳۷/۶۳	۱۴۸۸	۱۸۹۹	-	-
بصره	عراق	۴۰۶۸۹	۴۷/۷۰	۳۰/۴۰	۲	۱۹۲۱	-	-
بحرین	بحرین	۴۱۱۵۰	۵۰/۶۵	۲۶/۲۷	۲	۱۹۰۲	-	-
شارجه	امارات	۴۱۱۹۶	۵۵/۵۰	۲۵/۳۰	۳۴	۱۹۳۳	-	



شکل۱- موقعیت ایستگاه‌های پاسخو پیشگو: مشهد، جاسک و بوشه (با دایره‌های سیاه) و ایستگاه‌های منتخب به عنوان متغیر مستقل برای ترمیم دمای ماهانه (با دایره‌های قرمز)

داده‌های گمشده یکی از مشکلات جدی در علوم مختلف به ویژه در آب و هواشناسی است. امروزه روش‌های کارآمدی برای رفع مشکل داده‌های گمشده ارائه شده است که به سازوکار[۹] داده‌های گمشده بستگی دارد. کل داده‌ها (y) را می‌توان به دو قسمت گمشده (y_{mis}) و مشاهده‌ای (y_{obs}) تفکیک کرد. قوانین احتمالی بر y، y_{mis} و y_{obs} حاکم است.

داده‌های گمشده از لحاظ سازوکار به سه دسته تقسیم می‌شوند:

۱- گمشدگی کاملاً تصادفی MCAR[۱۰] یعنی توزیع داده‌های گمشده نه وابسته به قسمت مشاهده شده باشد و نه وابسته به قسمت گمشده.

۲- گمشدگی تصادفی MAR[۱۱] یعنی توزیع گمشدگی به قسمت مشاهده شده‌ی داده‌ها وابسته است.

۳- گمشدگی غیرتصادفی MNAR[۱۲] درصورتی‌که توزیع گمشدگی به داده گمشده نیز وابسته باشد (Scheffer،۲۰۰۲).

• رگرسیون

رگرسیون تابعی است که وابستگی یك متغیر (پاسخ) به یک یا چند متغیر (پیشگو) را ارائه می‌دهد. این تابع می‌تواند خطی، غیرخطی، ساده و چندگانه باشد. رگرسیون چندگانه ابزاری سودمند در ترمیم و گسترش داده‌های مفقود است (ارقامی و همکاران، ۱۳۸۰). بررسی باقی‌مانده‌های الگوی رگرسیونی(آسیب‌شناسی)[۱۳] با اینکه یکی از نقاط قوت رگرسیون است، اما عدم بررسی و تایید آنها نتایج را با چالش مواجه می‌کند. رگرسیون امید ریاضی شرطی Y به شرط متغیرهای X_1 تا X_k مطابق رابطه $Y=E(Y|X_1=x_1,\ldots,X_k=x_k)$ است. رابطه (۱) رگرسیون چندگانه خطی را براساس k متغیر پیشگو نشان می‌دهد. β_0 تا β_k پارامترهای الگو هستند که باید با داده‌های در دسترس برآورد شوند. u مولفه خطاست که از توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 پیروی می‌کند. σ^2 نیز باید توسط داده‌ها برآورد شود (رضایی پژند و بزرگ نیا، ۱۳۸۱؛ Ranhao و همکاران، ۲۰۰۸).

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k+u \qquad (۱)$$

پذیره‌های پایه: پذیرش الگوی رگرسیونی خطی به دو مرحله تفکیک می‌شود. مرحله اول برآورد پارامترها، آزمون‌های لازم، تحلیل واریانس، ضریب تعیین و غیره است. در رابطه خطی (۱)، βها ضرایب یا پارامترهای الگو هستند که باید برآورد شوند. برآورد این پارامترها با روش‌هایی چون حداقل مربعات خطا، حداکثر درست‌نمایی و ... انجام می‌شود. روش حداقل مربعات معمول‌تر است. ضرایب الگو باید به گونه‌ای برآورد شوند که مجموع مربعات خطا کمینه شود. پارامتر در الگو وقتی پذیرفته می‌شود که آماره به دست آمده از t جدول بزرگ‌تر یا میزان احتمال (p-value) به اندازه کافی کوچک باشد.

اگر الگو در مرحله اول رد نشد، باید مرحله دوم را که برقراری پذیره‌های زیر است بررسی کرد. این پذیره‌ها در مورد مولفه تصادفی است. قبول نهایی الگو مشروط به برقراری این فرض‌ها است.

۱) میانگین خطاها صفر است ($E(Z_i=0)$).

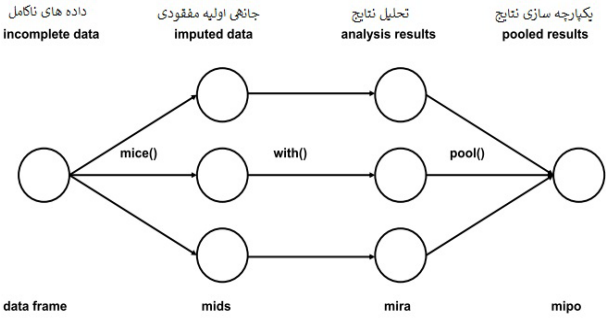
۲) واریانس خطاها ثابت است ($Var(Z_i)=\sigma^2$). مقدار آماره کای-دو در آزمون بروش-پاگان[۱۴] از کای-دو جدول بزرگتر باشد.

۳) خطاها توزیع نرمال دارند ($Z\sim N(0,\sigma^2)$). آزمون شاپیرو-ویلک و نمودار چندکی این فرض را بررسی می‌کند.

۴) توزیع خطاها مستقل است. یعنی ناخودهمبسته مرتبه یک باشد ($Cov(Zi,Zj)=0;i\neq j$). آزمون دوربین-واتسون (D-W) این فرض را بررسی می‌کند. آماره به دست آمده باید در محدوده خاصی قرار بگیرد.
۵) متغیرهای پیشگو همخطی نداشته باشند (عامل تورم واریانس یا آماره VIF[15] کمتر از ۱۰ باشد).
۶) داده پرت از الگوی رگرسیونی تا حد ممکن بررسی و حذف شود. آزمون بون فرنی داده‌های پرت فرین را نشان می‌دهد.
بررسی برقراری این فرض‌ها در خصوص باقی‌مانده‌ها با آزمون و روش‌های توصیفی مناسب صورت می‌گیرد. الگو پس از تایید این موارد رد نمی‌شود (قبول است). (رضایی پژند و بزرگ نیا، ۱۳۸۱).

• الگوریتم MICE

به‌منظور مدیریت مشکل افزایش خطای[16] حاصل از جانهی، روبین (۱۹۸۷) روشی را برای متوسط کردن نتایج در امتداد چندین مجموعه داده جانهی‌شده توسعه داد. الگوریتم چند جانهی MICE 1.0 حالت پیشرفته‌تر EM و MI بوده که در سال ۲۰۰۰ و در Splus نوشته شد. این بسته در سال ۲۰۰۱ به زبان R طراحی شد. سپس نسخه MICE 2.9 در سال ۲۰۱۱ توسط Van Buuren ارائه شد. این بسته نرم‌فزاری از زنجیره مونت کارلو مارکف (MCMC) استفاده می‌کند و نام‌های مختلفی در منابع مختلف دارد. این الگوریتم امروزه در زمینه‌های متنوعی از قبیل اقتصاد، کشاورزی، زیست‌شناسی و نظایر آن به کار می‌رود.
همه‌ی روش‌های جانهی چندگانه از سه گام زیر پیروی می‌کنند:
۱. جانهی: مشابه با جانهی تکی، مقدارهای گم‌شده جانهی می‌شوند. به‌هرحال، مقدارهای جانهی‌شده به‌جای یک بار، mبار از یک توزیع آماری به دست می‌آیند. در پایان این گام، باید m مجموعه داده کامل وجود داشته باشد.
۲. تحلیل: هرکدام از m مجموعه داده تحلیل می‌شود، در پایان این گام بایستی m تحلیل وجود داشته باشد.
۳. تجمیع: m نتیجه با محاسبه میانگین، واریانس و بازه اطمینان متغیر مورد نظر در این گام، در یک نتیجه تلفیق می‌شوند.
یکی از مزیت‌های جانهی چندگانه به نسبت جانهی تکی و روش‌های ساده این است که جانهی چندگانه انعطاف‌پذیر است و می‌تواند در سناریوهای متفاوتی استفاده شود. جانهی چندگانه می‌تواند در مواردی که داده‌ها به کلی یا به تصادف گم‌شده‌اند یا حتی مواردی که به تصادف گم‌نشده‌اند، استفاده شود. در هرصورت، روش اصلی جانهی چندگانه، جانهی چندگانه بوسیله معادلات زنجیری (MICE) است. نکته مهم لازم به‌ذکر این است که MICE تنها زمانی می‌تواند پیاده‌سازی شود که داده‌های گم‌شده از سازوکاری تصادفی پیروی کند.
غفلت از عدم قطعیت در جانهی، ممکن است منجر به خطا در نتیجه‌گیری شود. با چند مرتبه جانهی، جانهی چندگانه مطمئناً منجر به کاهش عدم قطعیت و به‌وجودآمدن بازه‌ای از مقدارهایی که مقدار واقعی شامل آن است می‌شوند.
جانهی چندگانه نیز پیاده‌سازی چندان دشواری ندارد. تعداد زیادی از بسته‌های آماری در نرم‌افزارهای آماری وجود دارند که به راحتی امکان اجرای جانهی چندگانه را ممکن می‌کند. برای مثال بسته MICE به کاربران امکان جانهی به‌روش MICE را در نرم‌افزار R می‌دهد.
بسته نرم‌افزاری MICE 2.9 از نسخه‌های قبلی کامل‌تر است. توابع جدید، تولید خودکار ماتریس پیشگو و موارد دیگر افزوده شده است. سه تابع مهم در این بسته عبارتند از ()mice برای جانهی داده‌های چند سطحی، ()with برای ورود الگو و تحلیل داده‌های کلی در داده‌های جانهی شده، ()pool برای ترکیب سطوح مختلف و ایجاد برترین سری پیشگو. این سه مرحله اصلی با عنوان جاگذاری چندگانه (جانهی اولیه مفقودی)، تجزیه و تحلیل نتایج و یکپارچه‌سازی در شکل (۲) آمده است.



شکل ۲- گام‌های اصلی اجرای الگوریتم MICE

سمت چپ شکل (۲) نشان می‌دهد که تجزیه و تحلیل با مجموعه داده‌های ناقص/ناکامل مشاهده شده Y_{obs} آغاز می‌شود. به طور کل، مسئله این است که ما نمی‌توانیم پارامتر تابع چگالی داده‌ها را از Y_{obs} بدون فرضیه‌های غیرواقعی درباره داده‌های غیرمشاهده‌ای تخمین بزنیم. جانهی چندگانه یک چارچوب کلی است که چندین نسخه منتخب داده را با جایگزین‌کردن مقادیر گمشده با مقادیر داده‌های احتمالی انجام می‌دهد. این مقادیر احتمالی از توزیع خاص اختصاص یافته برای هر درایه گمشده استخراج می‌شوند. تعداد مجموعه داده‌های منتسب اولیه با آرگومان m در تابع mice وارد می‌شود. این آرگومان به طور پیش‌فرض ۵ است. یعنی ۵ مجموعه داده Y_1، Y_2، ... ، Y_5 که برای داده‌های موجود یکسان و برای داده‌های جانهی شده متفاوت است. هر مدل جانهی سه شرط را باید داشته باشد: ۱- قانون‌مندی فرآیندی که داده گمشده را خلق می‌کند. ۲- حفظ وابستگی (ارتباط) بین داده‌ها. ۳- حفظ عدم قطعیت این روابط (Van Buurenو Groothuis، ۲۰۱۱).

نتایج و بحث

داده‌های طولانی مدت و کامل متغیر دما یکی از نیازهای مهم در تحلیل‌های قابل اعتماد در اکثر مطالعات آب و هواشناسی و علوم مرتبط است. آمار بیش از ۱۲۰ سال این متغیر برای سه ایستگاه ایران (مشهد، بوشهر و جاسک) اندازه‌گیری و در کتبی به نام World Weather Recrds ثبت شده است. این اطلاعات ماهانه بوده و در بازه‌هایی (متأثر از جنگ‌های جهانی اول و دوم) دارای داده گمشده است. هدف این پژوهش جمع‌آوری و ترمیم این آمار طولانی مدت است. زیرا نتایج تحلیل آمار کوتاه (کمتر از ۱۰۰ سال) یا ناقص از دیدگاه علم آمار اریب است. یعنی صحت نتایج زیر سوال است. این تحقیق در چند مرحله مطابق زیر انجام شده است.

ابتدا الگوهای چندگانه رگرسیونی به داده‌ها برازش داده شد. این‌کار مستلزم داشتن ایستگاه‌هایی با اطلاعات کامل به عنوان متغیر مستقل (توضیحی، پیشگو یا مبنا) در الگو است. با توجه به اینکه چنین ایستگاه‌هایی در ایران موجود نیست، لذا کلیه ایستگاه‌های کشورهای همسایه که دارای داده‌های قدیمی و همزمان با ایستگاه‌های پاسخ (وابسته) هستند انتخاب و بررسی شدند. ایستگاه‌ها بر اساس سه معیار انتخاب شدند. ۱- همبستگی بالا ۲- فاصله کم ۳- داشتن اطلاعات در خلأهای ایستگاه پاسخ. مشخصات ایستگاه‌های پاسخ و پیشگو و درصد مفقودی‌های سه ایستگاه ایران در جدول (۱) آمده است.

- الگوهای رگرسیونی

الگوهای رگرسیونی مختلف به بارش سه ایستگاه مشهد، جاسک و بوشهر برازش داده شد.

برای این منظور دمای ایستگاه‌های عشق‌آباد، بایرام علی، گودان، سرخس و تجن به عنوان متغیر مستقل در ساخت داده‌های گمشده دمای ماهانه مشهد بررسی شدند. سه عاملِ فاصله تا ایستگاه مشهد، همبستگی و وجود داده در ماه‌های گمشده در انتخاب این ایستگاه‌ها موثر بودند. دو ایستگاه سرخس و گودان در ترمیم دمای این ایستگاه انتخاب شدند. با توجه به همبستگی بالای دمای ایستگاه‌ها و در نتیجه هم خطی متغیرهای ورودی (همبستگی ایستگاه‌های پیشگو)، الگوهای خطی چند متغیره رد شدند و متغیرهای پیشگو به صورت تبدیل شده (لگاریتمی، توانی و ...) به الگوها وارد می‌شوند. به همین ترتیب برای ساخت داده‌های گمشده در دو ایستگاه بوشهر و جاسک نیز الگوهای مختلف بررسی شد. الگوهای مختلف بررسی و نتایج نهایی الگوهای دمای مشهد، بوشهر و جاسک به ترتیب در روابط با شماره (۲) تا (۴) به عنوان بهترین الگو انتخاب شدند.

متغیرهای دمای ماهانه ایستگاه‌های به کار رفته در این سه الگو به ترتیب با علائم اختصاری T_{Mas}(دمای مشهد)، T_{Ser} (دمای سرخس)، T_{Gud}(دمای گودان)، T_{Bus}(دمای بوشهر)، T_{Bah}(دمای بحرین)، T_{Bas} (دمای بصره)، T_{Jas}(دمای جاسک) و T_{Shar}(دمای شارجه)، تعریف می‌کنیم.

$$T_{Mas}=\beta_0+\beta_1 T_{Ser}+\beta_3 \log(T_{Gud}+6)+\varepsilon \quad (۲)$$

$$T_{Bus}=\beta_0+\beta_1 T_{Bah}+\beta_3 (T_{Bas}^{-2}) \quad (۳)$$

$$T_{Jas}=\beta_0+\beta_1 T_{Bah}+\beta_3 exp(T_{Shar}) \quad (۴)$$

نتایج حاصل از برازش الگوهای نهایی دمای سه ایستگاه مورد بررسی، شامل برآورد ضرایب و خطای معیار و آماره t هر ضریب، عامل تورم واریانس (VIF) و برخی آزمون‌های آسیب‌شناسی در جدول (۲) آمده است. ضریب تعیین و آماره F قدرت الگو را نشان می‌دهند.

جدول ۲- نتایج حاصل از برازش الگوهای نهایی دمای سه ایستگاه ایران شامل برآورد ضرایب و عامل تورم واریانس و آسیب‌شناسی

ایستگاه	ضرایب	متغیر پیشگو	برآورد ضریب	آماره t	همخطی (VIF)	ضریب تعیین (R^2_{adj})	پایایی واریانس	جذر مربع خطا (RMSE)	آماره F	دوربین-واتسون (D-W)	آزمون Shapiro	آماره کوک
مشهد	β_0	ثابت	-۴/۸۷	-۱۵/۶۲	-	۰/۹۹	۱/۸۲	۰/۷۱	۳۷۰۰۰	۱/۶۶	۰/۹۹۷	۰/۰۰۱۶
	β_1	سرخس	۰/۶۹	۷۰/۵۰	۷/۵							
	β_2	گودان	۲/۷۴	۱۶/۱۵	۷/۵							
بوشهر	β_0	ثابت	-۱/۸۲	-۴/۵۹	-	۰/۹۵۱	۰/۰۴۸	۰/۸۱	۵۱۵۰	۰/۷۷	۰/۹۹	۰/۰۵۴
	β_1	بحرین	۱/۰۱	۸۴/۵۴	۴/۵							
	β_2	بصره	-۱۸۳/۴۵	-۵/۳۵	۴/۵							
جاسک	β_0	ثابت	۹/۴۱	۱/۲۵	-	۰/۹۸۵	۷/۶۳	۰/۹۴	۱۸۲۱۰	۰/۹۷	۰/۹۹۷	۰/۰۳۹
	β_1	بحرین	۰/۶۸	۷/۷۱	۱/۸							
	β_2	شارجه	۷/۶×۱۶-۱۰	۴۲/۹۸	۱/۸							

آزمون‌های مناسب آسیب‌شناسی الگوها برای بررسی فرض‌های زیربنایی رگرسیون الگوهای دما انجام شد.

پایایی واریانس: مقدار آماره کای-دو در آزمون بروش-پاگن (دستور NCV test[18] از بسته نرم‌افزاری car در نرم‌افزار R) نشان می‌دهد که واریانس باقیمانده‌ها در الگوی ایستگاه مشهد پایا هستند (ستون هشتم جدول ۲). (مقدار آماره به دست آمده بیش از کای ۲ جدول و مقداراحتمال نزدیک به صفر است). نمودارها نیز تثبیت واریانس را به تقریب تأیید کرد (مقادیر پیش‌بینی شده در برابر باقیمانده‌های استاندارد شده به صورت تقریباً مستطیلی توزیع شده‌اند). در مورد الگوهای بوشهر و جاسک پایایی واریانس رد می‌شود.

استقلال باقیمانده‌ها: آماره آزمون دوربین واتسون باید در محدوده ناهمبسته بودن مقادیر جدول دوربین واتسون (۲/۲۶-۱/۷۴) باشد. بنابراین استقلال باقیمانده‌ها در مشهد مورد تایید است و در بوشهر و جاسک استقلال باقیمانده‌ها رد می‌شود.

داده پرت: مقدار کم آماره میانگین فاصله کوک عدم وجود داده پرت را نشان می‌دهد. این آماره کمتر از ۱ باشد مورد تایید است.

نرمال‌بودن باقیمانده‌ها: بررسی نرمال‌بودن باقیمانده‌ها با آزمون شاپیرو و نمودار چندکی انجام می‌شود. نتایج نشان‌دهنده نرمال‌بودن تقریبی باقیمانده‌هاست. به دلیل حجم بالا از آوردن نمودارها صرف نظر و فقط به نتایج حاصل اکتفا شد.

هر کدام از فرض‌های زیربنایی فوق که رد شود نتایج الگوهای رگرسیونی را با تردید مواجه خواهد کرد و نمی‌توان به نتایج حاصل از روش کلاسیک رگرسیون اعتماد نمود.

- برازش الگوریتم MICE برای ترمیم دمای ماهانه سه ایستگاه

الگوریتم MICE به عنوان الگوریتم برگزیده از خانواده MI و EM در برآورد داده گمشده ماهانه دمای سه ایستگاه ایران استفاده شد. این الگوریتم به عنوان یک روش چند جاگذاری برای حل مشکل داده‌های گمشده طراحی شده است. ابتدا پنج ورودی (به عنوان پیش فرض) از داده‌های کامل به الگوریتم وارد شده و سپس این پنج ورودی تا همگرایی و رسیدن به نتیجه مطلوب (بیشترین شباهت به قانون احتمالی داده‌ها) تکرار می‌شود. معدل برآوردهای این پنج روش به عنوان خروجی ارائه می‌شود. دوباره این کار تکرار شده تا کمترین خطا در برآورد به دست آید که همان نتیجه نهایی برآورد داده‌های گمشده است (Van Buuren، ۲۰۱۸). یکی از دلایل انتخاب این الگوریتم استفاده از روش‌های مختلف برای ورودی‌های اولیه است. این تحقیق ورودی‌های مختلفی را به نرم‌افزار اعمال و بهترین را انتخاب کرده است. یکی از فرضیات در این پژوهش اینست که آیا ورود اطلاعات اولیه دقیق توسط کاربر منجر به نتایج بهتر می‌شود؟ معیار خطای RMSE به منظور مقایسه روش‌های مختلف به کار رفت. بسته MICE 2.9 در نرم‌افزار R جاگذاری‌های چندگانه داده‌های گمشده را تحلیل و ارائه می‌کند (Van Buuren و Groothuis، ۲۰۱۱). ایستگاه‌های مجاور در این روش نیز می‌توانند به عنوان متغیرهای مستقل به پیش‌بینی دقیق‌تر کمک کنند. این الگوریتم شامل سه مرحله اصلی جاگذاری چندگانه، تجزیه و تحلیل نتایج و یکپارچه‌سازی است (شکل ۲). این الگوریتم نسبت به ورودی‌های اولیه (جاگذاری داده‌های کامل اولیه) حساس است، بنابراین این نتایج نیز مقایسه و بهترین پیش‌بینی در این مرحله ارائه می‌شود. مزیت دیگر این روش اینست که علاوه بر کامل‌کردن متغیر پاسخ، داده‌های گمشده متغیرهای کمکی (مستقل) را ترمیم و ارائه می‌کند.

برازش الگوریتم MICE نیازمند ورود پیش‌فرض‌هایی است. تعداد سری‌های کامل اولیه معمولاً ۵ (مقدار آرگومان m) انتخاب می‌شوند. البته تعداد بیشتر نیز (۱۰ و ۲۰) آزمون شد که در بیشتر موارد نتایج بهتر نبود. تعداد بیشترین تکرار در الگوریتم که در تابع با آرگومان "maxit" مشخص می‌شود را به روش سعی و خطا ۵۰ انتخاب می‌کنیم. پیش فرض این آرگومان ۵ است. داده‌های کامل اولیه باید با یک روش ورودی ساخته شود. این روش در آرگومان "meth" مشخص می‌شود. پیش‌فرض روش "pmm" یا میانگین‌گیری تطبیق شده است. روش‌های مختلفی وجود دارد که همه آزمون و بهترین روش ساخت اولیه داده‌ها روش "norm.predict" یعنی رگرسیون خطی انتخاب می‌شود. همچنین می‌توانیم داده‌های اولیه را خودمان با استفاده از آرگومان "data.init" وارد کنیم. بقیه موارد مطابق پیش‌فرض انتخاب می‌شود. این روش شبیه رگرسیون از متغیرهای پیشگو استفاده می‌کند.

نتایج برازش این الگوریتم برای دمای ماهانه مشهد به تفصیل بیان می‌شود. پس از آزمون و خطا دمای ماهانه عشق‌آباد، بایرامعلی و گودان به عنوان متغیرهای کمکی در برازش الگوریتم MICE به منظور ترمیم دمای ماهانه مشهد به کار گرفته شدند. طی انجام آزمون‌های متعدد مشخص شد تعداد ورودی‌ها با m=5 و روش تکمیل داده‌های ورودی "norm.predict" بهترین نتیجه را داد. نمودارهای حاشیه‌ای این سه ایستگاه نسبت به مشهد در شکل (۳) آمده است. این نمودار نشان می‌دهد توزیع حاشیه‌ای مقادیر مشاهده‌ای و گمشده شباهت زیادی دارد؛ بنابراین گمشدگی‌ها را می‌توان کاملاً تصادفی (MCAR) فرض کرد. نمودار پراکنش همچنین نشان‌دهنده همبستگی قوی متغیر وابسته با متغیرهای پیشگو است. نتیجه نهایی برازش الگو در جدول (۲) آمده است. این روش یک بازه‌ای (با اطمینان ۹۵٪) برای ضریب تعیین و جذر میانگین خطا در نظر می‌گیرد. که در قسمت پایین جدول آمده است.

شکل (۳) نمودار حاشیه‌ای دمای ماهانه مشهد در مقابل متغیرهای پیشگو (دمای عشق‌آباد، بایرامعلی و گودان) است. موقعیت داده‌های مشاهده شده با رنگ آبی و گمشده‌ها با قرمز نشان داده شده است.

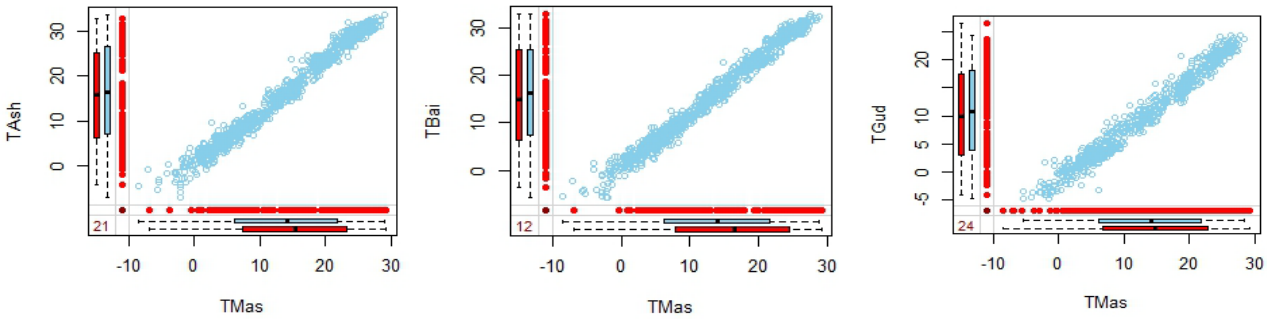
نمودار آسیب‌شناسی (شکل ۴)، توزیع باقیمانده‌های مقادیر

مشاهده‌ای (منحنی آبی) و پنج سری تولید شده توسط مدل با مقادیر جانهی شده (منحنی قرمز) را نشان می‌دهد. اولین نمودار (بالا چپ) مربوط به ایستگاه مشهد است که تطابق خوبی دارد؛ لذا نشان از مناسب‌بودن مدل است. تطابق در بقیه ایستگاه‌ها کمتر است مخصوصاً ایستگاه بایرامعلی که البته خروجی آنها مدنظر نیست و خللی به نتایج وارد نمی‌کند.

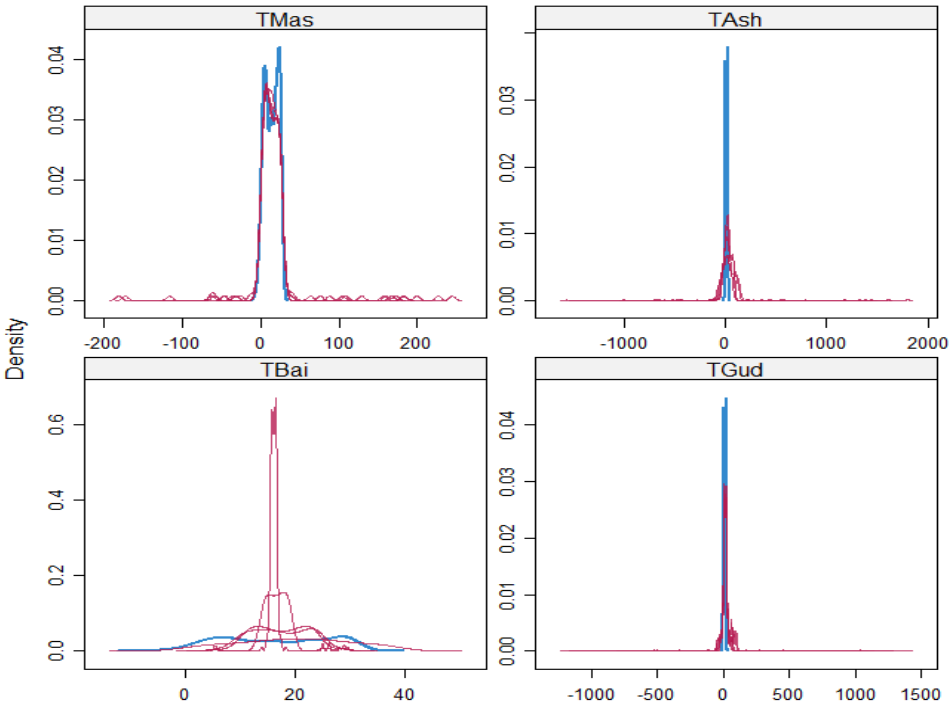
جدول ۲- نتایج اجرای تابع mice برای ترمیم دمای ماهانه مشهد

متغیر	برآورد	خطای استاندارد	آماره	درجه آزادی	p-مقدار
ضریب ثابت	۰/۴۲	۰/۰۵	۷/۸۰	۱۱۱۴/۸۵	۰/۰
TAsh	-۰/۱۶	۰/۰۶	-۲/۶۴	۴/۱۶	۰/۰۱
TBai	۰/۷۰	۰/۰۶	۱۱/۳۵	۱/۷۶	۰/۰
TGud	۰/۴۳	۰/۰۲	۲۲/۸۹	۱۰۴/۶۴	۰/۰

	کران پایین	مقدار متوسط	کران بالا
R۲	۰/۹۸۸	۰/۹۹۱	۰/۹۹۴
RMSE	۰/۳۹۱	۰/۵۱۵	۰/۶۸۰



شکل ۳- نمودار حاشیه‌ای دمای ماهانه مشهد در مقابل متغیرهای پیشگو (دمای عشق آباد، بایرامعلی و گودان)



شکل ۴- توزیع باقیمانده‌های مدل MICE برای دمای ماهانه مشهد (بالا-چپ) و ایستگاه‌های پیشگو

مقادیر مشاهده‌ای (منحنی آبی) و پنج سری تولید شده توسط مدل یا مقادیر جانهی شده (منحنی قرمز)

گمشده‌های ایستگاه جاسک بعد از ۱۹۵۰ است. بنابراین ایستگاه‌های داخلی از قبیل بندرعباس، زاهدان و کرمان نیز در ساخت الگوها بررسی شد. ترمیم دمای ایستگاه جاسک با دمای دو ایستگاه بحرین و شارجه با روش ورودی "norm.predict" با تعداد ورودی‌های اولیه (m=5) به منظور الگوسازی با الگوریتم MICE به کار رفت. پیش‌فرض‌های الگوریتم و نتایج ضریب تعیین و خطای RMSE ترمیم سه ایستگاه در جدول (۳) آمده است. نتایج RMSE الگوهای رگرسیونی نیز در ستون آخر جدول آمده است. با توجه به دوره گمشده موجود برای ساخت دمای جاسک در دوره ۱۹۵۷ تا ۱۹۶۷ از الگوی دیگری نیز استفاده شد. الگوی MICE با توجه به نتایج به دست آمده انتخاب شد. بهترین الگو از روی ایستگاه‌های مبنای بحرین، جیوانی پاکستان، کرمان و زاهدان به دست آمد. خطای RMSE در بهترین حالت ۰/۴۷ سانتیگراد بود که نشان‌دهنده دقت بالای این روش است.

دمای ماهانه بوشهر با دمای دو ایستگاه بحرین و بصره و به کمک الگوریتم MICE ترمیم شد. نتایج نشان داد روش ورودی "norm.predict" به دلیل ایجاد همخطی بالا برای این ایستگاه مناسب نیست. بنابراین روش‌های مختلف بررسی و روش "midastouch" یعنی میانگین وزنی به عنوان بهترین روش ورودی برای ایجاد داده‌های کامل اولیه انتخاب شد. همچنین تعداد جاگذاری‌های متفاوت یعنی آرگومان m به ازای ۵ و ۲۰ بررسی و مشخص شد تعداد m=۲۰ خطا را کمی کاهش داده بنابراین بهتر است.

جدول ۳- نتایج حاصل از برازش الگوریتم MICE به دمای ماهانه سه ایستگاه ایران

معیار	m	روش ورود اطلاعات	R۲			RMSEmice			RMSEreg
نام ایستگاه			کران پایین	متوسط	کران بالا	کران پایین	متوسط	کران بالا	متوسط
دمای مشهد	۵	norm.predict	۰/۹۹۴	۰/۹۹۷	۰/۹۹۸	۰/۳۹۱	۰/۵۱۵	۰/۶۸۰	۰/۷۱
دمای جاسک	۵	norm.predict	۰/۹۵۸	۰/۹۶۲	۰/۹۶۶	۰/۸۲۲	۰/۸۶۴	۰/۹۰۸	۰/۹۴
دمای بوشهر	۲۰	midastouch	۰/۹۸۴	۰/۹۸۶	۰/۹۸۸	۰/۷۲۷	۰/۷۸۱	۰/۸۳۸	۰/۸۱

نتیجه اینکه روش MICE دقت بالاتری دارد. دلایل دیگری نیز برای برتری و انتخاب روش MICE برای برآورد داده مفقود وجود دارد که در اینجا بیان می‌شود:

۱- روش رگرسیون به‌عنوان یک روش کلاسیک نیازمند بررسی پذیره‌های پایه است. اگر یک یا چند فرض پایه برقرار نباشد رگرسیون از نظر علمی رد می‌شود. همانطور که مشاهده شد برخی از فرض‌ها در اینجا برقرار نبود. در حالیکه روش MICE به این بررسی‌ها نیاز ندارد.

۲- روش MICE توانایی برآورد داده‌های مفقود ایستگاه‌های پیشگو را نیز دارند.

۳- اگر یک یا چند متغیر ورودی فاقد داده باشد در روش رگرسیون قادر به برآورد داده‌ها با یک الگو نخواهیم بود؛ اما در روش MICE از روابط درون داده‌ها استفاده کرده و همه مفقودی‌ها را برآورد می‌کند.

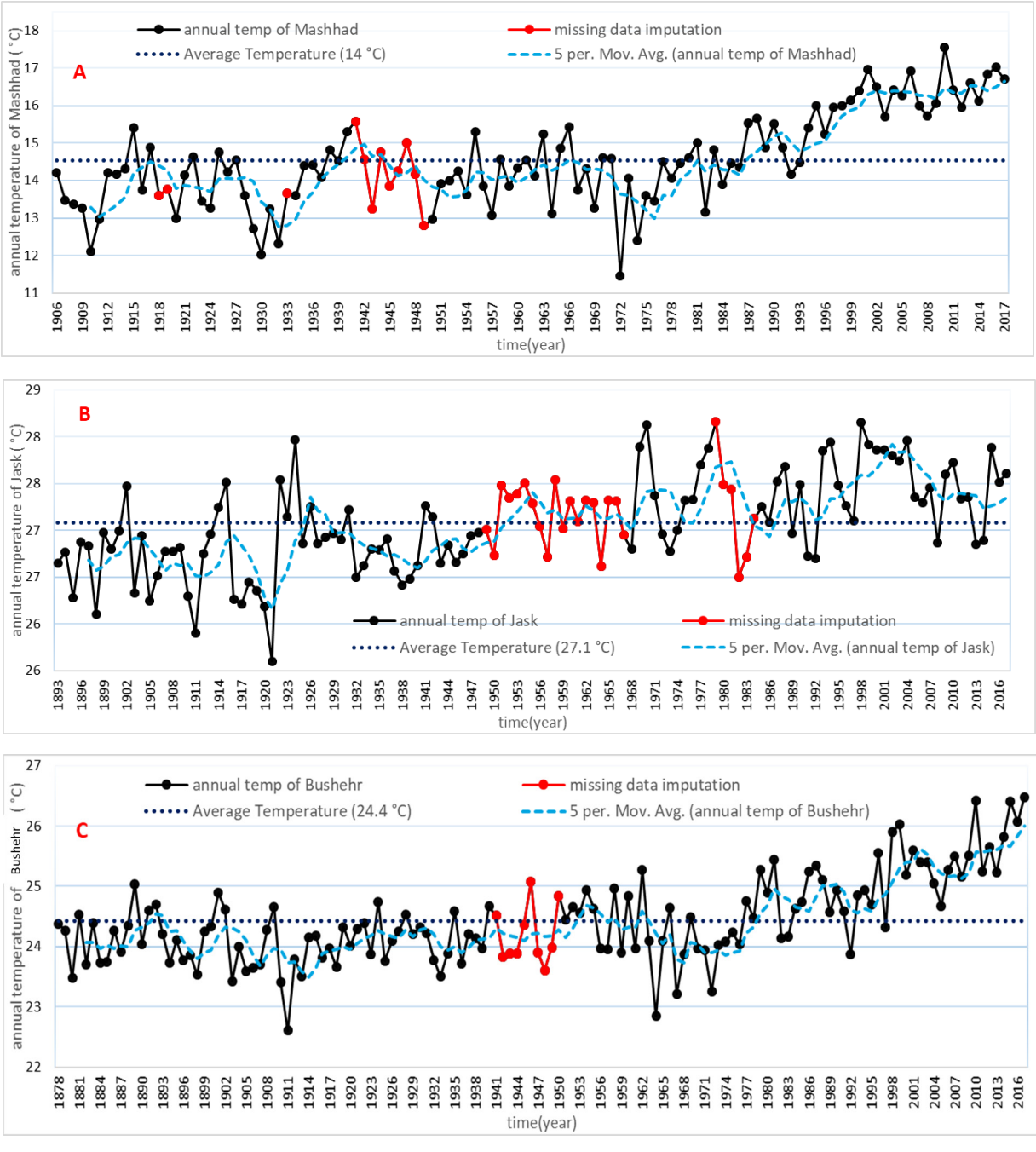
۴- بسته نرم‌افزاری MICE سری تکمیل شده متغیرهای پاسخ و پیشگو را تولید و فقط برآورد داده‌های مفقود را جاگذاری می‌کند. در حالیکه در روش رگرسیون کل داده‌ها با الگو تولید و سپس باید داده‌های برآوردی را در سری داده‌های موجود جاگذاری نمود.

۵-روش MICE بسیار سریع اجرا می‌شود. همچنین بندهای فوق خصوصاً ۳ و ۴ باعث صرفه‌جویی در زمان و هزینه می‌شود.

۶- روش MICE سازوکار داده‌های گمشده را در نظر می‌گیرد. این روش برای داده مفقود طراحی شده است و توزیع داده‌های موجود و مفقود را ارائه می‌دهد. منابع مختلف این روش را برای جاگذاری داده مفقود پیشنهاد داده‌اند. Deng و همکاران (۲۰۱۶) برتری روش MICE بر رگرسیون را با چند مثال تأیید کردند.

MICE یک روش اصولی و در عین حال انعطاف‌پذیر برای پرداختن به داده‌های از دست رفته ارائه می‌دهد که برای طیف گسترده‌ای از کاربران قابل دسترسی است (Melissa و همکاران، ۲۰۰۱).

با توجه به عملکرد خوب این روش در افزایش دقت برآورد دمای ماهانه ایستگاه مشهد جاگذاری[19] مفقودی‌های دما به روش MICE انجام و سری زمانی دراز مدت آن در شکل (۵) آمده است. مفقودی‌های برآورد شده با رنگ قرمز نمایش داده شده است.



شکل ۵- سری‌های زمانی کامل شده دمای سالانه سه ایستگاه ایران.

نقاط سیاه مربوط به بارش سالانه دراز مدت، نقاط قرمز مربوط به جاگذاری داده‌های گمشده با روش MICE، میانگین متحرک ۵ ساله با خط چین آبی و میانگین داده‌ها با خط‌چین سرمه‌ای نشان داده شده است. به ترتیب از بالا به پایین ایستگاه مشهد(A)، جاسک (B) و بوشهر (C)

نتیجه‌گیری

ترمیم و برآورد مفقودی‌های دمای ماهانه طولانی مدت مشهد، جاسک و بوشهر و معرفی یک روش کارامد برآورد داده مفقود، هدف این مقاله است. انتساب مقادیر مفقود متغیرهای هواشناسی با کمترین خطای ممکن همواره اهمیت بالایی داشته است. ایستگاه‌هایی از کشورهای مجاور به‌عنوان ایستگاه‌های مبنا انتخاب شدند. ترمیم داده‌های دما ابتدا با بهترین الگوهای رگرسیونی انجام شد (ریشه میانگین مربعات خطا ۰/۷۱تا ۰/۹۴). علیرغم استفاده از الگوهای غیرخطی، بررسی فرض‌های زیربنایی رگرسیون نشان از عدم تایید برخی از آنها مانند استقلال باقی‌مانده داشت. برآورد داده‌های مفقود با روش MICE نیز انجام شد (ریشه میانگین مربعات خطا ۰/۳۹تا ۰/۸۲). نتایج نشان داد به دلیل مزایای زیاد و دقتی که روش MICE دارد، نسبت به رگرسیون بهتر است. شایان ذکر است این روش برای برآورد و جانهی داده‌های مفقود طراحی شده، مشکلات رگرسیون را نداشته و قابلیت‌های زیادی دارد. لذا روش MICE برای ترمیم داده‌های گمشده آب و هواشناسی توصیه می‌شود. جمع آوری و اصلاح این آمار طولانی مدت و ارائه دمای ماهانه کامل آنها برای اولین بار در ایران می‌تواند مبنای ارزشمندی برای تحقیقات آب و هواشناسی باشد.

پی‌نوشت

1-Missing Data
2-Expectation–Maximization
3-Multiple Imputation
4-Multivariate Imputation by Chained Equations
5-Normal Ratio
6-Artificial Neural Network
7-Mont Carlo Markov Chain
8-World Weather Records
9-Mechanism
10-Missing Completely At Random
11-Missing At Random
12-Missing Not At Random
13-Diagnostic
14-Breusch–Pagan test
15-Variance Inflation factor
16-noise
17-History data
18-imputation

منابع

ارقامی، ن.ر.، سنجری، ن. و بزرگ‌نیا، الف. ۱۳۸۰. مقدمه‌ای بر بررسی‌های نمونه‌ای. چاپ چهارم، انتشارات دانشگاه فردوسی مشهد.

خلیلی، ع. و بذرافشان، ج. ۱۳۸۷. ارزیابی مخاطره تداوم خشك‌سالی با استفاده از داده‌های بارندگی سالانه قرن گذشته در ایستگاه‌های قدیمی ایران. مجله ژئوفیزیك ایران، ۲(۲): ۱۳- ۲۳.

رضایی پژند، ح. و بزرگ نیا. الف. ۱۳۸۱. تحلیل رگرسیون غیرخطی و کاربردهای آن. انتشارات دانشگاه فردوسی مشهد.

فرزندی، م.، رضایی پژند، ح. و ثنائی نژاد، ح. ۱۳۹۳. ترمیم و گسترش ۱۲۷ سال دمای ماهانه مشهد. مجله پژوهش‌های اقلیم‌شناسی، ۵(۱۷): ۱۱۱- ۱۲۳.

Deng Y, Chang C, Ido MS, Long Q. 2016, Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. Sci Rep. 6:21689 .

Edmond F.S., Victor A.K. and Khalid M. 1973. Floods and droughts, Water Resources Publications. Proceedings of the Second International Symposium in Hydrology, 679 pages.

Ghahraman B. and Ahmadi F. 2007. Applica tion of geo statistics in time series: Mashhad Annual Rainfall. Iran-Watershed Management Science & Engineering, 1(1):7-15.

Iqbal M., Wen J., Wang Sh., Tian Hu. and Adnan M. 2018. Variations of precipitation characteristics during the period 1960-2014 in the Source Region of the Yellow River, China. Journal of Arid Land, 10(3): 388-401.

Jacob D., Reed D.W. and Robson A.J. 1999. Choosing a pooling group. Flood Estimation Handbook. Vol. 3. Institute of Hydrology, Wallingford, UK.

Little R.JA. and Rubin D.B. 2002. Statistical analysis with missing data. John Wiley & Sons.

Melissa J. A, Elizabeth A. S., Constantine F., and Philip J. L., 2011, Multiple imputation by chained

1921-1930, Smithson. Miss C. Collect. pp 639.

Smithsonian Institution. 1947. World weather records, 1931-1940, Smithson. Miss C. Collect. pp 666

Smithsonian Institution. 1927. World weather records, 1750-1920, Smithson. Miss C. Collect. pp 1199.

Van Buuren S. 2018. Flexible Imputation of Missing Data. 2nd. Chapman & Hall/CRC Interdisciplinary Statistics.

Van Buuren S. and Groothuis-Oudshoorn K. 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3): 1-67.

Yozgatligil C., Aslan S., Iyigun C. and Batmaz I. 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theory Apply Climatology, 112: 143–167.

equations: what is it and how does it work?, Int J Methods Psychiatr Res. 20(1): 40–49.

Porto de Carvalho J.R, Boffinho Almeida Monteir, J.E., Nakai, A.M., Assad E.D., 2017, Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults. Revista Brasileira de Meteorologia, 32(4): 575-583.

Ranhao, S., Baiping, Z., and Jing, T., 2008, A Multivariate Regression Model for Predicting Precipitation in the Daqing Mountains, Mountain Research and Development, 28(3):318-325.

Scheffer J. 2002. Dealing with missing data. Research Letters in the Information and Mathematical Sciences, 3:153-160.

Smithsonian Institution. 1934. World weather records,