نشریه علوم و مهندسی
آب و فاضلاب
سال دوم، شماره ۴، زمستان ۱۳۹۶

علوم و مهندسی
آب و فاضلاب

Journal of Water & Wastewater
Science & Engineering (jwwse)
Winter 2018, Vol. 2, No. 4.
DOI: 10.22112/jwwse.2018.87485.1002

تخمین اکسیژن‌خواهی شیمیایی خروجی از فاضلاب صنعت پتروشیمی با استفاده از رگرسیون مقاوم

میلاد ابوذری[۱]، پرهام پهلوانی[۲*] و غلامرضا نبی بیدهندی[۳]

۱- کارشناسی ارشد مهندسی محیط زیست – آب و فاضلاب، گروه مهندسی محیط‌زیست، دانشگاه تهران، تهران، ایران
۲- استادیار، دانشکده مهندسی نقشه‌برداری و اطلاعات مکانی، پردیس دانشکده‌های فنی، دانشگاه تهران، تهران، ایران
۳- استاد، دانشکده محیط‌زیست، گروه مهندسی محیط‌زیست، دانشگاه تهران، تهران، ایران
* نویسنده مسئول، ایمیل: pahlavani@ut.ac.ir

تاریخ دریافت: ۱۳۹۶/۰۳/۷
تاریخ اصلاح: ۱۳۹۶/۱۲/۲۴
تاریخ پذیرش: ۱۳۹۷/۰۲/۱۶

چکیده

در راستای افزایش بهبود کیفیت پساب صنعتی و مدیریت بهتر آن‌ها، باید راه‌کاری ساده و با دقت مناسب برای تخمین فرآیندها ایجاد نمود. با توجه به این‌که فرآیندهای تصفیه به‌صورت سیستم جعبه سیاه[۱] می‌باشند و به دلیل تأثیرپذیری اکثر عوامل دخیل در سیستم و مشکلات جمع‌آوری داده در مدل‌های فیزیکی، استفاده از آمار و روش‌های رگرسیونی می‌تواند بسیار اثربخش باشد. بنابراین هرچه مدل ساده‌تر و با متغیرهای ورودی کمتری باشد مدل مربوطه اهمیت بیشتری خواهد داشت. ورودی مدل پیشنهادی شامل داده‌های خروجی واحد بیولوژیکی و پارامتر خروجی مدل، میزان اکسیژن‌خواهی شیمیایی[۲] واحد زلال‌ساز می‌باشد. همچنین برای مقایسه کارایی مدل‌ها از ضریب تبیین[۳]، میانگین مجموع خطاها[۴] و ضریب همبستگی[۵] استفاده می‌شود. هدف این تحقیق ارائه یک مدل داده‌مبنا و بهبود آن و سپس مقایسه آن با روش‌های مشابه است. در این تحقیق داده‌های ورودی و خروجی به‌صورت منظم و با مقیاس واقعی در طول ۱۳ ماه هستند. در نهایت یک رابطه خطی با مدل رگرسیون مقاوم۶ با شاخص‌های ماکزیمم مربعات خطای برابر ۰/۰۸۹، ضریب همبستگی برابر ۰/۷۸۴ و ضریب تبیین ۰/۶۰۹۶ ارائه شده است.

کلمات کلیدی: تصفیه‌خانه فاضلاب پتروشیمی فجر، مدل‌سازی خطی، رگرسیون مقاوم، اکسیژن‌خواهی شیمیایی.

Chemical Oxygen Demand (COD) Estimation in Petrochemical Industry Wastewater Effluent via Robusted Regression

Milad Abuzari[1], Parham Pahlavani[2*] and Gholamreza Nabi Bidhendi[3]

1- M.Sc., Faculty of Environmental Engineering, University of Tehran, Tehran, Iran
2- Assistant Professor, Faculty of Surveying Engineering, University of Tehran, Tehran, Iran
3- Professor, Faculty of Environmental Engineering, University of Tehran, Tehran, Iran
* Corresponding Author, Email: pahlavani@ut.ac.ir

Received: 28/5/2017
Revised: 15/3/2018
Accepted: 6/5/2018

Abstract

In order to increase the quality of industrial wastewater treatment and better manage of them, their approach should be simple and accurate for estimating process. Treatment processes for black box systems are due to the influence of many factors that involved in the system. Because of problems in using physical models, the use of statistics and regression methods could be helpful. Therefore, whatever model is simpler and less input variables so the model will be more important. Influent of the proposed model includes output data of biological unit and effluent is chemical oxygen demand of the clarifier. To compare the models performance three indicators of R-square, Correlation Coefficient(R) and Mean Square Error (MSE) are used. The aim of this study is creating linear data mining model and comparing them with similar methods for quality data. Finally, a linear robust regression with MSE = 0.089054, R = 0.784727 and R-Square = 0.6096 is proposed.

Keywords: Chemical Oxygen demand (COD), Fajr Petrochemical WWTP, Linear Models, Robust Regression.

۱- مقدمه

از مزایای عمده روش‌های داده‌مبنا، عدم نیاز به درک پیچیده از فیزیک و ریاضی در استفاده از آن‌ها است که این امر، احتمال خطای محاسباتی یا مفهومی را کاهش می‌دهد (Xiupeng and Andrew, 2015). یکی از ابزارهایی که در دهه اخیر به دلیل وفور جمع‌آوری داده‌ها، بسیار مورد استفاده قرار می‌گیرد روش‌های محاسبات هوشی[۷] می‌باشد. اگر بتوان مدلی کارآمد و مناسب برای تخمین پارامترهای خروجی هر واحد تصفیه‌خانه فاضلاب یافت در نتیجه با توجه به انواع فرایندهای فیزیکی و شیمیایی و بیولوژیکی، بهترین تصمیم‌گیری‌های فنی برای بهره‌برداری بهتر به‌کار گرفته خواهد شد. در این تحقیق تنها ابزار، داده‌های گذشته واحد مربوطه است که به با توجه به شرایط واقعی[۸] توسط ابزارهای اتوماسیونی یا روش‌های آزمایشگاهی و یا استفاده از ابزاردقیق و به‌صورت ۸ ساعتی و به مدت ۱۳ ماه اندازه‌گیری شده‌اند.

مدل‌های ریاضی در تخمین سیستم‌های فاضلاب از سال ۱۹۸۲ برای اولین بار توسط انجمن بین‌المللی تحقیقات و کنترل آلودگی آب (IAWPRC) ارائه شد (Ting Sie Chun and Malek, 2017) و از همان دهه مدل‌های تخمینی با مطالعات مشابهی توسط دانشمندان مختلف ارائه شد. همچنین با توجه به اهمیت تخمین پارامتر اکسیژن‌خواهی شیمیایی در سیستم‌های تصفیه پساب و نقش آن در بررسی عملکرد سیستم، مقاله‌های متعددی درخصوص تخمین این پارامتر در سیستم‌های آبی و فاضلابی با استفاده از روش‌های خطی کمترین مربعات و در مقایسه با روش‌هایی از قبیل شبکه‌های عصبی و ماشین‌های بردار پشتیبان وجود دارد.

Brayson et al. (2001) میزان بارش و رواناب‌های ناشی از آن با استفاده از الگوریتم مونت‌کارلو با استفاده از ۸ متغیر ورودی و با استفاده از دو نمونه با رواناب کم و زیاد را مطالعه کرده و در نهایت به این نتیجه رسیدند که استفاده از این الگوریتم در ترکیب با مدل‌های دیگر می‌تواند کیفیت برازش را بهبود بخشد.Curlin (2008) در تحقیق خود با استفاده از MLR و روش کمترین مربعات و رگرسیون تکه‌ای[۹] به تخمین میزان اکسیژن‌خواهی شیمیایی در تصفیه‌خانه فاضلاب واقع در ایالت زاگرب کرواسی پرداختند و داده‌های آن‌ها به صورت روزانه جمع‌اوری می‌شد و در نهایت مدل رگرسیونی برای قسمت مربوطه ارائه نمودند.Zare Abyaneh (2014) تفاوت میزان تخمین COD و BOD10 در تصفیه‌خانه فاضلاب شهری اکباتان تهران به دو روش استفاده از رگرسیون‌های خطی[۱۱] و شبکه‌های عصبی[۱۲] را بررسی کرد. همچنین میزان قدرت مدل‌ها را نیز با R و RMSE13 اندازه‌گیری نمود. نتایج حاکی از آن بود که قدرت ANN به‌مراتب از MLR بیشتر است. همچنین Ji Li et al.(2017) با توجه به اهمیت تخمین پارامتر اکسیژن‌خواهی شیمیایی در سیستم‌های فاضلاب در تحقیق خود روش‌های زیادی برای تخمین این پارامتر را مقایسه نموده و معایب و مزایای آن‌ها را بررسی کردند.

موردی که همه محققها در صدد آن بوده و در مورد آن کوتاهی نموده‌اند استفاده از داده‌های اصطلاحاً پرت برای متغیرهای ورودی و خروجی واحدهای تصفیه‌خانه فاضلاب می‌باشد که مقادیرشان با مقادیر دیگر مشاهدات متفاوت است. در واقع در پیش‌پردازش‌های داده‌های جمع‌آوری شده، قبل از استفاده از برخی روش‌های آزمون آماری، آن‌ها را کنار می‌گذارند، در حالی‌که ممکن است زنگ هشداری برای مدل‌سازی باشند و نباید این‌گونه عمل کرد (Dennis et al., 1992). در همین راستا باید در برخورد با داده‌های پرت ابتدا تحقیق کرد که وجود داده‌های پرت ناشی از خطای اندازه‌گیری یا نوشتاری نباشد و در صورتی‌که موردی مشاهده نشد ترجیح بر این است که نقطه پرت را در نظر گرفت. زیرا ممکن است داده معتبری باشد و در صورتی‌که امکان تکرار با درنظر گرفتن وقت و هزینه و امکان عمل نباشد، بهتر است داده غیرعادی در محاسبات حذف نشود. بنابراین در صورتی‌که با روش‌های مختلف نرمال‌سازی نتوان نرمال بودن داده‌های بعضی متغیرها را انجام داد، به‌جای فرایند حذف بهتر است از روش‌های مقاوم به داده‌های پرت استفاده کرد. همچنین استفاده از مدل‌های ریاضی و خطی هرچند ممکن است نسبت به مدل‌های غیرخطی از دقت کمتری برخوردار باشد، ولی درعوض ابزاری مناسب برای کارهای نظارتی است که در آن اپراتور با استفاده از کم‌ترین ابزارها می‌تواند تاثیر فرایندهای انجام شده در سیستم را تخمین بزند و حتی به یک خروجی ملموس

تابعی با یک‌سری متغیرهای ورودی برسد، در صورتی‌که این امکان برای مدل‌های غیرخطی شبیه شبکه‌های عصبی و یا ماشین‌های بردار پشتیبان وجود ندارد.

هدف اصلی این تحقیق تخمین خطی و ملموس داده‌های خروجی واحد زلال‌ساز تصفیه خانه فاضلاب فجر با ساده‌ترین و بهترین حالت و با استفاده از شناخت داده‌های ورودی تصفیه‌خانه در هر بخش است. به‌عبارتی در مدیریت فرایندها، به‌جای استفاده از ابزار آزمایشگاهی که نیازمند صرف وقت و هزینه است (حسنلو و همکاران، ۱۳۹۱) از ساده‌ترین حالت آماری استفاده می‌شود.

۲- مواد و روش‌ها

۲-۱- منطقه مورد مطالعه

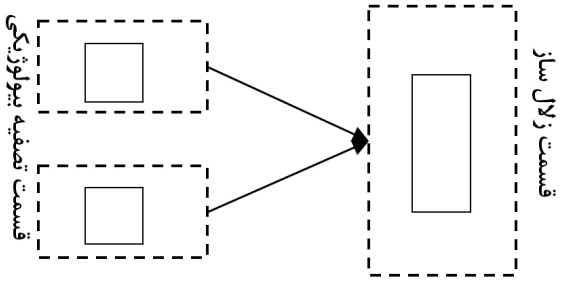
شرکت پتروشیمی فجر در سال ۱۳۷۷ با هدف تامین مواد موردنیاز مجتمع‌های منطقه ویژه اقتصادی پتروشیمی بندر امام خمینی به‌صورت متمرکز احداث شد. این منطقه در بین عرض جغرافیایی $30^{\circ}29'00.2"$ شمالی و طول جغرافیایی $49^{\circ}04'59.8"$ شرقی قرار دارد. واحد تصفیه پساب در این مجموعه دارای دو بخش شامل قسمت تصفیه پساب‌های روغنی با نمک پایین و فاضلاب بهداشتی و قسمت تصفیه پساب شیمیایی با نمک بالا می‌باشد. این مجموعه برای تصفیه انواع پساب‌های روغنی، شیمیایی و بهداشتی طراحی شده و ظرفیت آن ۴۶۰ مترمکعب بر ساعت است (شریعت‌زاده، ۱۳۸۸). فاضلاب ورودی در قسمت پساب شیمیایی با نمک پایین از طریق یک خط ۱۸ اینچ وارد واحد شده و به طرف یک چاله روغنی هدایت می‌شود. در این چاله همچنین آب‌های ناشی از دورریز سیستم‌ها و جریان‌های برگشتی از مخزن ذخیره نیز وارد می‌شود. این چاله مجهز به یک غربال میله‌ای برای جداسازی ذرات ریز است. جریان ورودی بعد از عبور از غربال از طریق نیروی ثقل به طرف چاله روغنی هدایت می‌شود. برای چاله روغن دو دریچه سرریز به‌طرف مخزن ذخیره طراحی شده است. جریان ورودی در حدود ۴۲۰ مترمکعب بر ساعت است که اگر بیشتر از این مقدار شود از طریق دریچه‌های سرریز به مخزن ذخیره هدایت می‌شود. حوضچه ذخیره‌سازی برای ذخیره مقادیر اضافی و سرریزها و نیز ذخیره پساب‌های خارج از طراحی مورد استفاده قرار می‌گیرد. این حوضچه دارای دو خط سرریز ۱۴ اینچی است که از طریق آن‌ها پساب اضافی به‌طرف خور هدایت می‌شود. به‌منظور کنترل پساب‌های ورودی، در ابتدای شیفت از ورودی واحد نمونه‌گیری شده و سپس برای انجام تست‌های کیفی به آزمایشگاه واحد ارسال می‌شود. چنانچه هریک از موارد مذکور خارج از محدوده مجاز باشد، سریعاً با نمونه‌گیری از خروجی پساب مجتمع‌ها، مجتمع ارسال‌کننده متخلف شناسایی و نسبت به بستن خروجی واحد مذکور اقدام خواهد شد. بعد از تصفیه اولیه فاضلاب در سیستم‌ها، پساب تصفیه شده در قسمت بیولوژیکی پس از عبور از حوضچه‌های هوادهی از طریق دیواره تیغه‌ای موجود در حوضچه‌ها سرریز شده و توسط نیروی ثقل به طرف زلال‌ساز ارسال می‌شود.

پساب از طریق یک خط به مرکز زلال‌ساز وارد می‌شود. در مرکز زلال‌ساز یک توزیع‌کننده بتنی قرار دارد که پساب را از طریق سوراخ‌های خود به صورت مساوی و در جهات مختلف وارد مخزن دایره‌ای شکل زلال‌ساز می‌کند. در حقیقت پساب به‌گونه‌ای وارد زلال‌ساز می‌شود که باعث برهم‌زدن آب نخواهد شد. لجن فعال پس از ورود به زلال‌ساز با زمان ماندی که به آن داده می‌شود به آرامی از پساب تصفیه جدا شده و ته‌نشین می‌شود.

پساب تصفیه شده نیز از طریق دیواره‌های کنگره‌ای که دورتادور زلال‌ساز قرار گرفته سرریز کرده و با پساب زلال‌شده زلال‌ساز دیگر ترکیب شده و به مرحله بعد ارسال می‌شود. خروجی زلال‌ساز از طریق نیروی ثقل به‌طرف حوضچه کلرزنی جریان می‌یابد. بر سر راه خط خروجی زلال‌سازها به حوضچه کلرزنی، یک خط کوچک تزریق آب ژاول برای تامین کلر آزاد وجود دارد. آب خروجی از حوضچه کلرزنی که اینک آب تصفیه‌شده خوانده می‌شود از طریق تیغه سرریز به خروجی هدایت شده و به خور می‌ریزد. از این آب می‌توان برای آب باغبانی و یا رقیق‌سازی سایر پساب‌های ورودی به واحد نیز استفاده کرد (مردانی، ۱۳۸۸). مشخصات کیفی و آماری اندازه‌گیری شده در سیستم مطابق جدول ۱ بوده و در شکل ۱ نیز وضعیت سیستم مورد نظر در این تحقیق نمایش داده شده است.

جدول ۱- توزیع آماری متغیرهای موجود در خروجی واحد بیولوژیکی و زلال‌ساز

متغیر	میانگین	میانه	مد	مینیمم	ماکزیمم	انحراف از معیار
Output COD (mg/l)	۱۰۰/۹۵	۹۰	۵۰	۳۸	۲۹۷	۴۵/۷۶۱۵
COD (A) (mg/l)	۱۳۵/۶۳	۱۲۱/۵	۵۰	۴۹	۸۲۵	۷۴/۶۶
N (A) (mg/l)	۷/۴۳	۷/۲	-	۰/۴	۲۳/۲	۴/۰۴
P (A) (mg/l)	۱/۶۶۶	۱/۲	۱/۱	۰/۱	۱۵۰	۵/۴
(PH) A	۷/۲	۷/۳	۷/۴	۳/۹	۸/۸	۰/۵۷۷
SST (A) (mg/l)	۲۷۱/۲۱	۲۵۰	-	۵۰	۸۵۰	۱۲۹/۵۶۷
COD (B) (mg/l)	۱۴۲/۸۶	۱۲۶/۵	۵۰	۴۳	۸۱۴	۸۳/۶۵
N (B) (mg/l)	۸/۱۶۷	۷/۲	۸/۳	۰/۴	۴۶۰	۱۶/۹۷
P (B) (mg/l)	۱/۵۹۹	۱/۲	-	۰/۱	۱۴۳	۵/۱۴۹۸
PH (B) (mg/l)	۷/۱۹	۷/۳	۷/۴	۳/۹	۷/۵	۰/۵۸۳۳
(SST (B) (mg/l)	۲۸۷/۰۸	۲۶۰	۲۰۰	۵۰	۷۶۰	۱۴۲/۴۹۶



شکل ۱- سیستم جداگانه مربوط به ارتباط واحد زلال‌ساز و قسمت تصفیه بیولوژیکی

۲-۲- روش‌های استفاده شده در تحقیق

به‌طور کلی، هدف از رگرسیون، برآورد یا پیش‌بینی پارامتری خاص با استفاده از یک سری متغیرهایی موجود است. رگرسیون‌های خطی مبنا بر این فرض بنا شده‌اند که ارتباط متغیرها در طبیعت به‌صورت خطی است در صورتی‌که در عمل این‌چنین نیست و با توجه به ماهیت غیرخطی سیستم‌های تصفیه‌خانه صنعتی و از طرفی درنظر گرفتن سادگی و اجرایی بودن رگرسیون‌های خطی مبنا در مقابل پیچیدگی و غیراجرایی بودن مدل‌های غیرخطی مطالعات گذشته، باید رابطه (۱) را به کمک روش‌های مختلف ریاضی به تابع خطی برازش داد.

$$COD_C = F(COD_A, PH_A, \dots, SST_B) \quad (۱)$$

این معادله بیان‌گر آن است که متغیرهای مستقل چگونه روی متغیر وابسته (COD) تاثیر می‌گذارند و ارتباط بین آن‌ها چگونه است.

۲-۲-۱- رگرسیون خطی چندگانه

تابع رگرسیونی خطی چندگانه (MLR) به‌صورت معادله (۲) است (Bryan, 2009):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (۲)$$

که i=1, 2, …, n و y متغیر وابسته یا متغیر پاسخ، x_i متغیر مستقل، β_0 عرض از مبدا و β_i شیب معادله با n مشاهده هستند. خطای ε یک متغیر تصادفی است، به عبارتی مقدار آن تحت کنترل تحلیل‌گر نیست و از تغییرپذیری طبیعی در ذات سیستم نشات می‌گیرد. یکی از روش‌های رگرسیونی معروف، روش حداقل مربعات معمولی[۱۴] می‌باشد که بر پایه مینیمم کردن اختلاف مشاهدات و برآوردها در هر مرحله است. حال ممکن است در داده‌ها شرایط به‌گونه‌ای باشد که فرض

واریانس ثابت و نرمال بودن باقیمانده‌ها ایفا نشود، بنابراین در این شرایط باید به‌دنبال روش‌های تعمیمی و جایگزین برای برازش تابع به داده‌ها رفت که این روش‌ها شامل روش‌های غیرپارامتریک رتبه‌ای یا خطوطی که به‌جز مربع باقیمانده‌ها را مینیمم کند می‌باشد.

۲-۲-۲- رگرسیون مقاوم

یکی دیگر از مدل‌های رگرسیونی موفق رگرسیون مقاوم است که به‌روش‌های رگرسیونی گفته می‌شود که رفتار باثبات و مقاومی در برابر وجود داده غیر معمول دارند. بعضی از روش‌های معمول رگرسیون مانند کمترین مربعات در صورت صدق فرض‌های آنان به‌خوبی کار می‌کنند، اما در مورد داده‌هایی که از فرض‌های آنان تخلف می‌کنند شاید به‌خوبی عمل نکنند. به‌ویژه، روش کمترین مربعات نسبت به داده پرت حساس است. مسئله دیگر وجود ناهم‌واریانسی در داده است. روش‌های پارامتری و ناپارامتری مختلفی برای رگرسیون باثبات پیشنهاد شده است. در همه آن‌ها، روش کمترین قدرمطلق نسبت به کمترین مربعات، باثبات‌تر است (James and LeSage, 1998).

$$s = \sum_{i=1}^{n} |y_i - f(x_i)| \qquad (۳)$$

همان‌گونه که در معادله (۳) مشخص است به‌جای توان دوم خطای رگرسیون، از قدر مطلق خطا استفاده می‌شود که وجود داده پرت تاثیر کمتری بر قدر مطلق خطا نسبت به مربع خطا دارد.

۲-۲-۳- رگرسیون ریج[۱۵]

رگرسیون ریج یک روش حل برای تحلیل داده‌هایی است که از چند بعد دارای شرایط غیرخطی باشند. وقتی که داده‌های سیستم Multicolinearity باشند، در نتیجه روش حداقل مربعات معمولی ناریب می‌شود و واریانس آن‌ها از مقدار واقعی فاصله می‌گیرد. بنابراین با افزودن درجه‌ای از بایاس در تقریب رگرسیون، خطای استاندارد داده‌ها کاسته می‌شود. رابطه رگرسیونی به صورت ماتریسی به‌صورت معادله (۴) است.

$$Y = XB + e \qquad (۴)$$

که Y متغیر وابسته، X متغیر مستقل، B ضریب رگرسیون و e خطای باقیمانده‌ها هستند. در رگرسیون ریج نرمالایز کردن متغیرهای مستقل و وابسته اولین قدم است که بعد از آن‌که رگرسیون محاسبه شد دوباره مقادیر به حالت اولیه خود تبدیل می‌شوند. در این رگرسیون ابتدا مانند کمترین مربعات، پارامتر $\hat{B}$ محاسبه می‌شود (Efron et al., 2004):

$$\hat{B} = (X'X)^{-1}X'Y \qquad (۵)$$

روش رگرسیون با اضافه کردن یک مقدار K به ماتریس کرولیشن عمل می‌کند (Efron et al., 2004):

$$\tilde{B} = (X'X + KI)^{-1}X'Y \qquad (۶)$$

می‌توان اثبات کرد یک مقداری برای K وجود دارد که خطای رگرسیون از خطای روش معمول حداقل مربعات کمتر شود. همچنین برای محاسبه K روابط زیادی وجود دارد که یکی از معروف‌ترین آن رابطه Hoerl and Kinnard است. (James and Lesage, 1999)

۲-۲-۴- استنباط بیزی

بررسی ویژگی مجموعه‌ای از داده‌های تصادفی را استنباط آماری می‌نامند. وقتی داده‌ها از یک مدل احتمالی $X \backsim f(x|\theta)$ تبعیت کنند در واقع فرض است که تمامی داده‌ها از توزیع احتمالی موردنظر برخوردارند.

در استنباط آماری عکس این قضیه اتفاق می‌افتد. دو دیدگاه کلی برای انجام استنباط آماری وجود دارد که شامل استنباط فراوانی‌گرا[۱۶] و استنباط بیزی[۱۷] می‌باشد.

دیدگاه فراوانی‌گرا بیان‌گر این است که احتمال یک پیشامد برابر است با فراوانی نسبی وقوع آن پیشامد ولی استنباط بیزی برخلاف روش فراوانی‌گرا، پارامتر مجهول به عنوان متغیر تصادفی فرض می‌شود. یعنی اگر توزیع احتمالی پارامتر، به یک قسمتی از فضای پارامتر وزن بیشتری بدهد، بنابراین از دیدگاه بیزی، اعتقاد بیشتری نسبت به تعلق آن پارامتر به آن محدوده از فضای پارامتری وجود دارد. این اعتقاد را اعتقاد پیشین[۱۸] و توزیع احتمال آن را توزیع پیشین[۱۹] می‌نامند. روش استنباط بیزی با استفاده از داده‌ها، توزیع پیشین را به‌روز کرده و توزیع احتمالی جدید به عنوان توزیع پسین[۲۰] جایگزین می‌شود.

۲-۲-۵- روش مونت کارلو با زنجیره مارکوف (MCMC)

به مجموعه روش‌هایی که با استفاده از ایجاد اعداد تصادفی سعی در حل مسئله داشته باشند را روش مونت کارلو می‌گویند. در برآورد مارکوف ابتدا معادله (۷) در نظر گرفته می‌شود (Lawless and Wang, 2004):

$$\theta = \int_a^b g(x)\,dx \qquad (۷)$$

با درنظر گرفتن توزیع دلخواهی در بازه [a,b] با چگالی f(x) و همچنین پارامتر $h(x) = \frac{g(x)}{f(x)}$ معادله (۸) تشکیل می‌شود.

$$= E(h(x)) = \int_a^b \frac{g(x)}{f(x)} f(x)dx \qquad (۸)$$

حال از چگالی f نمونه تصادفی استخراج می‌شود. سپس $h(x_1),h(x_2),\ldots,h(x_n)$ نیز تصادفی بوده و معادله (۹) به‌صورت زیر تبدیل خواهد شد:

$$\Sigma_{i=1}^{n} h(x_i)) = \theta \qquad (۹)$$

در واقع برتری این روش نسبت به روش‌های عددی معمولی، در حل مسائل با بیش از یک بعد نمایان می‌شود.

۲-۲-۶- رگرسیون وایت با رویکرد کمترین مربعات[۲۱]

روش رگرسیونی وایت با رویکردی مشابه ریج عمل می‌کند به‌این صورت که اگر معادله زیر یک معادله رگرسیونی باشد:

$$y = XB + \varepsilon \qquad (۱۰)$$

که $E(\varepsilon) = 0$ و $E(\varepsilon\varepsilon') = \Phi$ که یک مقدار مثبت است. بنابراین به‌جای محاسبه OLS به‌صورت رابطه $\hat{B} = (X'X)^{-1}X'y$ می‌توان از رابطه (۱۱) استفاده کرد:

$$Var(\hat{B}) = (X'X)^{-1}X'\Phi X(X'X)^{-1} \qquad (۱۱)$$

که اگر خطاها دارای واریانس همسانی باشند رابطه $\Phi = \sigma^2 I$ برقرار می‌شود و معادله بالا ساده می‌شود. حال با توجه به $e_i = y_i - x_i\hat{B}$ و با توجه به این‌که x_i ردیف iام از X است بنابراین ماتریس کواریانس حداقل مربعات را طبق رابطه (۱۲) می‌توان تقریب زد:

$$OLSCM = \frac{\sum {e_i}^2}{N-K}(X'X)^{-1} \qquad (۱۲)$$

که N تعداد نمونه و K تعداد متغیرها هستند. رابطه OLSCM برای بررسی اطمینان از همگنی پراکنش یا همسانی واریانس داده‌های رگرسیون به‌کار می‌رود که درصورت صحیح بودن طبق روابط بالا و ساده شدن محاسبه Φ، محاسبه رگرسیون انجام می‌شود. در اغلب مواقع این اتفاق نمی‌افتد و باید از ماتریس HCCM[22] استفاده کرد. ایده‌ای که پشت این رابطه نهفته است، استفاده از ${e_i}^2$ برای تخمین Φ است که برابر رابطه $\hat{\Phi} = diag[{e_i}^2]$ می‌باشد و درنهایت معادله (۱۳) به‌دست می‌آید:

$$HC0 = (X'X)^{-1}X'\Phi X(X'X)^{-1} = (X'X)^{-1}X'diag[{e_i}^2]X(X'X)^{-1} \qquad (۱۳)$$

این معادله شایع‌ترین معادله HCCM است و همان‌طور که وایت و همکاران در سال ۱۹۸۰ بیان کردند، $HC0$ برای مواقعی که همسانی واریانس $\hat{B}$ نامشخص باشد کاربرد دارد. همچنین یک‌سری اصلاحات در زمان‌های بعد به این معادله تحت عنوان HC1,HC2 و HC3 توسط دیگر دانشمندان اعمال شد که در این تحقیق نمی‌گنجد (Bady and Baltagi, 2002).

۳- ارائه و تحلیل نتایج

با درنظر گرفتن داده‌های موجود در سیستم مطابق جدول ۱، قبل از استاندارد کردن داده‌های خروجی، یک گام[۲۳] هشت ساعتی به آن‌ها داده می‌شود. زیرا باتوجه به پیوستگی سیال در واحدها، کیفیت پساب ورودی سیستم صرفا به ورودی مایع در لحظه مشخصی نیست بستگی ندارد و ورودی‌های قبلی مایع هم روی داده‌های کیفی سیستم تاثیرگذار است. بنابراین داده‌های گذشته سیستم تا حد معین باید به مدل اضافه شود. بعد از اضافه کردن متغیرهای جدید که همان گام‌های متغیرهای قدیمی هستند و با استفاده از همان متغیرهای قبلی، مجموعا ۲۰ متغیر در دسترس خواهد بود. بعد از محاسبه ماتریس همبستگی آن‌ها و با درنظر گرفتن روابط متغیرها و خواص هریک از آن‌ها و ویژگی‌هایی که دارند، ۱۱ متغیر از ۲۰ متغیر به دلیل همبستگی بالا نسبت به متغیرهای دیگر از مدل حذف می‌شوند.

با کاهش متغیرها در مرحله اول، نتایج مورد بررسی قرار می‌گیرد و نهایتاً برای جلوگیری از اندازه‌گیری داده‌های

غیرضروری و کاهش زمان و هزینه‌ها، متغیرهایی که میزان همبستگی بالایی با هم دارند از ورودی مدل آماری کم می‌شود تا مدل خروجی ساده‌تر شود. در نهایت در انتهای پیش‌پردازش‌های انجام شده، تابع خطی دارای ۹ متغیر ورودی و ۲ متغیر خروجی مطابق جدول ۲ خواهد شد که در مرحله بعدی استاندارد خواهند شد.

جدول ۲- متغیرها و پارامترهای مورد استفاده در گزینش نهایی برای مدل

متغیرهای ورودی مدل	COD(A)
	N(A)
	P(A)
	PH(A)
	SST(A)
	COD(A) with Lag
	N with Lag (A)
	N(B)
	N with Lag (B)
متغیرهای خروجی مدل	COD(C)

بعد از انجام مراحل بالا با استفاده از روش نرمال سازی مطابق فرمول (۱۴) عمل می‌شود.

$$Z = 2 \times \frac{x_i - min\,(X)}{Max(X) - Min(X)} - 1 \qquad (۱۴)$$

که x_i بیانگر هرکدام از داده‌ها، X بیانگر ماتریس داده‌ها و Z معادل ماتریس داده‌های نرمال‌شده هستند. بعد از انجام پیش‌پردازش‌های مذکور برای ارزیابی مدل‌سازی‌های انجام شده، بعد از آن‌که داده‌ها به بخش‌های آموزشی و اعتبارسنجی و تست یا آزمایشی تقسیم شد، از داده‌های تست برای ارزیابی مدل‌های پیاده شده با استفاده از شاخص‌های ذکر شده در روابط زیر اقدام می‌شود:

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N}(y_i - y_i')^2}{\sum_{i=1}^{N} y_i^2}\right) \qquad (۱۵)$$

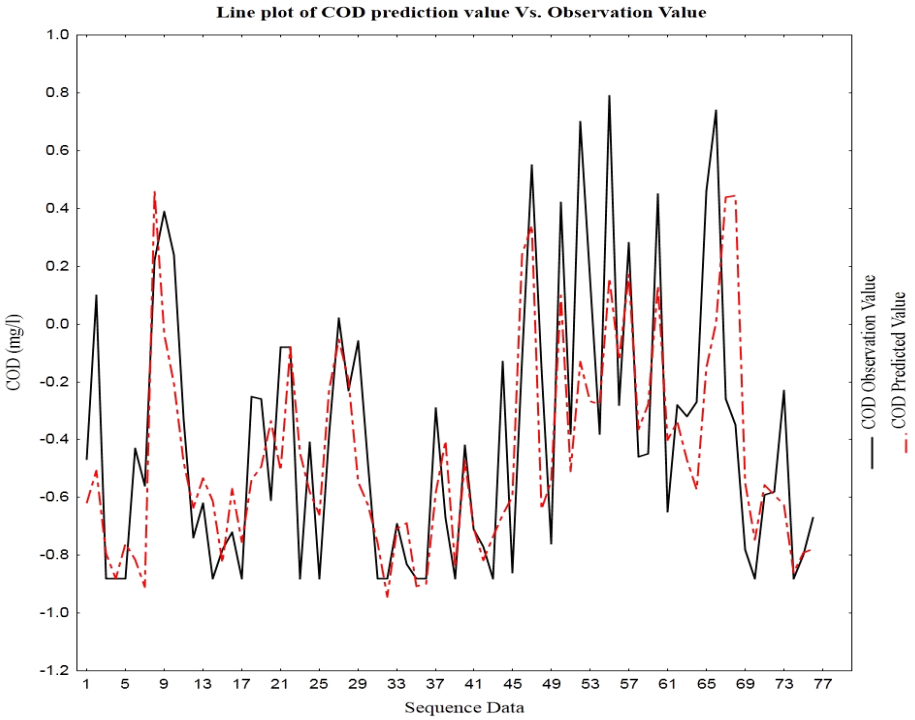
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - y')^2 \qquad (۱۶)$$

$$R = \frac{S_{XY}}{S_X S_Y} \qquad (۱۷)$$

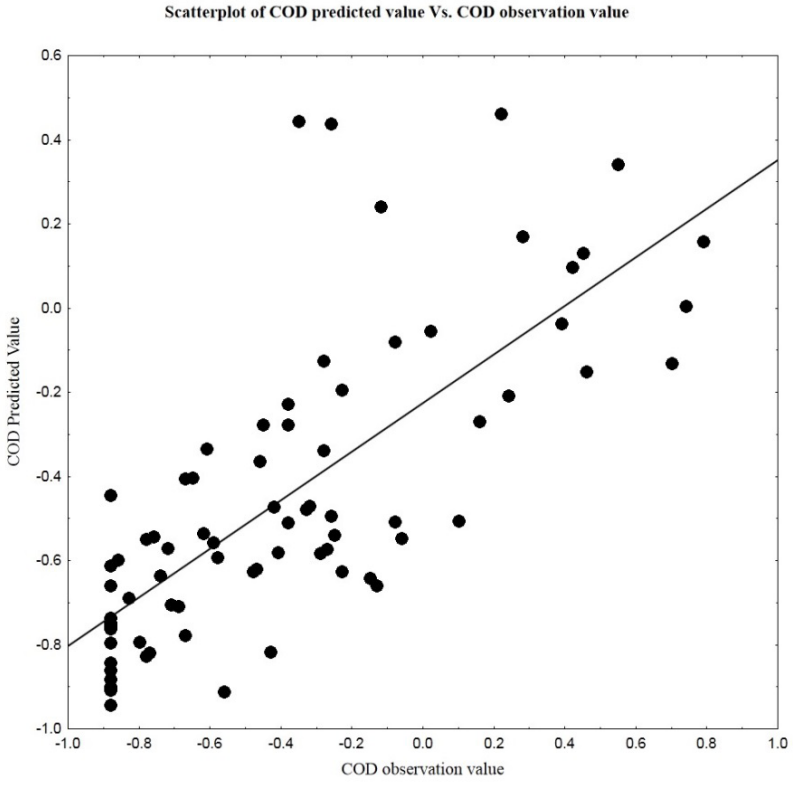
که y' نشان‌گر خروجی مدل و y نشان‌گر مشاهدات اندازه‌گیری هستند. در نهایت، همه محاسبات در برنامه MATLAB 2016-a و با شرایط مشابه پیاده شدند. خروجی مدل در جدول ۳ و شکل ۲ ارائه شده است.

جدول ۳- مقادیر شاخص‌های مدل ایجاد شده در تخمین اکسیژن‌خواهی شیمیایی واحد زلال‌ساز

ردیف	نوع مدل	R^2	MSE	R
۱	رگرسیون مقاوم	۰/۶۰۹۶	۰/۰۸۹۰	۰/۷۸۴۸
۲	رگرسیون ریج	۰/۵۷۳۴	۰/۰۹۰۰	۰/۷۵۰۰
۳	رگرسیون وایت با رویکرد کمترین مربعات	۰/۵۶۸۲	۰/۰۹۰۱	۰/۷۵۴۰
۴	کمترین مربعات با استفاده از روش مونت کارلو با زنجیره مارکوف و استنباط بیزی	۰/۵۵۶۲	۰/۰۹۱۵	۰/۷۴۹۷
۵	روش کمترین مربعات معمولی	۰/۵۶۲۶	۰/۰۹۰۲	۰/۷۵۱۲



(الف)



(ب)

شکل ۲- الف) مقادیر اکسیژن‌خواهی شیمیایی تخمینی و مشاهداتی

برای داده‌های تست، ب) نمودار خطی رابطه بین اکسیژن‌خواهی شیمیایی مشاهداتی و تخمینی مربوط به مدل رگرسیونی مقاوم

با توجه به این‌که مدل رگرسیون مقاوم با میزان کارایی R=0.784727 بهترین تقریب را در خصوص میزان اکسیژن‌خواهی شیمیایی واحد زلال‌ساز تصفیه‌خانه فاضلاب پتروشیمی فجر ارائه نمود، رابطه پیشنهادی این تحقیق با توجه به پارامترهای کیفی خروجی قسمت بیولوژیکی و با توجه به روابط (۴) الی (۱۳) به‌صورت رابطه (۱۸) است که t نشان‌گر زمان حاضر و $t\text{-}8hr$ بیان‌گر زمان مربوط به ۸ ساعت قبل هستند.

رگرسیون‌های معمول خطی، عملکرد ضعیفی دارد ولی با توجه به مزایای بی‌شمار رگرسیون‌های خطی مبنا در مقابل مدل‌های پیچیده و غیرملموس غیرخطی باید با اصلاح آن‌ها و استفاده از روش‌های مختلف روی داده‌ها برای ورود به مدل، به کارایی آن‌ها در تخمین پارامترهای کیفی سیستم همچون COD افزود که در این تحقیق این امر با دقت قابل قبولی حاصل شد.

$$COD_{(t)} = \left(0.20439 \times COD_{A(t-8hr)}\right) + \left(0.036955 \times N_{A(t-8hr)}\right) + \left(0.012338 \times P_{A(t-8hr)}\right) - \left(0.13179 \times PH_{A(t-8hr)}\right) + \left(0.060349 \times SST_{A(t-8hr)}\right) - \left(0.080425 \times COD_{A(t)}\right) + \left(0.665653 \times N_{A(t)}\right) - \left(0.060715 \times N_{B(t-8hr)}\right) + (0.117426 \times N_{B(t)}) \quad (۱۸)$$

۴- نتیجه‌گیری

در این تحقیق، ۹ رگرسیون خطی با ویژگی‌های ریاضی مختلف و با دو سناریو مختلف تصحیح و عدم تصحیح داده‌ها با استفاده از نرم‌افزار MATLAB محاسبه و اجرا شدند تا هم عملکرد رگرسیون‌ها برای یافتن بهترین تابع بررسی شود و هم توانایی آن‌ها در برابر داده شوک یا پرت مورد سنجش قرار گیرد. همچنین با استفاده از ماتریس کرولیشن، قبل از برازش، شماری از متغیرهای ورودی با ضریب همبستگی بالا مطابق جدول ۳ حذف شدند تا برای سادگی مدل، تعداد متغیرهای ورودی به کمترین مقدار ممکن برسد.

در نهایت، مدل رگرسیون مقاوم در مقایسه با دیگر روش‌های رگرسیونی موجود، میزان اکسیژن خواهی شیمیایی را تا میزان R=۰/۷۸۵ تخمین زد. یکی از دلایل موفقیت این مدل، ساختار ریاضی آن در مواجهه با داده‌های شوک بود که از آن می‌توان به‌عنوان یک رابطه خطی چندگانه در تخمین میزان اکسیژن‌خواهی شیمیایی واحد زلال‌ساز استفاده کرد.

همان‌گونه که از جدول ۳ مشخص است روش مرسوم رگرسیون خطی، همچون روش کمترین مربعات و روش رگرسیون وایت، در برابر داده‌های اصلاح نشده مقاوم مقاومت کمتری داشتند.

در نهایت می‌توان گفت که استفاده از بسیاری از

۵- پی‌نوشت‌ها

1- Black Box
2- Chemical Oxygen Demand (COD)
3- Coefficient of Determination (R-Square)
4- Mean square error (MSE)
5- Correlation of Coefficient (R)
6- Robust regression
7- Intelligent methods
8- Full-Scale
9- Particular least square (PLR)
10- Biological Oxygen Demand
11- Multiple Linear Regression (MLR)
12- Artificial Neural Network (ANN)
13- Relative Mean Square Error
14- Ordinary Least Square (OLS)
15- Ridge Regression Estimates
16- Frequentistic Inference
17- Bayesian Inference
18- Prior Belief
19- Prior distribution

20- Posterior Distribution

21- White›s adjusted heteroscedastic consistent Least--squares Regression

22- Heteroscedasaticity Consistent Covariance Matrix

23- Lag

۶- مراجع

حسنلو، ح.، مهردادی، ن.، نائب، ح. و گلبابایی، ف.، (۱۳۹۱)، «استفاده از روش تحلیل عاملی در مدل‌سازی عصبی واحد تصفیه پساب با نمک پایین تصفیه‌خانه فجر»، *هفتمین همایش ملی و نمایشگاه تخصصی مهندسی محیط‌زیست*، تهران، دانشکده محیط‌زیست.

شریعت‌زاده، م.، (۱۳۸۸)، «سیمای زیست‌محیطی پتروشیمی فجر»، *نشریه داخلی پتروشیمی فجر*.

مردانی، ن.، (۱۳۸۸)، «معرفی فرایند تصفیه پساب تصفیه‌خانه فجر»، *نشریه داخلی پتروشیمی فجر*.

Badi, H., and Baltagi, (2002), *Econometrics*, 3rd Edition, Springer.

Bates, B.C., and Campbell, E.P., (2001), "A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling", *Water Recourses Research*, 37(4), 937-947.

Bryan, F., and Manly, J., (2009), *Statistics for environmental science and management*, Taylor and Francis Group, International Standard Book Number-13: 978-1-4200-6147-5 (Hardcover).

Curlin, M., Bevetek, A., Ležajić, Z., Deverić Meštrović, B., and Kurtanjek, Z., (2008), "Modelling of activated sludge wastewater treatment process in municipal plant in Velika Gorica", *Chemistry in Industry*, 57(2), 59-67.

Helsel, D.R., and Hirsch, R.M., (1992), *Statistical methods in water resource*, USGS.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., (2004), "Least angle regression", *Annals of Statistics*, 32(2), 409-499.

LeSage, J.P., (1998), *Spatial econometrics*, Department of Economics, University of Toledo.

LeSage, J.P., (1999), *Applied econometrics using MATLAB*, Department of Economics University of Toledo.

Li, J., Luo, G., He, L.J., Xu, J., and Lyu, J., (2017), "Analytical approaches for determining chemical oxygen demand in water bodies, A review", *Critical Reviews in Analytical Chemistry*, 48(1), 47-65.

Lawless, J.F., and Wang, P., (1976), "A simulation study of Ridge and other regression estimators", *Communications in Statistics, Part A-Theory and Method*, 5, 307-323.

Chun, T.S., M. A. Malek, M.A., and Ismail, A.R., (2017), "A review of wastewater treatment plant modelling: Revolution on modelling technology", *American Journal of Environmental and Resource Economics*, 2(1), 22-26.

Wei, X., and Kusiak, A., (2015), "Short-term prediction of influent flow in wastewater treatment plant", *Stochastic Environmental Research Risk Assessment*, 29(1), 241-249.

Zare Abyaneh, H., (2014), "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters", *Journal of Environmental Health Science Engineering*, 12(40), 1-8.